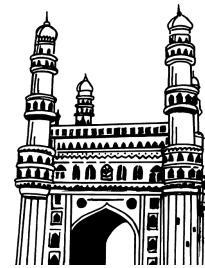


***Rahul's*** ✓  
*Topper's Voice*



# M.B.A.

***I Semester***  
***(Osmania University)***

**Latest 2022 Edition**

## STATISTICS FOR MANAGEMENT

- ☞ Study Manual
- ☞ FAQ's and Important Questions
- ☞ Short Question & Answers
- ☞ Exercise Problems
- ☞ Choose the Correct Answers
- ☞ Fill in the blanks
- ☞ Solved Model Papers
- ☞ Solved Previous Question Papers

- by -

**WELL EXPERIENCED LECTURER**



***Rahul Publications***<sup>TM</sup>

Hyderabad. Ph : 66550071, 9391018098

All disputes are subjects to Hyderabad Jurisdiction only

**M.B.A.**  
***I Semester***  
***(Osmania University)***

**STATISTICS FOR  
MANAGEMENT**

*Inspite of many efforts taken to present this book without errors, some errors might have crept in. Therefore we do not take any legal responsibility for such errors and omissions. However, if they are brought to our notice, they will be corrected in the next edition.*

© No part of this publications should be reporduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording and/or otherwise without the prior written permission of the publisher

***Price ` 199-00***

---

**Sole Distributors :**

**☎ : 66550071, Cell : 9391018098**

**VASU BOOK CENTRE**

**Shop No. 2, Beside Gokul Chat, Koti, Hyderabad.**

**Maternity Hospital Opp. Lane, Narayan Naik Complex, Koti, Hyderabad.**

**Near Andhra Bank, Subway, Sultan Bazar, Koti, Hyderabad -195.**

# STATISTICS FOR MANAGEMENT

## STUDY MANUAL

FAQ's and Important Questions	V - IX
Unit - I	1 - 102
Unit - II	103 - 138
Unit - III	139 - 186
Unit - IV	187 - 240
Unit - V	241 - 314
Tables	315 - 325

## SOLVED MODEL PAPERS

Model Paper - I	326 - 327
Model Paper - II	328 - 329
Model Paper - III	330 - 331

## SOLVED PREVIOUS QUESTION PAPERS

December - 2020	332 - 333
November - 2020	334 - 335
May / June - 2019	336 - 336
July - 2018	337 - 338
July / August - 2017	339 - 340

# SYLLABUS

## UNIT - I

**Introduction to Statistics:** Overview, origin and development and Managerial Applications of statistics, Measures of Central Tendency, Dispersion, Skewness and Kurtosis.

**Introduction to probability:** Concepts and Definitions of Probability – Classical, Relative, frequency, subjective and axiomatic. Addition and Multiplication theorems, Statistical independence, Marginal, Conditional and Joint Probabilities.

**Bayes' theorem** and its applications.

## UNIT – II

**Probability Distribution:** Random Variable (RV), Expectation and Variance of a RV. Probability distribution, function, properties, Continuous and Discrete Probability distribution functions.

**Discrete Probability distributions:** Binomial Distribution, Properties and applications; Poisson distribution, properties and applications.

**Continuous Probability Distributions:** Normal Distribution, Standard Normal Distribution properties, applications and importance of Normal Distribution.

## UNIT – III

**Sampling Theory :** The basics of sampling-Sampling procedures-Random and Non-Random methods- Sample size determination-Sampling distribution, Standard Error, Central Limit Theorem.

**Hypothesis Testing :** Statistical Estimation, Point and Interval Estimation, Properties of a Good Estimator, confidential interval.

**Large Sample tests:** Test for one and two proportions, Test for one and two means, Test for two S.D's.

## UNIT – IV

**Small Sample Tests- t-Distribution:** properties and applications, testing for one and two means, paired t-test.

**Analysis of Variance:** One Way and Two ANOVA (with and without Interaction).

**Chi-square distribution:** Test for a specified Population variance, Test for Goodness of fit, Test for Independence of Attributes.

## UNIT – V

**Correlation Analysis:** Scatter diagram, Positive and negative correlation, limits for coefficient of correlation, Karl Pearson's coefficient of correlation, Spearman's Rank correlation, concept of multiple and partial Correlation.

**Regression Analysis:** Concept, least square fit of a linear regression, two lines of regression, properties of regression coefficients.

**Time Series Analysis:** Components, Models of Time Series-Additive, Multiplicative and Mixed models; Trend analysis-Free hand curve, Semi averages, moving averages, Least Square methods



# Contents

## UNIT - I

Topic	Page No.
1.1 Introduction to Statistics .....	1
1.1.1 Overview .....	2
1.1.2 Origin and Development .....	3
1.2 Managerial Applications of Statistics .....	9
1.3 Measures of Central Tendency .....	10
1.3.1 Arithmetic Mean .....	11
1.3.2 Mode .....	19
1.3.3 Median .....	25
1.3.4 Relationship among Mean, Median and Mode .....	29
1.4 Dispersion .....	33
1.4.1 Absolute and Relative Measures .....	34
1.4.1.1 Range .....	35
1.4.1.2 Quartile Deviation .....	37
1.4.1.3 Mean Deviation .....	42
1.4.1.4 Standard Deviation .....	46
1.5 Skewness .....	52
1.5.1 Karl Pearson's Coefficient of Skewness - Bowley's Coefficient of Skewness - Kelly's Measure of Skewness .....	53
1.6 Kurtosis .....	61
1.6.1 Mesokurtosis, Platy Kurtosis and Leptokurtosis .....	61
1.7 Introduction to Probability .....	62
1.7.1 Concepts and Definitions of Probability .....	62
1.8 Approaches to Probability .....	67
1.8.1 Classical .....	67
1.8.2 Relative / Empiricals .....	68
1.8.3 Subjective .....	69
1.8.4 Axiomatic .....	69
1.9 Theorems of Probability .....	71
1.9.1 Addition .....	71
1.9.2 Multiplication .....	76
1.10 Statistical Independence .....	79
1.10.1 Marginal, Conditional and Joint Probabilities .....	79
1.11 Bayes' Theorem .....	81

Topic	Page No.
1.11.1 Applications .....	81
➤ <b>Short Question and Answers</b> .....	<b>87 - 95</b>
➤ <b>Exercise Problems</b> .....	<b>96 - 99</b>
➤ <b>Choose the Correct Answers</b> .....	<b>100 - 101</b>
➤ <b>Fill in the Blanks</b> .....	<b>102 - 102</b>

## UNIT - II

2.1	Probability Distribution .....	103
2.1.1	Random Variable (RV) .....	103
2.1.2	Expectation and Variance of a RVs .....	105
2.2	Probability Distribution Function .....	107
2.2.1	Properties .....	107
2.3	Types of Probability Distribution Function .....	108
2.4	Discrete Probability Distribution Functions .....	109
2.4.1	Binomial Distribution .....	109
2.4.1.1	Properties .....	110
2.4.1.2	Applications .....	110
2.4.2	Poisson Distribution .....	113
2.4.2.1	Properties .....	114
2.4.2.2	Applications .....	115
2.5	Continuous Probability Distributions .....	126
2.5.1	Normal Distribution .....	126
2.5.1.1	Standard Normal Distribution Properties .....	127
2.5.1.2	Applications .....	128
2.5.1.3	Importance of Normal Distribution .....	128
➤	<b>Short Question and Answers</b> .....	<b>132 - 135</b>
➤	<b>Exercise Problems</b> .....	<b>136 - 136</b>
➤	<b>Choose the Correct Answers</b> .....	<b>137 - 137</b>
➤	<b>Fill in the Blanks</b> .....	<b>138 - 138</b>

## UNIT - III

3.1	Sampling Theory .....	139
3.1.1	The Basics of Sampling .....	139
3.1.2	Sampling Procedures .....	141
3.2	Random and Non-Random Methods .....	141
3.3	Sample Size Determination .....	145

Topic	Page No.
3.4 Sampling Distribution .....	146
3.5 Standard Error .....	147
3.6 Central Limit Theorem .....	148
3.7 Hypothesis Testing .....	148
3.8 Statistical Estimation .....	153
3.8.1 Point and Interval Estimation .....	153
3.8.2 Properties of a Good Estimator .....	154
3.8.3 Confidential Interval .....	155
3.9 Large Sample Tests .....	161
3.9.1 Test for One Proportion .....	161
3.9.2 Test for Two Proportions .....	165
3.9.3 Test for One Mean .....	170
3.9.4 Test for Two Means .....	172
3.9.5 Test for two S.D's .....	174
➤ <b>Short Question and Answers</b> .....	<b>177 - 181</b>
➤ <b>Exercise Problems</b> .....	<b>182 - 182</b>
➤ <b>Choose the Correct Answers</b> .....	<b>183 - 184</b>
➤ <b>Fill in the Blanks</b> .....	<b>185 - 186</b>

#### UNIT - IV

4.1 Small Sample Tests .....	187
4.1.1 t-Distribution .....	187
4.1.2 Properties .....	188
4.1.3 Applications .....	188
4.1.4 Testing for One and Two Means .....	188
4.1.5 Paired T-test .....	195
4.2 Analysis of Variance .....	199
4.2.1 One Way ANOVA .....	200
4.3 Chi-square Distribution .....	217
4.3.1 Test for a Specified Population Variance .....	218
4.3.2 Test for Goodness of Fit .....	219
4.3.3 Test for Independence of Attributes .....	226
➤ <b>Short Question and Answers</b> .....	<b>233 - 236</b>
➤ <b>Exercise Problems</b> .....	<b>237 - 237</b>
➤ <b>Choose the Correct Answers</b> .....	<b>238 - 239</b>
➤ <b>Fill in the Blanks</b> .....	<b>240 - 240</b>

Topic	Page No.
<b>UNIT - V</b>	
5.1 Correlation Analysis .....	241
5.1.1 Types of Correlation .....	243
5.1.1.1 Positive and negative correlation .....	243
5.1.2 limits for coefficient of correlation .....	246
5.2 Methods of Correlation .....	246
5.2.1 Scatter diagram .....	247
5.2.2 Karl Pearson's coefficient of correlation .....	249
5.3 Spearman's Rankcorrelation .....	259
5.4 Concept of Multiple and Partial Correlation .....	264
5.5 Regression Analysis .....	265
5.5.1 Concept .....	265
5.5.2 Least square fit of a linear regression .....	268
5.5.3 Two lines of regression .....	269
5.6 Time Series Analysis .....	285
5.6.1 Components .....	286
5.6.2 Models of Time Series .....	288
5.6.2.1 Additive, Multiplicative and Mixed Models .....	288
5.7 Trend Analysis .....	289
5.7.1 Free Hand Curve .....	290
5.7.2 Semi Averages .....	291
5.7.3 Moving Averages .....	293
5.7.4 Least Square Methods .....	299
➤ <b>Short Question and Answers</b> .....	<b>304 - 309</b>
➤ <b>Exercise Problems</b> .....	<b>310 - 311</b>
➤ <b>Choose the Correct Answers</b> .....	<b>312 - 313</b>
➤ <b>Fill in the Blanks</b> .....	<b>314 - 314</b>

## Frequently Asked & Important Questions

### UNIT - I

1. Explain briefly about the scope of statistics.

*Ans :* (Imp.)

Refer Unit-I, Q.No. 5.

2. State the various limitations of statistics.

*Ans :* (Imp.)

Refer Unit-I, Q.No. 7.

3. What are the Managerial Applications of Statistics?

*Ans :* (Imp.)

Refer Unit-I, Q.No. 8.

4. For the following data calculate mean, median and mode and comment on the same.

Class Interval	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
Frequency	7	14	18	20	11	10

*Sol :* (July.-18, Imp.)

Refer Unit-I, Prob. No. 16.

5. Explain Addition theorem of probability.

*Ans :* (Dec.-20, June-19, July-18)

Refer Unit-I, Q.No. 54.

6. Explain the concept of independent event. Discuss the different types of probabilities under Statistical Independence.

*Ans :* (Dec.-20, Aug.-17)

Refer Unit-I, Q.No. 56.

7. State and explain Baye's probability theorem.

*Ans :* (Nov.-20, Imp.)

Refer Unit-I, Q.No. 57.

### UNIT - II

1. What is Binomial Distribution? State the assumptions of Binomial Distribution.

*Ans :* (June-19, Imp.)

Refer Unit-II, Q.No. 9.

2. State the Properties of Binomial Distribution.

*Ans :* (June-19, Imp.)

Refer Unit-II, Q.No. 10.

3. Define Poisson distribution. State the assumptions of Poisson distribution.

*Ans :* (June-19, Imp.)

Refer Unit-II, Q.No. 12.

4. In a Research Methodology Book, the following frequency mistakes per page were observed. Fit a Poisson distribution.

No. of Mistakes	0	1	2	3	4	5
No. of Pages	620	180	80	60	40	80

*Sol :* (Dec.-20, Imp.)

Refer Unit-II, Prob. No. 6.

5. An automatic machine makes paper clips from coils of wire. On the average. 1 in 400 paper clips is defective. If the paper clips are packed in boxes of 100. What is the probability that any given box of clips will contain (i) no defective, (ii) one or more defective and (iii) less than two defectives.

*Sol :* (Aug.-17, Imp.)

Refer Unit-II, Prob. No. 10.

### UNIT - III

1. Explain various probabilistic sampling methods.

*Ans :* (Nov.-20, July-18, Aug.-17)

Refer Unit-III, Q.No. 6.

2. Differentiate between random and non-random sampling methods.

*Ans :* (July-18, Imp.)

Refer Unit-III, Q.No. 7.

3. State the concept of Central Limit Theorem.

*Ans :* (Dec.-20, Imp.)

Refer Unit-III, Q.No. 11.

4. Define Hypothesis. What are the characteristics of Hypothesis ?

*Ans :* (Dec.-20, June-19, Imp.)

Refer Unit-III, Q.No. 12.

5. Explain the procedure generally followed in testing of hypothesis.

*Ans :* (Dec.-20, June-19, Imp.)

Refer Unit-III, Q.No. 13.

6. State the properties of a good estimator.

*Ans :* (Nov.-20, June-19)

Refer Unit-III, Q.No. 16.

#### UNIT - IV

1. Discuss in detail about Paired t-test.

*Ans :* (July-18, Imp.)

Refer Unit-IV, Q.No. 6.

2. What is ANOVA? State the assumptions and applications of ANOVA.

*Ans :* (Imp.)

Refer Unit-IV, Q.No. 7.

3. Explain briefly about One Way ANOVA.

*Ans :* (Imp.)

Refer Unit-IV, Q.No. 8.

4. A manufacturing company wishes to test the average life of the four brands of electric bulbs. The company uses all brands in a randomly selected production plants. The records showing the lives (in "00" hours) of bulbs are as given in the table below:

Brand 1	Brand 2	Brand 3	Brand 4
22	21	23	17
25	17	21	19
20	19	22	18
19	22	19	20
	18	18	

Test the hypothesis that the average life for each brand of bulbs is the same. Assume alpha 1%.

*Sol :* (July-18, Imp.)

Refer Unit-IV, Prob. No. 11.

5. Four technicians analyzed three samples each of the moisture content in the sample. The results are given below :

Samples	Technicians			
	A	B	C	D
X	9	12	10	11
Y	12	11	15	12
Z	9	10	12	14

Analyze the data and comment. Use 5% significance level.

*Sol :*

(Aug.-17, Imp.)

Refer Unit-IV, Prob. No. 12.

6. Explain briefly about Chi-square test as test of goodness of fit.

*Ans :*

(Nov.-20, Imp.)

Refer Unit-IV, Q.No. 12.

7. Explain the Chi-Square Test for independence of attributes.

*Ans :*

(Imp.)

Refer Unit-IV, Q.No. 13.

## UNIT - V

1. How do you say that the correlation between the two variables is significant (or) not.

*Ans :*

(June-19, Imp.)

Refer Unit-V, Q.No. 2.

2. What are the different Types of correlations.

*Ans :*

(Imp.)

Refer Unit-V, Q.No. 5.

3. What is Karl Pearson's Coefficient of Correlation? Explain properties of Coefficient of Correlation.

*Ans :*

(Imp.)

Refer Unit-V, Q.No. 10.

4. Find Karl Pearson's coefficient correlation to the following:

X	40	45	53	55	38	42	45	62
Y	71	78	87	73	74	71	76	75

*Sol :*

(July-18, Imp.)

Refer Unit-V, Prob. No. 6.



5. What are the differences between Correlation and Regression.

*Ans :* (July-18, Imp.)

Refer Unit-V, Q.No. 26.

6. What do you mean by line of regression? Derive the equations of lines of regression.

*Ans :* (Imp.)

Refer Unit-V, Q.No. 28.

7. Find both regression lines to the following:

Mean (X) = 15

Mean (Y) = 110

Variance (X) = 25

Variance (Y) = 62.5 and  $r = 0.81$ .

*Sol :* (July-18, Imp.)

Refer Unit-V, Prob. No. 23.

8. What are the components of time series analysis.

*Ans :* (Imp.)

Refer Unit-V, Q.No. 33.

9. What is least square method and explain its advantages and disadvantages?

*Ans :* (Imp.)

Refer Unit-V, Q.No. 39.

# UNIT I

- (i) **Introduction to Statistics:** Overview, origin and development and Managerial Applications of statistics, Measures of Central Tendency, Dispersion, Skewness and Kurtosis.
- (ii) **Introduction to probability:** Concepts and Definitions of Probability – Classical, Relative, frequency, subjective and axiomatic. Addition and Multiplication theorems, Statistical independence, Marginal, Conditional and Joint Probabilities.
- (iii) Bayes' theorem and its applications.

## 1.1 INTRODUCTION TO STATISTICS

**Q1. Explain the meaning of the word statistics as used in different sense.**

(OR)

**Define the term statistics in both Plural and Singular Sense.**

(OR)

**Define statistics in singular sense.**

(OR)

**Define the statistics in plural sense.**

*Ans :*

### Statistics in Singular Sense

In singular sense, statistics acts as a helpful device used for the purpose of collecting, classifying, presenting and interpreting the data. It is otherwise known as 'Analytical statistics'.

Some of the definitions of statistics are presented as follows,

#### Definitions

- (i) **According to A.L. Bowley,** "Statistics is the science of counting".
- (ii) **According to A.L. Bowley,** "Statistics may rightly be called the science of averages".
- (iii) **According to Turtle,** "Statistics is a body of principles and techniques of collecting, classifying, presenting, comparing and interpreting the quantitative data".

Thus, the above definitions specify that statistics in singular sense, is a science which comprises different statistical methods used-for

collection, organization, classification, presentation and interpretation of data.

### Statistics in Plural Sense

In plural sense, the term "Statistics" implies systematical collection of numerical facts like statistics of population, production, price-level, national income and so on.

#### Definitions

- (i) **According to 'Dr. A.L. Bowley,** "statistics are numerical statements of facts in any department of enquiry placed in relation to each other".
- (ii) **According to 'L.R. Comor',** "Statistics are measurements, enumeration or estimates of natural or social phenomena, systematically arranged so as to exhibit their inter-relations".
- (iii) **According to 'Prof. Yule and Prof. Kendall',** "By statistics we mean quantitative data affected to a marked extent by multiplicity of causes".
- (iv) **According to 'Horace Secrist',** "By statistics we mean aggregate of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other".
- (v) **According to Webster,** "Statistics are classified facts respecting the conditions of the people in a state - especially those facts which can be stated in numbers or in tables of numbers or in any other tabular or classified arrangement."

- (vi) **According to Achenwall**, "A collection of noteworthy facts concerning state, both historical and descriptive."
- (vii) **According to H. Secrist**, By statistics we mean aggregate of facts affected to a marked extent by a multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other."

### 1.1.1 Overview

**Q2. What are the characteristics of statistics?**

(OR)

**Explain the characteristics of statistics.**

(OR)

**Enumerate the characteristics of a statistics.**

*Ans :*

#### 1. Statistics are aggregates of facts

Single and unconnected figures are not statistics, for example, if it is stated that Raj, a student, secured 50 marks, it is not statistics. However, marks secured by the students of the class would constitute statistics. A single figure relating to production income, marks height, etc. cannot be regarded as statistics but aggregates of such facts would be regarded as statistics.

#### 2. Statistics are affected to a marked extent by multiplicity of causes

Numerical facts are affected by a multiplicity of factors. For example, the price, of a commodity is affected by number of factors such as supply, demand, imports, exports, money in circulation, competitive products in the market and so on. It is very difficult to study separately the effect of these factors on the price of the commodity. In physical sciences, it is possible to isolate the effect of

various factors on a single item, but statistics are commonly used in social sciences and in social sciences it is very difficult to study the effect of any one factor separately. In statistical methods, the effects of various factors affecting a particular phenomenon are generally studied in a combined form, though attempts are also made to study the effects of different factors separately as well.

#### 3. Statistics are numerically expressed

Only numerical data constitutes statistics. Qualitative expressions like good, bad, young, old etc. cannot be regarded as statistics. Statement like the standard of living of people in India has improved or production of petroleum products has increased in India do not constitute statistics. But Statements like "Rice production in India in 2003-04 is expected to be 86.4 million tones as against 72.66 million tons in 2002-03" (Source Economic Survey 2003-2004) is statistics.

#### 4. Statistics are enumerated or estimated according to reasonable standards of Accuracy

Numerical information can be either enumerated or estimated. If they are enumerated i.e., actually counted or measured, they are supposed to be exact and accurate. If complete enumeration is not possible because of the large size of high cost, then data is estimated by using sampling technique. Estimated figures cannot be absolutely accurate and precise. The degree of accuracy depends to a large extent on the particular purpose for which are information is collected and the nature of the particular problem about which the data is collected. There cannot be uniform standard of accuracy for all types of data collection. For example, while calculating the marks of students in an entrance examination or the number of votes received by a candidate in an election, it is necessary that it should be absolutely accurate, but while calculating the number of persons watching a cricket match on Television, the numbers need not be accurate.

**5. Statistics are collected in a systematic manner**

The data should be collected in a very systematic manner. For any socioeconomic survey, a proper schedule depending on the object of enquiry should be prepared and trained investigators should be used to collect the data. An attempt should be made to reduce personal bias to the minimum. Data collected in haphazard manner may give inaccurate results.

**6. Statistics are collected for a predetermined purpose**

The Purpose of collecting the data should be pre-determined. Otherwise, the data collected may not serve any purpose and may become useless. One should not waste time and money in collecting information that is irrelevant for the enquiry. For example, if the purpose of enquiry is to measure the cost of living of low-income group people, we should collect information about the items that are generally consumed by them. Hence for such an index it is useless to collect information on items such as cars, refrigerators, cell-phones etc.

**7. Statistics should be placed in relation to each other**

Statistical information is collected mostly for the purpose of comparison. If the data collected cannot be compared, then much of the purpose of collection will be lost. The information collected should be homogeneous and not heterogeneous in character. Statistical data are often compared period wise or region-wise. For example, the data relating to population of a country for different years or population of different countries in some fixed period will be regarded as statistics. But data relating to the size of shoe of an individual and his intelligence quotient do not constitute statistics.

Statistics should contain the above characteristics. In the absence of such characteristics, numerical data cannot be called Statistics. Hence, "all statistics are numerical " 'statements of facts, but all numerical statements of facts are not statistics."

**1.1.2 Origin and Development**

**Q3. Explain the Origin and Development of Statistics.**

(OR)

**State the evolution of statistics.**

*Ans :*

**Introduction**

The word 'statistic' has been derived from the Latin word 'status' or the Italian word 'statistic' or the German word 'statistic' or the French word 'statistique' all of which mean a political state'. The 'State' collected data pertaining to population and their wealth, in order to plan new taxes and to fund wars. Hence, statistics was also known as 'science of state craft'.

**Evolution**

Statistics is said to be as old as recorded history. One of the earliest known statistics is the census conducted by the Emperors of Egypt in connection with construction of pyramids. Statistics have been found to be use in India even before 300 B.C. Historical evidences about the prevalence of a good system of collection of data are available in Kautilya's Arthshastra'.

Most of the data collection activity in the medieval times relates to population and output. There have been some studies on the movement of heavenly bodies by Tycho Brahe and Johannes Kepler in the sixteenth century, which led to discovery of three laws relating to movement of planets. This helped Sir Isaac Newton to formulate the theory of gravitation. The use of statistics increased in the seventeenth century, which saw the birth of 'Vital Statistics' Captain John Graunt of London, known as the Father of Vital Statistics, introduced a systematic study of birth and death statistics. The computation of mortality tables and calculation of life expectancy at different stages led to the establishment of the life Insurance Institution

in London in 1698. However, all these studies were carried out under the name of 'Political Arithmetic'.

The first study of the theory of probability was also done in the mid seventeenth century, inspired by an attempt to estimate the chances of winning (or) losing in a gamble. The study of J. Bernoulli, which contained and 'Law of Large numbers', was published in 1713. De-Moivre published his famous "Doctrine of Chance" in 1718 and also discovered the normal probability curve. The term 'statistics' is said to have been used as a 'subject matter' in 1749 by Gottfried Ache wall. The modern theory of statistics is said to have been formulated by L. A.J. Quelled who introduced the concept of 'average' and deviation from average. He discovered the principles of 'Constancy of Great Numbers' which forms the basis of sampling. There have been some grate contribution to the science of statistics in the nineteenth and twentieth centuries. Karl Pearson who conceptualized 'Chi-Square test.' founded the greatest statistical laboratory in England Sir Ronald Fisher pioneered the study of Estimation Theory' and also applied statistics to a variety of diversified fields such as genetics, biometry, psychology, agriculture and education. The application of scientific methods has considerably widened in the last 2 decades and there is hardly and field that does not make use of statistics. The use of information technology had made it easy and cost effective to use statistics.

Originally, statistics was restricted to recording and classification of data. Today, it has bloomed into a science and an analytical tool with significant applications in almost all disciplines ranging from Mathematics and Economics to areas such as Sociology, Psychology, etc.

#### **Q4. Explain the importance of statistics.**

**(OR)**

**Explain the importance of statistics in economic analysis and planning.**

*Ans :*

#### **1. Importance to the State**

We know that the subject of statistics originated for helping the ancient rulers in the assessment of their military and economic

strength. Gradually its scope was enlarged to tackle other problems relating to political activities of the State. In modern era, the role of State has increased and various governments of the world also take care of the welfare of its people. Therefore, these governments require much greater information in the form of numerical figures for the fulfillment of welfare objectives in addition to the efficient running of their administration.

In a democratic form of government, various political groups are also guided by the statistical analysis regarding their popularity in the masses. Thus, it can be said that it is impossible to think about the functioning of modern state in the absence of statistics.

#### **2. Importance in economics**

Statistics is an indispensable tool for a proper understanding of various economic problems. It also provides important guidelines for the formulation of various economic policies.

Almost every economic problem is capable of being expressed in the form of numerical figures, e.g., the output of agriculture or of industry, volume of exports and imports, prices of commodities, income of the people, distribution of land holding, etc. In each case, the data are affected by a multiplicity of factors. Further, it can be shown that the other conditions prescribed for statistical data are also satisfied. Thus, we can say that the study of various economic problems is essentially the one of a statistical nature."

#### **3. Importance in national income accounting**

The system of keeping the accounts of income and expenditure of a country is known as national income accounting. These accounts contain information on various macro-economic variables like national income, expenditure, production, savings, investments, volume of exports and imports, etc. The national income accounts of a country are very useful in having an idea about the broad features of its economy or of a particular region. The preparation of

these accounts require data, regarding various variables, at the macro-level. Since such information is very difficult, if not impossible, to obtain, is often estimated by using techniques and principles of statistics.

#### 4. Importance in planning

Planning is indispensable for achieving faster rate of growth through the best use of a nation's resources. It also requires a good deal of statistical data on various aspects of the economy. One of the aims of planning could be to achieve a specified rate of growth of the economy. Using statistical techniques, it is possible to assess the amounts of various resources available in the economy and accordingly determine whether the specified rate of growth is sustainable or not. The statistical analysis of data regarding an economy may reveal certain areas which might require special attention, e.g., a situation of growing unemployment or a situation of rising prices during past few years. Statistical techniques and principles can also guide the Government in adopting suitable policy measures to rectify such situations. In addition to this, these techniques can be used to assess various policies of the Government in the past. Thus, it is rather impossible to think of a situation where planning and evaluation of various policies can be done without the use of statistical techniques. In view of this it is sometimes said that, "Planning without statistics is a ship without rudder and compass Hence statistics is an important tool for the quantification of various planning policies.

#### 5. Importance in business

With the increase in size of business of a firm and with the uncertainties of business because of cut throat competition, the need for statistical information and statistical analysis of various business situations has increased tremendously. Prior to this, when the size of business used to be smaller without much complexities, a single person, usually the owner or manager of a firm, used to take all decisions regarding its business. For example, he used to decide, from where the necessary

raw materials and other factors of production were to be obtained, how much of output will be produced, where it will be sold, etc. This type of decision making was usually based on experience and expectations of this single man and as such had no scientific basis.

#### Q5. Explain briefly about the scope of statistics.

*Ans :*

(Imp.)

Statistics is viewed not just as a device for collecting numerical data, but as a means of sound techniques for their handling, analysis and for drawing valid inference from them. From this perspective, the scope (or) subject matter to statistics can be broadly studied under 2 hands namely.

(A) Statistical Methods and

(B) Applied Statistics

##### (A) Statistical Methods

Statistical methods are the tools that are in the hands of the statistician. They include all the general principles and techniques that are commonly used in the collection, analysis and interpretation of data. These methods are applicable to all kinds of data. The stage involved in study of any kind of data are:

(i) Observation and Collection

(ii) Organization

(iii) Presentation

(iv) Analysis and

(v) Interpretation.

##### (B) Applied Statistics

Applied Statistics deals with application of Statistical methods of specific problems or concrete forms. To illustrate, if a software services firm is experiencing attrition (loss of personnel), it may be worthwhile to investigate the reasons for the same. Special techniques can be employed to understand the underlying trends. For example, a correlation analysis could probably indicate that increasing opportunities (growth in the industry) and attrition are highly correlated.

All such techniques and the results obtained by employing such techniques form part of applied Statistics.

Applied Statistics can be further classified into different categories. These categories are

- (i) **Descriptive Statistics:** Descriptive Statistics deals with data that is known. The describe the main features of such data. They are 'basic' to any statistical analysis and bring out such characteristics of data which could have escaped attention of the statistician. Tabulation, Averages and trends are examples of such descriptive statistics.
- (ii) **Scientific Statistics:** Scientific statistics deals with formulation of statistical laws. These laws are based on quantitative data and generally hold true. For example, when we refer to the properties of standard normal distribution. Scientific statistics are heavily used for the purpose of business forecasting.
- (iii) **Analytical Statistics:** This includes methods such as correlation, regression, etc that help in establishing functional relationship between variables. Two or more sets of data are compared and analyzed to arrive at the relationship between them. Such relationship could be continuous or at a defined period of time.
- (iv) **Inferential Statistics:** Statistical methods that help us arrive at certain conclusion based on study of sample data are part of inferential statistics. It must be noted that no guess work is being made. The size of the sample, criteria for inclusion of a person/ activity/thing into the sample, etc are all clearly defined. Based on the results of the study on the sample, inferences about the total population can be drawn.
- (v) **Inductive Statistics:** Statistical methods that help in arriving at general consensus based on a study of random observations are part of inductive statistics. Unlike inferential statistics, the sample chosen under inductive statistics may not 'representative'. Hence, the conclusions arrived at need to be double checked.

#### Q6. What are the functions of statistics?

(OR)

**Explain the various functions of statistics.**

*Ans :*

The, following are the main functions of statistics:

##### 1. Presents facts in numerical figures

The first function of statistics is to present a given problem in terms of numerical figures. We know that the numerical presentation helps in having a better understanding of the nature of a problem. Facts expressed in words are not very useful because they are often vague and are likely to be understood differently by different people. For example, the statement that a large proportion of total work force of India is engaged in agriculture, is vague and uncertain. On the other hand, the statement that 70% of the total work force is engaged in agriculture is more specific and easier to grasp. Similarly, the statement that the annual rate of inflation in a country is 10% is more convincing than the statement that prices are rising.

##### 2. Presents complex facts in a simplified form

Generally a problem to be investigated is represented by a large mass of numerical figures which are very difficult to understand and remember. Using various statistical methods, this large mass of data can be presented in a simplified form. This simplification is achieved by the summarization of data so that broad features of the given problem are brought into focus. Various statistical techniques such as presentation of data in the form of diagrams, graphs, frequency distributions and calculation of average, dispersion, correlation, etc., make the given data intelligible and easily understandable.

##### 3. Studies relationship between two (or) more phenomena

Statistics can be used to investigate whether two or more phenomena are related. For example, the relationship between income

and consumption, demand and supply, etc., can be studied by measuring correlation between relevant variables. Furthermore, a given mathematical relation can also be fitted to the given data by using the technique of regression analysis.

**4. Provides techniques for the comparison of phenomena**

Many a times, the purpose of undertaking a statistical analysis is to compare various phenomena by computing one or more measures like mean, variance, ratios, percentages and various types of coefficients. For example, when we compute the consumer price index for a particular group of workers, then our aim could be to compare this index with that of previous year or to compare it with the consumer price index of a similar group of workers of some other city, etc. Similarly, the inequalities of income in various countries may be computed for the sake of their comparison.

**5. Enlarges individual experiences**

An important function of statistics is that it enlarges human experience in the solution of various problems. In the words of A.L. Bowley, "the proper function of statistics, indeed is to enlarge individual experience." Statistics is like a master key that is used to solve problems of mankind in every field. It would not be an exaggeration to say that many fields of knowledge would have remained closed to the mankind forever but for the efficient and useful techniques and methodology of the science of statistics.

**6. Helps in the formulation of policies**

Statistical analysis of data is the starting point in the formulation of policies in various economic, business and government activities. For example, using statistical techniques a firm can know the tastes and preferences of the consumers and decide to make its product accordingly. Similarly, the Government policies regarding taxation, prices, investments, unemployment, imports and exports, etc., are also guided by statistical studies in the relevant areas.

**7. Helps in forecasting**

The success of planning by the Government or of a business depends to a large extent upon the accuracy of their forecasts. Statistics provides a scientific basis for making such forecasts. Various techniques used for forecasting are time series analysis, regression analysis, etc.

**8. Provides techniques for testing of a hypothesis**

A hypothesis is a statement about some characteristics of a population (or universe). For example, the statement that average height of students of a college is 66 inches, is a hypothesis. Here students of the college constitute the population. It is possible to test the validity of this statement by the use of statistical techniques.

**9. Provides techniques for making decisions under uncertainty**

Many a times we face an uncertain situation where any one of the many alternatives may be adopted. For example, a person may face a situation of rain or no rain and he wants to decide whether to take his umbrella or not. Similarly, a businessman might face a situation of uncertain investment opportunities in which he can lose or gain. He may be interested in knowing whether to undertake a particular investment or not. The answer to such problems are provided by the statistical techniques of decision-making under uncertainty.

**Q7. Discuss the limitations of statistics.**

(OR)

**State the various limitations of statistics.**

(OR)

**List out the limitations of statistics.**

*Ans :* (Imp.)

**1. Statistics deals with numerical facts only**

Broadly speaking there are two types of facts:

- (a) quantitative and
- (b) qualitative.



- (a) **Quantitative:** Quantitative facts are capable of being represented in the form of numerical figures and therefore, are also known as numerical facts. These facts can be analysed and interpreted with the help of statistical methods.
- (b) **Qualitative:** Qualitative facts, on the other hand, represent only the qualitative characteristics like honesty, intelligence, colour of eyes, beauty, etc. and statistical methods cannot be used to study these types of characteristics. Sometimes, however, it is possible to make an indirect study of such characteristics through their conversion into numerical figures. For example, we may assign a number 0 for a male and 1 for a female, etc.
2. **Statistics deals only with groups and not with individuals**
- Statistical studies are undertaken to study the characteristics of a group rather than individuals. These studies are done to compare the general behaviour of the group at different points of time or the behaviour of different groups at a particular point of time. For example, the economic performance of a country in a year is measured by its national income in that year and by comparing national income of various years, one can know whether performance of the country is improving or not. Further, by comparing national income of different countries, one can know its relative position vis-a-vis other countries.
3. **Statistical results are true only on the average**
- Statistical results give the behaviour of the group on the average and these may not hold for an individual of that very group. Thus, the statement that average wages of workers of a certain factory is ` 1,500 p.m. does not necessarily mean that each worker is getting this wage. In fact, some of the workers may be getting more while others less than or equal to ` 1,500. Further, when value of a variable is estimated by using some explanatory

variable, the estimated value represents the value on the average for a particular value of the explanatory variable. In a similar way, all the laws of statistics are true only on the average.

4. **Statistical results are only approximately true**

Most of the statistical studies are based on a sample taken from the population. Under certain circumstances the estimated data are also used. Therefore, conclusions about a population based on such information are to be true only approximately.

5. **Statistical methods constitute only one set of methods to study a problem**

A given problem can often be studied in many ways. Statistical methods are used to simplify the mass of data and obtain quantitative results by its analysis. However, one should not depend entirely on statistical results. These results must invariably be supplemented by the results of alternative methods of analyzing the problem. It should be kept in mind that statistics is only a means and not an end.

6. **Statistics are liable to be misused**

Statistical data are likely to be misused to draw any type of conclusion. If the attitude of the investigator is biased towards a particular aspect of the problem, he is likely to collect only such data which give more importance to that aspect. The conclusions drawn on the basis of such information are bound to be misleading. Suppose, for example, the attitude of the Government is biased and it wants to compute a price index which should show a smaller rise of prices than the actual one. In such a situation, the Government might use only those price quotations that are obtained from markets having lower prices.

7. **Statistics must be used only by experts**

Statistics, being a technical subject, is very difficult for a common man to understand. Only the experts of statistics can use it correctly and derive right conclusions from the analysis.

**1.2 MANAGERIAL APPLICATIONS OF STATISTICS****Q8. What are the Managerial Applications of Statistics?**

*Ans :* (Imp.)

Statistics influence the operations of business and management in many dimensions. Statistical applications include the area of production, marketing, promotion of product, financing, distribution, accounting, marketing research, manpower planning, forecasting, research and development and so on.

As the organizational structure has become more complex and the market highly competitive, it has become necessary for executives to base their decisions on the basis of elaborate information systems and analysis instead of intuitive judgment. In such situations, statistics are used to analyze this vast data base for extracting relevant information. Some of the typical areas of business operations where statistics have been extensively and effectively used are as follows:

**1. Entrepreneuring**

If we are opening a new business or acquiring one, it is necessary to study the market as well as the resources from statistical point of view to ensure success of the new venture. A shrewd businessman must make a proper and scientific analysis of the past records and current market trends in order to predict the future course for business conditions. The analysis of the needs and wants of the consumers, the number of competitors in the market and their marketing strategies, availability of resources and general economic conditions and trends would all be extremely helpful to the entrepreneur. A number of new enterprises have failed either due to unreliability of data or due to faulty interpretations and conclusions.

**2. Production**

The production of any item depends upon the demand of that item and this demand must be accurately forecast using statistical techniques. Similarly, decisions as to what to produce and how much to produce are based

largely upon the feedback of surveys that are analyzed statistically.

**3. Marketing**

An optimum marketing strategy would require a skillful analysis of data on population, shifts in population, disposable income, competition, social and professional status of target market, advertising, quality of sales people, easy availability of the product and other related matters. These variables and their inter-relationships must be statistically studied and analyzed.

**4. Purchasing**

The purchasing department of an organization makes decisions regarding the purchase of raw materials and other supplies from different vendors. The statistical data in the cost structure would assist in formulating purchasing policies as to where to buy, when to buy, at what price to buy and how much to buy at it given time.

**5. Investment**

Statistics have been almost indispensable in making a sound investment whether it be in buying or selling of stocks and securities or real estate. The financial newspapers are full of tables and graphs analyzing the prices of stocks and their movements. Based upon these statistical data, a good investor will buy when the prices are at their lowest and sell when the prices are at their highest. Similarly, buying an apartment building would require that an investor take into consideration the rent collected, rate of occupancy, any rent control laws, cost of the mortgage obtained and the age of the building before making\* a decision about investing in real estate.

**6. Banking**

Banks are highly affected by general economic and market conditions. Many banks have research departments which gather and analyze information not only about general

economic conditions but also about businesses in which they may be directly or indirectly involved. They must be aware of money markets, inflation rates, interest rates and so on, not only in their own vicinity but also nationally and internationally. Many banks have lost money in international operations, sometimes in as simple a matter as currency fluctuations because they did not analyze the international economic trends correctly. Many banks have failed because they over-extended themselves in making loans without properly analyzing the general business conditions.

## 7. Quality Control

Statistics are used in quality control so extensively that even the phenomenon itself is known as statistical quality control. Statistical quality control (SQC) consists of using statistical methods to gather and analyze data on the determination and control of quality. This technique primarily deals with the samples taken randomly and as representative of the entire population, then these samples are analyzed and inferences made concerning the characteristics of the population from which these random samples were taken. The concept is similar to testing one spoonful from a pot of stew and deciding whether it needs more salt or not. The characteristics of samples are analyzed by statistical quality control and the use of other statistical techniques.

## 8. Personnel

Study of statistical data regarding wage rates, employment trends, cost of living indexes, work related accident rates, employee grievances, labor turnover rates, records of performance appraisal and so on and the proper analysis of such data assist the personnel departments in formulating the personnel policies and in the process of manpower planning.

### 1.3 MEASURES OF CENTRAL TENDENCY

**Q9. Define Measures of Central Tendency. State the characteristics of Measures of Central Tendency.**

*Ans :*

(Dec.-20)

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

#### Definitions

- (i) "A measure of central tendency is a typical value around which other figures congregate".
- (ii) "An averages stands for the whole group of which forms a part yet represents the whole".
- (i) "One of the most widely used set of summary figures is known as measures of location".

#### Need

The measures of central tendency are necessary for two reasons:

1. The average represents *all* the measurements made on a group and gives a *concise* description of the group as a whole.
2. When two or more groups are measured, the central tendency provides the basis of comparison between them.

### Characteristics

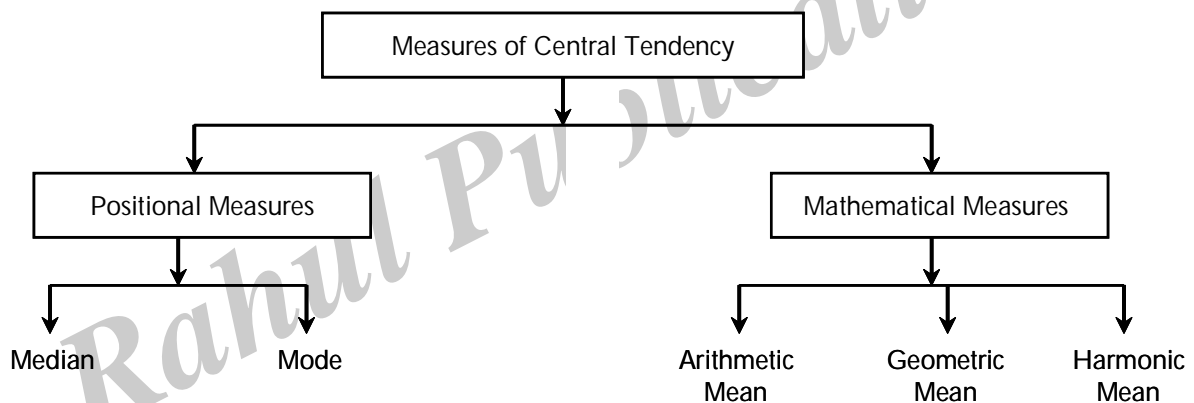
The following properties should be possessed by an ideal average :

- It should be clearly defined.
- It should be easy to understand and compute.
- It should be based on all items in the data.
- Its definition shall be in the form of a mathematical formula.
- It should be capable of further algebraic treatment.
- It should have sampling stability.
- It should be capable of being used in further statistical computations or processing
- A good averages should represent maximum characteristics of the data.
- Its value should be nearest to the most items of the given series.

### Q10. What are the various measures of central tendency?

*Ans :*

Various measures of averages can be classified into the following two categories :



#### 1.3.1 Arithmetic Mean

### Q11. What is arithmetic mean? State the merits and demerits of arithmetic mean.

*Ans :*

#### Meaning

Arithmetic Average (or) Mean of a series is the figure obtained by dividing the total "Values of the various items by their number. In other words it is the sum of the values divided by their number. Arithmetic means is the most widely used measure of central tendency.

#### Merits

1. It is rigidly defined. Hence, different interpretations by different persons are not possible.
2. It is easy to understand and easy to calculate. In most of the series it is determinate and its value is definite.
3. It takes all values into consideration. Thus, it is more representative.

4. It can be subjected to further mathematical treatment. The properties of Arithmetic mean are separately explained elsewhere in the chapter.
5. It is used in the computation of various other statistical measures.
6. It is possible to calculate arithmetic average even if some of the details of the data are lacking. For example, it can be known even when only the number of items and their aggregate value are known, and details of various items are not available. Similarly, the aggregate value of items can be calculated if the number of items and the average are known.
7. Of all averages, arithmetic average is least affected by fluctuations of sampling. Thus, it is the most stable measure of central tendency.
8. It provides a good basis for comparison.
9. Arithmetic mean is impacted by every observation. It gives weight to all items in direct proportion to their size.

### Demerits

While Arithmetic mean satisfies most of the conditions of an ideal average, it suffers from certain drawbacks. Some of the demerits or limitations of Arithmetic Mean are listed below:

1. It cannot be determined by inspection.
2. It cannot be located graphically.
3. It cannot be used in the study of qualitative phenomena.
4. It can be significantly impacted by extreme values and may lead to erroneous conclusions. Abnormal items may considerably affect this average, particularly when the number of items is not large. Thus, it is desirable not to use arithmetic average when the distribution is unevenly spread.

### Q12. How to calculate arithmetic mean for individual series ?

*Ans :*

#### Individual Series

The process of computing mean in case of individual observations (i.e., where frequencies are not given) is very simple. Add together the various values of the variable and divide the total by the number of items. Symbolically :

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \dots + X_n}{N}$$

(OR)

$$\bar{X} = \frac{\sum X}{N}$$

Here

$\bar{X}$  = Arithmetic Mean

$\sum X$  = Sum of all the values of the variable  $X$ , i.e.,  $X_1, X_2, X_3, \dots, X_n$

$N$  = Number of observations.

**Steps**

The formula involves two steps in calculating mean :

- i) Add together all the values of the variable X and obtain the total, i.e.,  $\Sigma X$ .
- ii) Divide this total by the number of the observations, i.e.,  $N$ .

**Short-cut method**

The arithmetic mean can be calculated by using what is known as an arbitrary origin. When deviations are taken from an arbitrary origin, the formula for calculating arithmetic mean is

$$\bar{X} = A + \frac{\Sigma d}{N}$$

where

A is the assumed mean and

d is the deviation of items from assumed mean

i.e.,  $d = (X - A^*)$ .

**Steps**

1. Take an assumed mean.
2. Take the deviations of items from the assumed mean and denote these deviations by d.
3. Obtain the sum of these deviations,
4. Apply the formula :  $\bar{X} = A + \frac{\Sigma d}{N}$ .

**Q13. How to calculate Arithmetic mean for discrete series ?**

*Ans :*

In discrete series arithmetic mean may be computed by applying

- i) Direct method (or)
- ii) Short-cut method

**i) Direct method**

The formula for computing mean is  $\bar{X} = \frac{\Sigma fX}{N}$

Where,

f = Frequency;

X = The variable in question;

$N^*$  = Total number of observations, i.e.,  $\Sigma f$ .

**Steps**

- i) Multiply the frequency of each row with the variable and obtain the total  $\Sigma fX$ .
- ii) Divide the total obtained by step (i) by the number of observations, i.e., total frequency.

**ii) Short-cut Method**

According to this method,

$$\bar{X} = A + \frac{\sum fd}{N}$$

where A = Assumed mean; d = (X – A); N - Total number of observations, i.e.,  $\sum f$ .

**Steps**

- i) Take an assumed mean.
- ii) Take the deviations of the variable X from the assumed mean and denote the deviations by d.
- iii) Multiply these deviations with the respective frequency and take the total  $\sum fd$ .
- iv) Divide the total obtained in third step by the total frequency.

**Q14. How to calculate arithmetic mean for continuous series win an example?**

*Ans :*

In continuous series, arithmetic mean may be computed by applying any of the following methods:

- i) Direct method,
- ii) Short-cut method.

**i) Direct Method**

When direct method is used

$$\bar{X} = \frac{\sum fm}{N}$$

where

m = Mid-point of various classes

f = The frequency of each class

N = The total frequency.

**Steps**

- i) Obtain the mid-point of each class and denote it by m.
- ii) Multiply these mid-points by the respective frequency of each class and obtain the total  $\sum fm$ .
- iii) Divide the total obtained in step (i) by the sum of the frequency, i.e., N.

**Example**

From the following data compute arithmetic mean by direct method :

Marks:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
No. of students :	5	10	25	30	20	10

*Sol. :*

**Calculation of Arithmetic Mean by Direct Method**

Marks	Mid-point (m)	No. of Students (f)	fm
0 – 10	5	5	25
10 – 20	15	10	150
20 – 30	25	25	625
30 – 40	35	30	1,050
40 – 50	45	20	900
50 – 60	55	10	550
		N = 100	Σfm = 3,300

$$\bar{X} = \frac{\sum fm}{N} = \frac{3,300}{100} = 33.$$

**ii) Short-cut Method**

When short-cut method is used, arithmetic; computed by applying the following formula :

$$\bar{X} = A + \frac{\sum fd}{N}$$

where A - assumed mean; d = deviations of mid-points from assumed mean, i.e., (m - A); N = total number of observations.

**Steps**

- Take an assumed mean.
- From the mid-point of each class deduct the assumed mean.
- Multiply the respective frequencies of each class by these deviations and obtain the total Σfd.
- Apply the formula:  $\bar{X} = A + \frac{\sum fd}{N}$

Calculate arithmetic mean by the short-cut method from the data of above example.

**Calculation of Arithmetic Mean**

Marks	Mid-points (m)	No. of Students (f)	(m – 35) d	fd
0 – 10	5	5	–30	–150
10 – 20	15	10	–20	–200
20 – 30	25	25	–10	–250
30 – 40	35–A	30	0	0
40 – 50	45	20	+10	+200
50 – 60	55	10	+20	+200
		N = 100		Σ fd = – 200



$$\bar{X} = A + \frac{\sum fd}{N}$$

$$\begin{aligned}\bar{X} &= 35 + \frac{-200}{100} \\ &= 35 + (-2) = 33\end{aligned}$$

### PROBLEMS

1. Calculate the arithmetic mean of the monthly incomes of the families in a certain locality in Delhi.

Family	A	B	C	D	E	F	G	H	I	J
Income (₹)	85	70	10	75	500	8	42	250	40	36

*Sol:*

$$\begin{aligned}\text{Arithmetic Mean } (\bar{X}) &= \frac{\sum X}{N} = \frac{85 + 70 + 10 + 75 + 500 + 8 + 42 + 250 + 40 + 36}{10} \\ &= 1116/10 = 111.6\end{aligned}$$

∴ The arithmetic mean of the monthly incomes in the locality is Rs. 111.6.

2. Calculate arithmetic mean of the following data by direct, short-cut methods :

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks (Out of 100)	50	65	37	29	92	43	81	36	52	45

*Sol:*

Calculation of Arithmetic mean by

- i) **Direct Method**

$$\text{Arithmetic mean } (\bar{X}) = \frac{\sum X}{N} = \frac{50 + 65 + 37 + 29 + 92 + 43 + 81 + 36 + 52 + 45}{10} = \frac{530}{10} = 53$$

∴ Average marks of the class = 53 marks.

**Short Cut Method**

X	X - A = d
50	- 42
65	- 27
37	- 55
29	- 63
<b>92 A</b>	0
43	- 49
81	- 11
36	- 56
52	- 40
45	- 47
	- 390

$$\begin{aligned}\bar{X} &= A + \frac{\sum d}{N} = 92 + \frac{-390}{10} \\ &= 92 + (-39) \\ &= 53.\end{aligned}$$

3. Calculate Arithmetic mean from the following data.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
No. of Students	33	53	108	221	153	322	439	526	495	50

*Sol :*

Calculation of Arithmetic mean :

Marks	No. of students (f)	Midvalue (M)	fM
0-10	33	5	165
10-20	53	15	795
20-30	108	25	2700
30-40	221	35	7735
40-50	153	45	6885
50-60	322	55	17,710
60-70	439	65	28,535
70-80	526	75	39,450
80-90	495	85	42,075
90-100	50	95	4750
	$\Sigma f = 2400$		$\Sigma fM = 1,50,800$

$$\text{Arithmetic mean} = \frac{\Sigma fM}{\Sigma f} = \frac{1,50,800}{2400} = 62.83$$

4. Find the missing frequency from the following data :

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students	5	15	20	-	20	10

The arithmetic mean is 34 marks.

*Sol :*

Let the missing frequency be " $x_1$ "

## Calculation of Missing Frequency

Class (x)	Frequency (f)	Mid Values (m)	fm
0 – 10	5	5	25
10 – 20	15	15	225
20 – 30	20	25	500
30 – 40	$x_1$	35	$35x_1$
40 – 50	20	45	900
50 – 60	10	55	550
	<b><math>N = 70 + x_1</math></b>		<b><math>\Sigma fm = 2200 + 35x_1</math></b>

Given, Mean ( $\bar{X}$ ) = 34

$$\bar{X} = \frac{\Sigma fm}{N}$$

$$34 = \frac{2200 + 35x_1}{70 + x_1}$$

$$34(70 + x_1) = 2200 + 35x_1$$

$$2380 + 34x_1 = 2200 + 35x_1$$

$$2380 - 2200 = 35x_1 - 34x_1$$

$$\therefore x_1 = 180$$

$\therefore$  Missing frequency  $x_1$  would be 180.

5. The mean of the following series is 115.86. Find the missing figures.

House Rent (₹)	110	112	113	117	9	125	128	130
Number of Houses	25	17	13	15	14	8	6	2

Sol:

Let missing figure be ' $x_1$ '

## Calculation of Missing Figure

House Rent (₹) (x)	No. of Houses	fx
110	25	2,750
112	17	1,904
113	13	1,469
117	15	1,755
$x_1$	14	$14x_1$
125	8	1,000
128	6	768
130	2	260
	<b><math>N = 100</math></b>	<b><math>\Sigma fx = 9,906 + 14x_1</math></b>

Given,

$$\text{Mean } (\bar{x}) = 115.86$$

$$\bar{x} = \frac{\sum fx}{N}$$

$$115.86 = \frac{9,906 + 14x_1}{100}$$

Cross multiply the above information,

$$115.86 \times 100 = 9,906 + 14x_1$$

$$11,586 = 9,906 + 14x_1$$

$$-14x_1 = 9,906 - 11,586$$

$$-14x_1 = -1,680$$

$$x_1 = \frac{-1,680}{-14}$$

$$x_1 = 120$$

$\therefore$  Missing figure  $x_1$  is 120.

### 1.3.2 Mode

**Q15. Define mode.**

*Ans :*

#### Meaning

Mode may be defined as the value that occurs most frequently in a statistical distribution or it is defined as that exact value in the ungrouped data if each sample which occurs most frequently.

#### Definitions

- (i) **According to Croxton and Cowden**, "The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded as the most typical of a series of values."
- (ii) **According to A.M. Tuttle**, "Mode is the value which has the greatest frequency density in its immediate neighbourhood."
- (iii) **According to Zizek**, "The mode is the value occurring most frequently in a series of items and around which the other items are distributed most densely."

Every distribution cannot have a unique value of Mode. It can have two or even more than two modal values. Such distributions are known as Uni-Modal, Bimodal and Multi Modal.

Its graphical representation is given below.

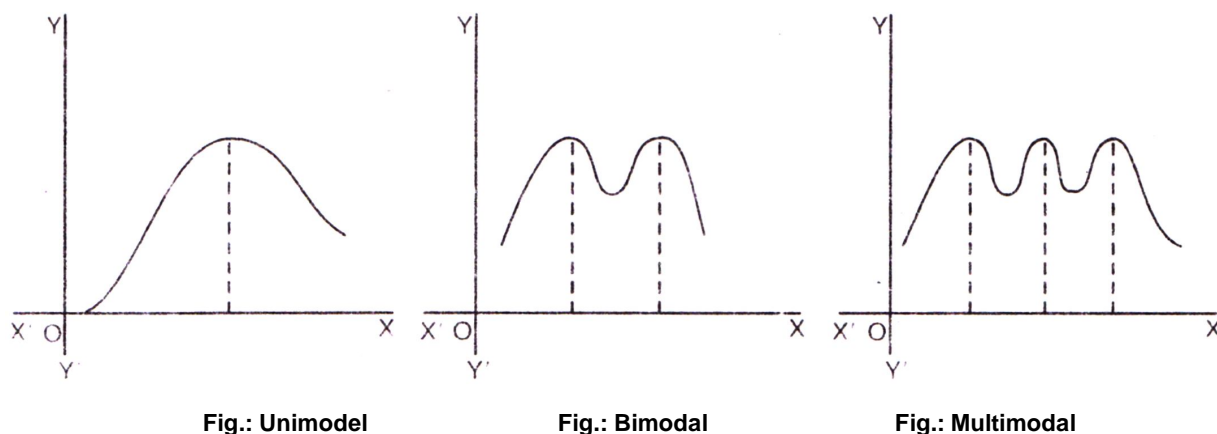


Fig.: Unimodal

Fig.: Bimodal

Fig.: Multimodal

**Q16. Explain the merits and demerits of Mode.**

*Ans :*

#### Merits

1. Mode is easy to understand and calculate.
2. It is not influenced much by items on the extremes.
3. It can be located even if the class-intervals are of unequal magnitudes, provided the modal class and the preceding and succeeding it are of the same magnitude.
4. It can be computed for distributions which have open end classes.
5. Mode is not an isolated value like the median. It is the term that occurs most in the series.
6. Mode is not a fictional value that is not found in the series.
7. It can be determined by inspection.
8. It can be located graphically
9. It has wide business application

#### Demerits

1. Calculation of Mode does not consider all the items of the series. Thus, it is not fully representative of the entire data.
2. It is not rigidly defined.
3. It is not capable of further mathematical treatment.
4. Mode is sometimes indeterminate. There may be 2 (Bi-modal) or more (Multimodal) values.
5. Mode is significantly impacted by fluctuations of sampling. Hence, it is less reliable.
6. Mode is considerably influenced by the choice of grouping. A change in the size of the class interval will change the value of the mode. It is a very unstable average and its true value is difficult to determine.

**Q17. How mode is calculated for individual and discrete series.**

*Ans :*

**Determination of Mode in Individual Series**

The steps involved in calculating mode for individual series are as follows,

1. First arrange the data in ascending or descending order.
2. Check which value is repeating maximum number of times. The value repeating maximum number of times is considered as the value of mode.

**Determination of Mode in Discrete Series**

In discrete series the value of mode can be determined in two ways,

- (i) Inspection
- (ii) Grouping and analysis table.

**(i) Inspection: Steps**

- (a) Maximum value or highest value is selected from frequency column
- (b) The value corresponding to highest frequency value is considered as mode.

**(ii) Grouping and Analysis Table**

If there exists an error of maximum frequency, frequency has small value preceding or succeeding it and items are highly focused on any one side then under such case grouping and analysis table are prepared.

**Grouping Table**

It consists of six columns as follows,

- In column 1-highest frequency is selected and highlighted with a circle or a tick (✓) marks.
- In column 2-frequency values are grouped in two's
- In column 3-Ignore first frequency and group the remaining in two's.
- In column 4-frequencies are grouped in three's
- In column 5-Ignore first frequency and group the remaining frequency values in three's.
- In column 6-Ignore first two frequencies and group remaining frequencies in three's

After preparing the grouping table the maximum value in each column should be marked with a circle.

**Analysis Table**

- Analysis table is prepared by taking the column number on left side and probable values of mode on right side of the table.
- The values of variable corresponding to the highest frequencies are taken at the top of the analysis table and probable values of mode are determined.
- Finally, the maximum marked variable is considered as 'Mode'.

**Q18. How mode is calculated for continuous series.**

*Ans :*

The steps involving in calculating mode in continuous series are as follows,

1. In continuous series the class of model class can be determined in two ways,
  - (i) By Inspection (or)
  - (ii) By preparing grouping and analysis table.
2. Value of mode is ascertained by using the following formula,

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

Where,

L = Lower limit of model class

$f_1$  = Frequency of model class

$\Delta_1 = f - f_1$

$\Delta_2 = f - f_2$

C = Class interval.

**6. Compute the modal value for the following observations and give your reasons.**

10, 12, 13, 10, 18, 16, 15, 10, 11, 17, 10, 16, 11, 10, 12, 18, 19, 20, 15, 9

*Sol :*

Arrange the data in ascending order,

Observations	No.of times repeated (or) appeared
9	1
10	5
11	2
12	2
13	1
15	2
16	2
17	1
18	2
19	1
20	1

Among of all the above number 10 is repeated at maximum i.e. 5 times than the other numbers.  
Hence mode = 10.

7. Find the mode of the following data.

0, 1, 6, 7, 2, 3, 7, 6, 6, 2, 6, 0, 5, 6, 0

*Sol:*

Observations	No. of times repeated
0	3
1	1
2	2
3	1
5	1
6	5
7	2

Among of all the above number 6 is repeated at maximum times i.e., 5 times.

Hence mode = 6.

8. Data on monthly income of 70 persons are given below. Calculate value of mode.

Income	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
Persons	8	14	19	17	12

*Sol:*

Mode can be calculated using the given formula,

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

Where

$$\Delta_1 = f - f_1$$

$$\Delta_2 = f - f_2$$

From the above distribution, 19 is the highest frequency and hence 20 – 30 is the modal class.

Income	Persons
0 – 10	8
10 – 20	14
20 – 30	19
30 – 40	17
40 – 50	12

L

( $f_1$ )

( $f$ )

( $f_2$ )

Here,

$$L = 20 \quad f = 19 \quad f_1 = 14 \quad f_2 = 17$$

$$\Delta_1 = f - f_1$$

$$19 - 14 = 5$$



$$\Delta_2 = f - f_2$$

$$19 - 17 = 2$$

$$\text{Mode} = 20 + \frac{5}{5+2} \times 10$$

$$= 20 + \frac{5}{7} \times 10$$

$$= 20 + 7.142 = 27.142$$

∴ The calculated value of mode is 27.142.

9. Calculate mode from the following data :

Marks No. of	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
Frequency	5	15	20	20	32	14	14

Sol :

Marks	Frequency
0 - 10	5
10 - 20	15
20 - 30	20
30 - 40	20 $f_1$
40 - 50	32 $f$
50 - 60	14 $f_2$
60 - 70	14

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

$$\Delta_1 = f - f_1$$

$$\Delta_2 = f - f_2$$

$$\Delta_1 = 32 - 20 = 12$$

$$\Delta_2 = 32 - 14$$

$$= 18$$

$$40 + \frac{12}{12+18} \times 10$$

$$40 + \frac{120}{30}$$

$$40 + 4 = 44$$

### 1.3.3 Median

**Q19. Define median. What are the characteristics of median?**

*Ans :*

#### Meaning

If a group of N observations is arranged in ascending or descending order of magnitude, then the middle value is called median of these observations and is denoted by M.

That is,  $M = \frac{N+1}{2}$  th observation.

#### Definition

**According to Croxton and Cowden,** "The median is that value which divides a series so that one half or more of the items are equal to or less than it and one half or more of the items are equal to or greater than it."

#### Characteristics

- Unlike the arithmetic mean, the median can be computed from open-ended distributions. This is because it is located in the median class-interval, which would not be an open-ended class.
- The median can also be determined graphically whereas the arithmetic mean cannot be ascertained in this manner.
- As it is not influenced by the extreme values, it is preferred in case of a distribution having extreme values.
- In case of the qualitative data where the items are not counted or measured but are scored or ranked, it is the most appropriate measure of location.

---

**Q20. What are the advantages and disadvantages of median?**

*Ans :*

#### Advantages

- (i) The median, unlike the mean, is unaffected by the extreme values of the variable.
- (ii) It is easy to calculate and simple to understand, particularly in a series of individual observations and a discrete series. .
- (iii) It is capable of further algebraic treatment. It is used in calculating mean deviation.
- (iv) It can be located by inspection, after arranging the data in order of magnitude.
- (v) It can be determined graphically.
- (vi) Median can be calculated in case of open-end classes.
- (vii) Median is defined rigidly.

#### Disadvantages

- (i) For calculation, it is necessary to arrange the data, other averages do not need any such arrangement.
- (ii) It is amenable to algebraic treatment in a limited sense. Median cannot be used to calculate the combined median of two or more groups, like mean.
- (iii) Median is affected more by sampling fluctuations than the mean.
- (iv) It is not based on all observations. So it is a positional average.

**Q21. How median is calculated for individual series ?***Ans :***Steps**

- (i) Arrange the data in ascending or descending order of magnitude. (Both arrangements would give the same answer.)
- (ii) In a group composed of an **odd** number of values such as 7, add 1 to the total number of values and divide by 2. Thus,  $7 + 1$  would be 8 which divided by 2 gives 4 the number of the value starting at either end of the numerically arranged groups will be the median value.

$$\text{Median} = \text{Size of } \frac{N+1}{2} \text{th item}$$

- (iii) In a group composed an even number of values, then median =  $\frac{N}{2} + 1^{\text{th}}$  item

(or)

$$\frac{\text{Two middle values}}{2}$$

**Q22. How median is calculated for discrete series?***Ans :***Steps**

- (i) Arrange the data in ascending or descending order of magnitude,
- (ii) Find out the cumulative frequencies.
- (iii) Apply the formula: Median = Size of  $\left(\frac{N+1}{2}\right)^{\text{th}}$  item
- (iv) Now look at the cumulative frequency column and find that total which is either equal to Size of  $\frac{N+1}{2}$  or next higher to that and determine the value of the variable corresponding to it. That gives the value of median.

**Q23. How median is calculated for continuous series ?***Ans :*

1. Arrange the given data in ascending order
2. Calculate cumulative frequency
3. Find  $N_1$  as  $N_1 = N/2$  for Median
4. In C.f find the value equal to or just greater than  $N_1$
5. Find  $f$  and  $L$  just one step below the C.f

$$\text{Median} = L + \frac{\frac{N}{2} - F}{f} \times C$$

L = Lower limit of the median class,

F = Cumulative frequency of the class preceding the median class (or) sum of the frequencies of all classes lower than the median class.

f = Simple frequency of the median class

C = The class interval of the median class.

### **PROBLEMS ON MEDIAN**

**10. Calculate Median from the following.**

61 62 63 61 63 64 64 60 65 63 64 65 66 64

*Sol:*

**Procedure**

**Step 1**

Arrange the data in ascending order for the given i.e.,

60, 61, 61, 62, 63, 63, 63, 64, 64, 64, 65, 65, 65, 66

**Step 2**

Apply the formulae for :

$$\begin{aligned} \text{Median} &= \text{Size of } \frac{N+1}{2} \text{th item} \\ &= \frac{14+1}{2} = \text{Size of 7.5th item} \end{aligned}$$

when 7.5 is equalled to 7th and 8th items of the data. Hence 7<sup>th</sup> = 63, 8th item = 64

$$\text{median} = \frac{63+64}{2} = 63.5$$

**11. Calculate median from the following.**

22, 26, 14, 30, 18, 11, 35, 41, 12, 32

*Sol:*

**Step I :** Arrange the data in ascending order i.e.,

11, 12, 14, 18, 22, 26, 30, 32, 35, 41

**Step II :** Calculation of given ascending to formulae

$$\begin{aligned} \text{Median} &= \text{Size of } \frac{N+1}{2} \text{th item} \\ &= \frac{10+1}{2} = \frac{11}{2} = \text{Size of 5.5th item} \end{aligned}$$

$$\text{Median} = \frac{5\text{th item} + 6\text{th item}}{2} = \frac{22+26}{2} = \frac{48}{2} = 24.$$

12. From the following data, find the value of median.

Income (₹)	200	250	130	270	300	230
No. of Persons	34	36	26	30	16	40

*Sol :*

**Procedure :**

**Step I :** Arrange the data in ascending order.

Income (Rs.) (x)	No. of Persons (F)	Cumulative Frequency (CF)
130	26	26
200	34	60
230	40	100
250	36	136
270	30	166
300	16	182
	N = 182	

**Step II:** Apply the formulae for determination of media.

$$\text{Median} = \text{Size of } \frac{N+1}{2} \text{th item}$$

$$\text{Median} = \text{Size of } \frac{182+1}{2} \text{th item}$$

$$\begin{aligned} \text{Median} &= \text{Size of } \frac{183}{2} \\ &= 91.5 \text{ item} \end{aligned}$$

Median Size of 91.5 is representing in 100 at cumulative frequency which is representing in corresponding (x) column is 230 in income level.

$\therefore$  Median = 230.

13. From the following data, compute Median

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	15	100	170	120	40

*Sol :*

Marks	Frequency	Cumulative Frequency (CF)
0 - 10	15	15
10 - 20	100	115 F
L 20 - 30	170 f	285
30 - 40	120	405
40 - 50	40	445

$$\text{Median} = L + \frac{\frac{N}{2} - F}{f} \times C$$

Median class interval = Size of  $\frac{N}{2}$ th item

Median class interval = Size of  $= \frac{445}{2} = 222.50$  item

$\therefore$  222.50th item consisting in CF 285 which is representing in corresponding class interval is 20-30.

$$\text{Median} = 20 + \left[ \frac{\frac{N}{2} - F}{f} \right] \times C = 20 + \left( \frac{222.50 - 115}{170} \right) \times 10$$

$$\text{Median} = 20 + 0.632 \times 10$$

$$\text{Median} = 20 + 6.32$$

$$\text{Median} = 26.32$$

#### 1.3.4 Relationship among Mean, Median and Mode

**Q24. Describe the relationship between Mean, Median and Mode.**

*Ans :*

Mode touches the peak of the curve indicating maximum frequency. Median divides the area of the curve in two equal halves and Mean is the centre of gravity. The three values are inter-related. The following points list the relationship between the various averages.

- In a distribution, the relative positions of mean, median, and the mode depend upon the Skewness of the distribution. If the distribution is symmetrical, then Mean = Median = Mode.
- If the distribution is positively skewed (the longer tail of the distribution is towards the right), the Mode will be lesser, than the Median, which in turn will be lower than the Arithmetic Mean. In case of a negatively skewed distribution, the Mode will be greater than the Median, which in turn will be greater than the Arithmetic Mean.
- In a given distribution, if all the observations are positive, Arithmetic Mean is greater than Geometric Mean, which in turn is greater than Harmonic Mean. With a distribution of moderate Skewness, median tends to be approximately  $1/3^{\text{rd}}$  as far away from the mean as from the mode.
- Mode = 3 Median - 2 Mean
- Mean - Mode = 3 (Mean - Median). Thus, the difference between Mean and Mode is three times, the difference between mean and median.

14. If the Mode and Mean of a moderately skewed series are 30.2 and 20.4 respectively, what would be its Median?

*Sol:*

We know that  $3\text{Median} - 2\text{Mean} = \text{Mode}$

Given mode = 30.2 & Mean = 20.4

Median = ?

$3\text{Median} = \text{Mode} + 2\text{Mean}$

$3\text{Median} = 30.2 + 2(20.4)$

$3\text{Median} = 30.2 + 40.8$

$3\text{Median} = 71.0$

$$\text{Median} = \frac{71}{3} = 23.6$$

15. If mode = 16, mean = 15.6 find the median.

*Sol:*

Given that,

Mean = 15.6 and Mode = 16

Mode =  $3\text{Median} - 2\text{Mean}$

$16 = 3\text{Median} - (2 \times 15.6)$

$16 + 31.2 = 3\text{Median}$

$47.2 = 3\text{Median}$

$$\text{Median} = \frac{47.2}{3}$$

Median = 15.733

16. For the following data calculate mean, median and mode and comment on the same.

Class Interval	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
Frequency	7	14	18	20	11	10

*Sol:*

(July.-18)

Class Interval	Frequency (f)	Midpoint (m)	fm
10 - 20	7	15	105
20 - 30	14	25	350
30 - 40	18	35	630
40 - 50	20	45	900
50 - 60	11	55	605
60 - 70	10	65	650
	80		3240

$$\bar{X} = \frac{\Sigma fM}{N} = \frac{3240}{80} = 40.5$$

**Calculation of Median**

Class Interval	Frequency (f)	Cumulative Frequency (CF)
10 – 20	7	7
20 – 30	14	21
30 – 40	18	39 F
L 40 – 50	20 f	59
50 – 60	11	70
60 – 70	10	80
	80	

$$L + \frac{\frac{N}{2} - F}{f} \times C$$

$$40 + \frac{\frac{80}{2} - 39}{20} \times 20$$

$$40 + \frac{10}{20}$$

$$40 + 0.5 = 40.5$$

**Calculation of Mode**

Morts	Frequency
10 – 20	7
20 – 30	14
30 – 40	18 $f_1$
L 40 – 50	20 f
50 – 60	11 $f_2$
60 – 70	10

$$L = \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

$$\Delta_1 = f - f_1$$



$$20 - 18 = 2$$

$$D_2 = f - f_2$$

$$20 - 11 = 9$$

$$\frac{40 + 2 \times 10}{2 + 9} = \frac{40 + 20}{11}$$

$$= 40 + 1.82 = 41.82$$

17. Find the mean and median of the following distribution.

Wages (Rs.)	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of workers	3	5	20	10	5

Sol.:

(Aug.-17)

Class Interval of wages	No. of workers (f)	Mid Value (m)	Deviation f in assumed mean m-45	Step deviation $d' = \frac{m-45}{10}$	fd'
20 – 30	3	25	-20	-2	-6
30 – 40	5	35	-10	-1	-5
40 – 50	20	45	0	0	0
50 – 60	10	55	10	+1	10
60 – 70	5	65	20	+2	10
	N = 43				$\Sigma fd' = 9$

$$\text{Mean } \bar{X} = A + \left( \frac{\Sigma fd'}{N} \right) \times C$$

$$= 45 + \left( \frac{9}{43} \right) \times 10$$

$$= 45 + \frac{90}{43}$$

$$= 45 + 2.09$$

$$= 47.09$$

Calculation of Median

Wages	No. of workers (f)	Cumulative frequency (C.f)
20 – 30	3	3
30 – 40	5	8 F
L 40 – 50	20 f	28
50 – 60	10	38
60 – 70	5	43

Median is the value of  $\frac{1}{2}$  (43) or 21.5 i.e., which lies in the 40 – 50 group. This is a case of inclusive class interval. They should be made exclusive and the median be deemed to lie in 39.5 – 49.5 group. Thus the value of  $L_1$  should be 39.5 and hof 28.

$$\begin{aligned}
 &= L + \frac{\frac{N}{2} - F}{f} \times C \\
 &= 39.5 + \frac{21.5 - 8}{20} \times 10 \\
 &= 39.5 + \frac{13.5 \times 10}{20} \\
 &= 39.5 + \frac{135}{20} \\
 &= 39.5 + 6.75 \\
 &= 46.25
 \end{aligned}$$

#### 1.4 DISPERSION

**Q25. Define Dispersion. What are the objectives of measuring dispersion?**

*Ans :*

(Dec.-20)

##### Meaning

The concept of dispersion is related to the extent of scatter or variability in observations. The variability, in an observation, is often measured as its deviation from a central value. A suitable average of all such deviations is called the measure of dispersion. Since most of the measures of dispersion are based on the average of deviations of observations from an average, they are also known as the averages of second order.

##### Definitions

As opposed to this, the measures of central tendency are known as the averages of first order. Some important definitions of dispersion are given below:

- (i) **According to A.L. Bowley**, "Dispersion is the measure of variation of the items."
- (ii) **According to Connor**, "Dispersion is the measure of extent to which individual items vary."
- (iii) **According to Simpson and Kafka**, "The measure of the scatteredness of the mass of figures in a series about an average is called the measure of variation or dispersion."
- (iv) **According to Spiegel**, "The degree to which numerical data tend to spread about an average value is called variation or dispersion of the data."

##### Objectives

The main objectives of measuring dispersion of a distribution are:

1. **To test reliability of an average:** A measure of dispersion can be used to test the reliability of an average. A low value of dispersion implies that there is greater degree of homogeneity among various items and, consequently, their average can be taken as more reliable or representative of the distribution.
2. **To compare the extent of variability in two or more distributions:** The extent of variability in two or more distributions can be compared by computing their respective dispersions. A distribution having lower value of dispersion is said to be more uniform or consistent.
3. **To facilitate the computations of other statistical measures:** Measures of dispersions are used in computations of various important statistical measures like correlation, regression, test statistics, confidence intervals, control limits, etc.
4. **To serve as the basis for control of variations:** The main objective of computing a measure of dispersion is to know whether the given observations are uniform or not.

---

**Q26. Explain the Significance of Measuring Dispersion.**

(OR)

**Explain the purpose of Dispersion.**

*Ans :*

Measures of variation are needed for four basic purposes :

1. To determine the reliability of an average.
2. To serve as a basis for the control of the variability.
3. To compare two or more series with regard to their variability.
4. To facilitate the use of other statistical measures.

---

**Q27. What are the Characteristics of a good Measures of Dispersion.**

*Ans :*

- (i) It should be easy to calculate.
- (ii) It should be easy to understand.
- (iii) It should be rigidly defined.
- (iv) It should be based on all the observations.
- (v) It should be capable of further mathematical treatment.
- (vi) It should not be unduly affected by extreme observations.
- (vii) It should not be much affected by the fluctuations of sampling.

---

#### **1.4.1 Absolute and Relative Measures**

**Q28. Explain the types of Measures of Dispersion.**

*Ans :*

There are different measures of dispersion. These measures can be classified into:

- i) Absolute measures and
- ii) Relative measures.

- i) An 'Absolute' measure is one that is expressed in terms of the same unit in which the variable (or given data) is measured.
- ii) A 'Relative' measure of dispersion is expressed as a pure number (without any units) which enables comparison of the levels of dispersion from a central tendency across different series (stated in different units). These measures are also called as "Co-efficient(s) of Dispersion". The important methods of studying variation are listed as under:

#### Absolute measures

- (i) The Range
- (ii) Inter Quartile Range and Quartile Deviation.
- (iii) The Mean Deviation (or) Average Deviation.
- (iv) The Standard Deviation and Variance
- (v) The Lorenz Curve

#### Relative measures

- (i) Coefficient of Range
- (ii) Coefficient of Quartile Deviation.
- (iii) Coefficient of Mean Deviation.
- (iv) Coefficient of Variation

#### 1.4.1.1 Range

**Q29. What is Range? Explain.**

*Ans :*

Range is the simplest method of studying dispersion. It is defined as the difference between the value of the smallest item and the value of the largest item included in the distribution. Symbolically,

$$\text{Range} = L - S$$

where

L = Largest item, and

S = Smallest item.

The relative measure corresponding to range, called the coefficient of range, is obtained by applying the following formula :

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

If the averages of the two distributions are about the same, a comparison of the range indicates that the distribution with the smaller range has less dispersion, and the average of that distribution is more typical of the group.

**Q30. What are the merits and limitations of Range ?**

*Ans :*

The merits and limitations of Range can be enumerated here.

**Merits**

- (i) Amongst all the methods of studying dispersion range is the simplest to understand and the easiest to compute.
  - It is one of those measures which are rigidity defined.
- (ii) It takes minimum time to calculate the value of range. Hence, if one is interested in getting a quick rather than a very accurate picture of variability one may compute range.
  - It gives us the total picture of the problem even with a single glance.
  - It is used to check the quality of a product for quality control. Range plays an important role in preparing R-charts, thus quality is maintained.

**Demerits**

1. Range is not based on all the terms. Only extreme items reflect its size. Hence range cannot be completely representative of the data as all other middle values are ignored.
2. Due to above reason range is not a reliable measure of dispersion.
3. Range does not change even the least even if all the terms and variables are changed.
4. Range is too much affected by fluctuation of sampling. Range changes from sample to sample. As the size of sample increases range increases and vice versa.
5. It does not tell us anything about the variability of other data.
6. For open-end intervals, range is indeterminate because lower and upper limits of first and last interval are not given.

- 
- 18. The following are the wages of 10 workers of a factory. Find the range of variation and also compute the coefficient of range.**

**275, 200, 370, 240, 100, 290, 400, 500, 180, 350**

*Sol.:*

**Step I**

Arrange the data in ascending order i.e.

100, 180, 200, 240, 275, 290, 350, 370, 400, 500

Range = L – S

where as = L = 500, S = 100

Range (R) = 500 – 100 = 400

$$\text{Coefficient of range} = \frac{L - S}{L + S} = \frac{500 - 100}{500 + 100}$$

$$\text{Coefficient of range} = \frac{400}{600} = 0.667.$$

19. The following are the marks of 100 students of a class. Find the range of variation of marks and also compute coefficient of range.

Marks	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of students	8	16	26	20	12	10	8

*Sol :*

From the given data

Highest (L) marks is 80 and

Lowest (S) marks is 10. Hence

Range (R) = L – S

$$(R) = 80 - 10 = 70$$

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

$$\text{Coefficient of Range} = \frac{80 - 10}{80 + 10}$$

$$\text{Coefficient of Range} = \frac{70}{90}$$

$$\text{Coefficient of Range} = 0.77.$$

#### 1.4.1.2 Quartile Deviation

##### Q31. What is Quartile Deviations?

*Ans :*

It is based on two extreme items and it fails to take account of the scatter within the range. From this there is reason to believe that if the dispersion of the extreme items is discarded, the limited range thus established might be more instructive.

For this purpose there has been developed a measure called the interquartile range, the range which includes the middle 50 per cent of the distribution. That is, one quarter of the observations at the lower end, another quarter of the observations at the upper end of the distribution are excluded in computing the interquartile range. In other words, interquartile range represents the difference between the third quartile and the first quartile.

$$\text{Inter quartile range} = Q_3 - Q_1$$

Very often the interquartile range is reduced to the form of the Semi-interquartile range or quartile deviation by dividing it by 2.

$$\text{Quartile Deviation or Q.D.} = \frac{Q_3 - Q_1}{2}$$

**Coefficients of Quartile Deviation**

Quartile deviation is an absolute measure of dispersion. Its relative measure is the coefficient of Quartile deviation.

Coefficients of Quartile Deviation

$$\begin{aligned} & \frac{Q_3 - Q_1}{2} \\ = & \frac{Q_3 + Q_1}{2} \\ & \frac{Q_3 - Q_1}{Q_3 + Q_1} \end{aligned}$$

It is used to compare the degree of variation in different series.

**Q32. Explain merits and demerits of quartile deviations.**

*Ans :*

**Merits**

1. It is the simple to calculate and very easy to understand.
2. It is not impacted by extreme values
3. It can be computed for open ended distributions and for data containing unequal classes

**Demerits**

1. It is impacted by sample size and composition of the sample.
2. It is not amenable to algebraic or statistical treatment
3. It does not tell us anything about the spread of the various data items across the measure of central tendency.

**Q33. How to calculate quartile deviations?**

*Ans :*

Low Quartile ( $Q_1$ ),

$$= \text{Size of } \frac{N+1}{4} \text{ th item}$$

N = No. of observations

Upper Quartile ( $Q_3$ ),

$$= \text{Size of } \frac{3(N+1)}{4} \text{ th item.}$$

**Discrete Series**

i)  $Q_1 = \text{Size of } \frac{N+1}{4} \text{ th item}$

ii)  $Q_3 = \text{Size of } \frac{3(N+1)}{4} \text{ th item}$

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

**Continuous Series**

$$Q_1 = L + \frac{\frac{N}{4} - F}{f} \times C$$

$$Q_3 = L + \frac{\frac{3N}{4} - F}{f} \times C$$

20. Find out the value of quartile deviation and its coefficient for the following data :

Roll No	1	2	3	4	5	6	7
Marks	25	33	45	17	35	20	55

*Sol :*

**Calculation of Quartile Deviation**

Marks arranged in ascending order 17 20 25 33 35 45 55

$$Q_1 = \text{Size of } \left[ \frac{N+1}{4} \right]^{\text{th}} \text{ Item}$$

N = No. of observations.

$$= \text{Size of } \left[ \frac{7+1}{4} \right]^{\text{th}} \text{ Item} = 2^{\text{nd}} \text{ Item} = 20$$

$$Q_3 = \text{Size of } 3 \left[ \frac{N+1}{4} \right]^{\text{th}} \text{ Item}$$

$$= \text{Size of } 3 \left[ \frac{7+1}{4} \right]^{\text{th}} \text{ Item} = 6^{\text{th}} \text{ Item} = 45$$

$$\text{Quartile Deviation (Q.D)} = \frac{Q_3 - Q_1}{2} = \frac{45 - 20}{2} = \frac{25}{2} = 12.5$$

$$\text{Coefficient of Q.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{45 - 20}{45 + 20} = \frac{25}{65} = 0.38.$$

21. Find out the value of quartile deviation from the following data

Roll No.	1	2	3	4	5	6	7
Marks :	30	42	60	18	45	24	75

*Sol :*

Rearrange into ascending order

$$x = 18, 24, 30, 42, 45, 60, 75$$



Quartile deviation  $Q_1$  = size of  $\left(\frac{N+1}{4}\right)^{\text{th}}$  item

Size of  $\left(\frac{7+1}{4}\right)^{\text{th}}$  item =  $2^{\text{nd}}$  item

The size of  $2^{\text{nd}}$  item is 24 i.e.  $Q_1 = 24$

$Q_3$  = size of  $3\left(\frac{7+1}{4}\right)^{\text{th}}$  item

=  $6^{\text{th}}$  item = 60, then  $Q_3 = 60$

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

$$= \frac{60 - 24}{2} = \frac{36}{2} = 18$$

$$Q.D = 18$$

22. Calculate the Quartile Deviation and Coefficient of Quartile Deviation of the following distribution.

Marks (X)	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
No. of Students (f)	11	18	25	28	30	33	22	15	12	10

Sol.:

#### Calculation of Quartile Deviation

X	f	C.f
0-10	11	11
10-20	18	29 F
L 20-30	25 f	54 ( $Q_1$ )
30-40	28	82
40-50	30	112
50-60	33	145 F
L 60-70	22 f	167 ( $Q_3$ )
70-80	15	182
80-90	12	194
90-100	10	204

**Calculation of  $Q_1$** 

$$\begin{aligned}
 Q_1 &= \left( \frac{N}{4} \right)^{th} \text{ Item} \\
 &= \left( \frac{204}{4} \right)^{th} \text{ Item} \\
 &= 51^{th} \text{ Item}
 \end{aligned}$$

Size of  $51^{th}$  Item lies in CF 54. So  $Q_1$  class interval = 20–30

$$\begin{aligned}
 Q_1 &= L + \frac{\frac{N}{4} - F}{f} \times C \\
 &= 20 + \frac{51 - 29}{25} \times 10 \\
 &= 20 + \frac{22 \times 10}{25} \\
 &= 20 + \frac{220}{25} \\
 &= 20 + 8.8 = 28.8.
 \end{aligned}$$

**Calculation of  $Q_3$** 

$$\begin{aligned}
 &= Q_3 = 3 \left( \frac{N}{4} \right)^{th} \\
 &= 3 \left( \frac{204}{4} \right)^{th} \text{ Item} = 3(51)^{th} \text{ Item} = 153^{th} \text{ Item}
 \end{aligned}$$

Size of  $153^{th}$  Item lies in CF 167. So  $Q_3$  class interval = 60 – 70

$$\begin{aligned}
 Q_3 &= L_3 + \frac{\frac{3N}{4} - F}{f} \times C \\
 &= 60 + \frac{153 - 145}{22} \times 10 \\
 &= 60 + \frac{80}{22} \\
 &= 60 + 3.64 \\
 &= 63.64
 \end{aligned}$$

**Calculation of Quartile Deviation**

$$\text{Quartile Deviation (Q.D)} = \frac{Q_3 - Q_1}{2} = \frac{63.64 - 28.8}{2} = 17.4$$

$$\text{Coefficient of Q.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{63.64 - 28.8}{63.64 + 28.8} = \frac{34.84}{92.44} = 0.37$$

**1.4.1.3 Mean Deviation**

**Q34. Define mean deviation. State the merits and demerits of mean deviation.**

*Ans :*

**Meaning**

"Mean Deviation of a series is the arithmetic average of the deviations of various items from a measure of central tendency (either mean, median or mode)". Theoretically, deviations can be taken from any of the three averages mentioned above, but in actual practice it is calculated either from mean or from Median. While Calculating deviations algebraic signs are not taken into account.

**Merits**

- (i) It is rigidly defined
- (ii) It is not least impacted by sampling fluctuations
- (iii) It takes into account every single value in the series
- (iv) Mean deviation from Median is least impacted due to extreme values
- (v) It is extensively used in multiple fields such as Economics, Commerce, etc as it is the best measure for comparison of two or more series.

**Demerits**

- (i) It is relatively difficult to compute
- (ii) It is not amenable to further algebraic or statistical treatment.
- (iii) It is difficult for a layman to understand as to why or when a particular average should be considered for calculation of Mean Deviation. The Mean Deviations obtained by taking the Mean, Median and Mode as average differ widely
- (iv) It is not effective for open ended series, particularly when the average is Arithmetic Mean.'

**Q35. How to calculate mean deviation for individual series ?**

*Ans :*

**Individual Series**

If  $X_1, X_2, X_3, X_N$  are N given observations then the deviation about an average. A is given by

$$\frac{\sum |D|}{N}$$

**Steps**

- i) Compute the median of the series.
- ii) The deviations of items from median ignoring  $\pm$  signs and denote these deviations by  $|D|$ .
- iii) Obtain the total of these deviations, i.e.,  $\Sigma |D|$ .
- iv) Divide the total obtained in step
- v) By the number of observations.

The relative measure corresponding to the mean deviation, called the coefficient of mean deviation, is obtained by dividing mean deviation by the particular average used in computing mean deviation. Thus, if mean deviation has been computed from median, the coefficient of mean deviation shall be obtained by dividing mean deviation by median.

$$\text{Coefficient of M.D.} = \frac{\text{M.D.}}{\text{Median}}$$

If mean has been used while calculating the value of mean deviation, in such a case coefficient of mean deviation shall be obtained by dividing mean deviation by the mean.

**Q36. How to calculate mean deviation for discrete series ?**

*Ans :*

In discrete series the formula for calculating mean deviation is

$$\text{M.D.} = \frac{\Sigma f |D|}{N}$$

$|D|$  denotes deviation from median ignoring signs.

$N$  = Sum of frequency

**Steps**

- i) Calculate the median of the series.
- ii) Take the deviations of the items from median ignoring signs and denote them by  $|D|$ .
- iii) Multiply these deviations by the respective frequencies and obtain the total  $\Sigma f |D|$
- iv) Divide the total obtained in Step (ii) by the number of observations. This gives us the value of mean deviation.

**Q37. How to calculate mean deviation for continuous series ?**

*Ans :*

For calculating mean deviation in continuous series the procedure remains the same as discussed above. The only difference is that here we have to obtain the mid-point of the various classes and take deviations of these points from median. The formula is same, i.e.,

$$\text{M.D.} = \frac{\Sigma f |D|}{N}$$

**Steps**

- i) Calculate the median of the series.
- ii) Take the deviations of the items from median ignoring signs and denote them by  $|D|$ .
- iii) Multiply these deviations by the respective frequencies and obtain the total  $\Sigma f |D|$ .
- iv) Divide the total obtained in step (ii) by the number of observations. This gives us the value of mean deviations.

23. Calculate mean deviation of the following from mean and median

2, 6, 11, 14, 16, 19, 23

Sol. :

X	$X - \bar{X} =  D $	$X - \text{Median} =  D $
2	11	12
6	7	8
11	2	3
14	1	0
16	3	2
19	6	5
23	10	9
91	40	39

(a) Mean  $(\bar{X}) = \frac{\sum X}{N} = \frac{91}{7} = 13$

MD through mean =  $\frac{\sum |D|}{N} = \frac{40}{7} = 5.71$

(b) Median = size of  $\left(\frac{N+1}{2}\right)^{\text{th}}$  item

= size of  $\left(\frac{7+1}{2}\right)$

=  $\frac{8}{2} = 4^{\text{th}}$  observation = 14

(c) Median = size of 4th Item = 14

Mean Deviation (MD) through

Median =  $\frac{\sum |D|}{M}$

Mean Deviation (MD) =  $\frac{39}{7}$

Mean Deviation (MD) = 5.57

24. From the following data calculate mean deviation from the median.

C.I.	16 - 20	21 - 25	26 - 30	31 - 35	36 - 40	41 - 45	46 - 50	51 - 55	56 - 60
Frequency	8	15	13	20	11	7	3	2	1

Sol :

Calculation of Mean Deviation from Median

C.I	Frequency (F)	C.F	Mid-Value (m)	$ M - A  =  D $ $ M - 31.5 $	$F M - A  = f D $
16 - 20	8	8	18	13.5	108.0
21 - 25	15	23	23	8.5	127.5
26 - 30	13	36 F	28	3.5	45.5
L 31 - 35	20 f	56	33	1.5	30.0
36 - 40	11	67	38	6.5	71.5
41 - 45	7	74	43	11.5	80.5
46 - 50	3	77	48	16.5	49.5
51 - 55	2	79	53	21.5	43.0
56 - 60	1	80	58	26.5	26.5
	<b>N = 80</b>				<b><math>\Sigma f D  = 582</math></b>

$$\begin{aligned}\text{Median} &= \text{size of } \frac{N^{\text{th}}}{2} \text{ item} \\ &= \frac{80}{2} = 40^{\text{th}} \text{ item}\end{aligned}$$

C.F is just greater than 40 is 56

Median lies in the class 31-35. However, the real limit of this class under exclusive method is 30.5 – 35.5.

$$\text{Median} = L + \frac{\frac{N}{2} - F}{f} \times C$$

$$L = 30.5, F = 36, f = 20 \quad C = 5$$

$$= 30.5 + \frac{40 - 36}{20} \times 5$$

$$= 30.5 + \frac{4}{20} \times 5$$

$$= 30.5 + [0.2] \times 5$$

$$= 30.5 + 1 = 31.5$$

$$\text{M.D from Median} = \frac{\Sigma f|D|}{N} = \frac{582}{80} = 7.275$$

#### 1.4.1.4 Standard Deviation

**Q38. What is standard deviation? Explain its merits and demerits.**

*Ans :*

**(Imp.)**

The standard deviation concept was introduced by Karl Pearson in 1823. It is by far the most important and widely used measure of studying dispersion. Its significance lies in the fact that it is free from those defects from which the earlier methods suffer and satisfies most of the properties of a good measure of dispersion. Standard deviation is also known as root mean square deviation for the reason that it is the square root of the mean of the squared deviation from the arithmetic mean. Standard deviation is denoted by the small Greek letter  $\sigma$  (read as sigma).

The standard deviation measures the absolute dispersion (or variability of distribution; the greater the amount of dispersion or variability), the greater the standard deviation, the greater will be the magnitude of the deviations of the values from their mean. A small standard deviation means a high degree of uniformity of the observation as well as homogeneity of a series; a large standard deviation means just the opposite.

#### Merits

1. It is rigidly defined
2. It takes into account every single value in the series
3. It is amenable to further algebraic or statistical treatment.
4. It is extensively used in various other statistical calculations such as correlation, regression, sampling etc

#### Demerits

1. It is relatively difficult to compute
2. It is calculated with only Arithmetic Mean as the average. Standard deviation from other averages such as Median is not an effective measure of dispersion.

---

**Q39. How to compute the standard deviations for individual series ?**

*Ans :*

#### Individual Series

In case of individual observations standard deviations may be calculate by applying any of the following two methods, i.e.

- i) By taking deviations of the items from the actual mean.
- ii) By taking deviations of the items from an assumed mean.

Deviation taken from actual mean : when deviations taken from actual mean the following formula is applied.

$$\sigma = \sqrt{\frac{\sum x^2}{N}}$$

where as

$$x = (X - \bar{X})$$

N = Total number of observations.

**Steps**

- i) Calculate the actual for the series i.e.  $\bar{X}$
- ii) Find  $X$  deviation from  $\bar{X}$  i.e.  $x = (X - \bar{X})$
- iii) Square these deviations and obtain the total  $\Sigma x^2$
- iv) Divide  $\Sigma X^2$  by the total number of observations and extract the square root.

**Q40. How to compute the standard deviations for discrete series?***Ans. :*

For calculating standard deviation in discrete series, any of the following methods may be applied:

- (a) Actual mean method.
- (b) Assumed mean method.
- (c) Step deviation method.

- (a) Actual Mean Method.** When this method is applied, deviations are taken from the actual mean, i.e., we find  $(X - \bar{X})$  and denote these deviations by  $x$ . These deviations are then squared and multiplied by the respective frequencies. The following formula is applied :

$$\sigma = \sqrt{\frac{\Sigma fx^2}{N}}, \text{ where } x = (X - \bar{X})$$

$N$  = Sum of observation/Frequency.

However, in practice this method is rarely used because if the actual mean is in fractions the calculations take a lot of time.

**Steps**

- i) Take the deviations of the items from an assumed mean and denote these deviations by  $d$ .
- ii) Multiply these deviations by the respective frequencies and obtain the total,  $\Sigma fd$ .
- iii) Obtain the squares of the deviations, i.e., calculate  $d^2$ .
- iv) Multiply the squared deviations by the respective frequencies, and obtain the total,  $\Sigma fd^2$ .
- v) Substitute the values in the above formula.

- (b) Step Deviation Method:** When this method is used we take deviations of midpoints from an assumed mean and divide these deviations by the width of class interval, i.e., 'C'. In case class intervals are unequal, we divide the deviations of midpoints by the lowest common factor and use 'C' in the formula for calculating standard deviation.

The formula for calculating standard deviation is :

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

where,

$$d = \frac{(X - A)}{C} \text{ and } C = \text{class interval.}$$

The use of the above formula simplifies calculations.



**Q41. How to compute the standard deviations for continuous series ?***Ans :*

In continuous series any of the methods discussed above for discrete frequency distribution can be used. However, in practice it is the step deviation method that is most used. The formula is

$$= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times C$$

where  $d = \frac{(m - A)}{C}$ ,  $C$  = class interval

**Steps**

- Find the mid-points of various classes.
- Take the deviations of these mid-points from an assumed mean and denote these deviations by  $d$ .
- Wherever possible take a common factor and denote this column by  $d$ .
- Multiply the frequencies of each class with these deviations and obtain  $\sum fd$ .
- Square the deviations and multiply them with the respective frequencies of each class and obtain  $\sum fd^2$ .

**Q42. Define Coefficient of Variation.***Ans :*

Coefficient of Variation (CV) was proposed by Karl Pearson. It is used to compare the variability of two (or) more distributions. A distribution with greater C.V. is considered as more variable or less consistent, less unit form, less stable or less homogeneous distribution and the distribution with less C.V is considered as less variable or more consistent, more uniform, more stable or more homogeneous distribution.

$$\text{Coefficient of variation} = \frac{\text{Standard deviation}}{\text{Arithmetic mean}} \times 100$$

**25. Calculate standard deviation for the following data:**

Sl. No.	Weekend Income (₹)
1	270
2	350
3	258
4	282
5	218
6	202
7	364
8	184

*Sol :***Calculation of Standard Deviation**

Sl.No.	X	$x = [X - \bar{X}]$	$x^2$
1	270	4	16
2	350	84	7056
3	258	- 8	64
4	282	16	256
5	218	- 48	2304
6	202	- 64	4096
7	364	98	9604
8	184	- 82	6724
	<b><math>\Sigma X = 2128</math></b>		<b><math>\Sigma X^2 = 30120</math></b>

(i) Arithmetic Mean ( $\bar{X}$ ) =  $\frac{\Sigma X}{N} = \frac{2128}{8}$   
 = 266

(ii) Standard Deviation ( $\sigma$ ) =  $\sqrt{\frac{\Sigma x^2}{N}}$   
 =  $\sqrt{\frac{30120}{8}}$   
 =  $\sqrt{3765}$   
 = 61.35

26. From the following information calculate standard deviation.

X	4.5	14.5	24.5	34.5	44.5	54.5	64.5
F	1	5	12	22	17	9	4

*Sol :***Calculation of Standard Deviation**

X	f	$X - A = d$	$d^2$	fd	$fd^2$
4.5	1	-30	900	-30	900
14.5	5	-20	400	-100	2000
24.5	12	-10	100	-120	1200
34.5A	22	0	0	0	0
44.5	17	10	100	170	1700
54.5	9	20	400	180	3600
64.5	4	30	900	120	3600
	<b>70</b>			<b>220</b>	<b>13,000</b>

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

$$\sigma = \sqrt{\frac{13000}{70} - \left(\frac{220}{70}\right)^2}$$

$$\sigma = \sqrt{185.71 - (3.14)^2}$$

$$\sigma = \sqrt{185.71 - 9.8596}$$

$$\sigma = \sqrt{175.85}$$

$$\sigma = 13.26$$

27. Find out the mean and standard deviation of the following data :

Age under	10	20	30	40	50	60	70	80
No. of persons dying	15	30	53	75	100	110	115	125

Sol :

Calculation of Standard Deviation

Age	f	Cf	Midvalue (m)	m-A=d	d	d <sup>2</sup>	fd <sup>2</sup>	fd
0 - 10	15	15	5	-30	-3	9	135	-45
10 - 20	15	30	15	-20	-2	4	60	-30
20 - 30	23	53	25	-10	-1	1	23	-23
30 - 40	22	75	35 A	0	0	0	0	0
40 - 50	25	100	45	10	1	1	25	25
50 - 60	10	110	55	20	2	4	40	20
60 - 70	5	115	65	30	3	9	45	15
70 - 80	10	125	75	40	4	16	160	40
	125						488	-2

i) Mean  $\bar{X} = A + \left(\frac{\sum fd'}{N}\right) \times C$

$$\text{Mean } \bar{X} = 35 + \left(\frac{25}{125}\right) \times 10$$

$$\text{Mean } \bar{X} = 35 + (0.016) \times 10$$

$$\text{Mean } \bar{X} = 35 + 0.16 = 35.16$$

$$35 + \left(\frac{-2}{125}\right) \times 10$$

$$35 - 0.16 = 34.84$$

$$\begin{aligned}
 \text{ii) Standard deviation } (\sigma) &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times C \\
 &= \sqrt{\frac{488}{125} - \left(\frac{2}{125}\right)^2} \times 10 \\
 &= \sqrt{3.904 - (0.016)^2} \times 10 \\
 &= \sqrt{3.904 - 0.0003} \times 10 \\
 &= \sqrt{3.9037} \times 10 = 1.9757 \times 10 = 19.76.
 \end{aligned}$$

28. Calculate Standard Deviation and Coefficient of Variation from the following data:

Marks	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80
No. of Students	5	7	14	28	12	9	6	2

Sol :

Calculation of Standard Deviation

Marks	Frequency	Midvalue	d	d <sup>2</sup>	fd <sup>2</sup>	fd
0 - 10	5	5	- 3	9	45	- 15
10 - 20	7	15	- 2	4	28	- 14
20 - 30	14	25	- 1	1	14	- 14
30 - 40	28	35 A	0	0	0	0
40 - 50	12	45	1	1	12	12
50 - 60	9	55	2	4	36	18
60 - 70	6	65	3	9	54	18
70 - 80	2	75	4	16	32	8
	83				221	13

$$\text{Mean} = A + \frac{\sum fd}{N} \times C$$

$$\begin{aligned}
 &= 35 + \frac{13}{83} \times 10 \\
 &= 35 + 1.5 = 36.5
 \end{aligned}$$

$$\text{S.D} = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times C$$

$$= \sqrt{\frac{221}{83} - \left(\frac{13}{83}\right)^2} \times 10$$

$$= \sqrt{2.66 - 0.024} \times 10$$

$$= \sqrt{2.636} \times 10$$

$$= 1.623 \times 10$$

$$= 16.23$$

$$\text{Coefficient of variation} = \frac{\text{S.D}}{\text{Mean}} \times 100$$

$$= \frac{16.23}{36.5} \times 100$$

$$= 0.44 \times 100 = 44.46$$

### 1.5 SKEWNESS

**Q43. Explain briefly about Skewness.**

*Ans :*

#### Introduction

The word skewness refers to lack of symmetry. Non-normal or asymmetrical distribution is called skew distribution. Two frequency distributions may have the same mean and standard deviation and yet may differ with respect to another characteristics- the skewness or, asymmetry of the distribution. Any measure of skewness indicates the difference between the manner in which items are distributed in a particular distribution compared with a normal distribution. Lack of symmetry or skewness in frequency distributions is due to the existence of a longer tail on one side (either to the left or to the right), which has no counterpart on the other side. If the larger tail is on the right, we say that the distribution is positively skewed; whereas if the longer tail is on the left side.

#### Definitions

Some important definitions of skewness are as follows :

- **According to Corxton & Cowden** "When a series is not symmetrical it is said to be asymmetrical or skewed."
- **According to Morris Hamburg** "Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution."
- **According to Simpson & Kafka** "Measures of skewness tell us the direction and the extent of skewness. In symmetrical distribution the mean, median and mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness."
- **According to Garrett** "A distribution is said to be 'skewed' when the mean and the median fall at different points in the distribution, and the balance (or centre of gravity) is shifted to one side or the other – to left or right."

#### Types

##### 1. Positively Skewed Distribution

If the longer tail of the distribution is towards the higher values (or) right hand side, the skewness is positive. Positive skewness occurs when mean is increased by some unusually high values thus satisfying the following properties,

- (i) Mean > Median > Mode
- (ii) Right tail is longer than its left tail
- (iii)  $(Q_3 - \text{Median}) > (\text{Median} - Q_1)$ .

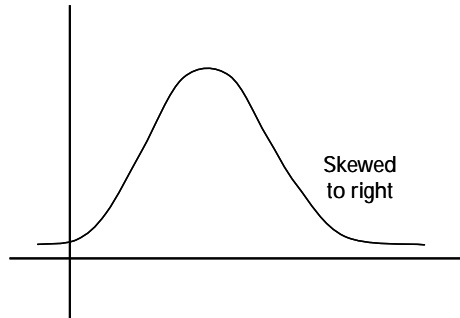


Fig.: Positively Skewed Distribution

## 2. Negatively Skewed Distribution

If the longer tail is towards the lower value or left hand side, the skewness is negative.

Negative skewness arises when mean is decreased by some extremely low values, thus satisfying the following properties,

- (i) Mean < Median < Mode
- (ii) Left tail is longer than its right tail
- (iii)  $(Q_3 - \text{Median}) < (\text{Median} - Q_1)$

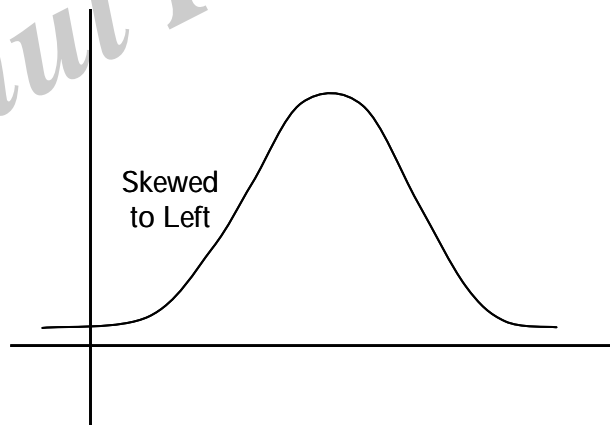


Fig.: Negatively Skewed Distribution

### 1.5.1 Karl Pearson's Coefficient of Skewness - Bowley's Coefficient of Skewness - Kelly's Measure of Skewness

**Q44. Explain the different measures of skewness.**

*Ans :*

Skewness can be measured absolutely and relatively. Absolute measures are also known as measures of skewness whereas relative measures are termed as the coefficients of skewness.

## 1. Absolute Measures of Skewness

In a skewed distribution the three measures of central tendency differ. Accordingly skewness may be worked out in absolute amount with the help of the following formulae :

$$\text{Absolute Skewness } S_K = \bar{X} - \text{Mode}$$

$$\text{Absolute Skewness } S_K = \bar{X} - \text{Median}$$

$$\text{Absolute Skewness } S_K = \text{Median} - \text{Mode}$$

## 2. Relative Measures of Skewness

The following are the four important measures of relative skewness, termed as coefficients of skewness :

- (i) The Karl Pearson's Coefficient of Skewness.
- (ii) The Bowley's Coefficient of Skewness.
- (ii) The Kelly's Coefficient of Skewness.

The results obtained by these formulae will generally lie between +1 and -1. When the distribution is positively skewed, the coefficient of skewness will have plus sign and when it is negatively skewed it will have negative sign. It should be remembered that the value of the coefficient will never exceed 1.

### (i) Karl Pearson's Coefficient of Skewness

This method of measuring skewness, also known as Pearsonian Coefficient of Skewness, was suggested by Karl Pearson, a great British Biometrician and Statistician. It is based upon the difference between mean and mode. This difference is divided by standard deviation to give a relative measure. The formula thus becomes :

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

$Sk_p$  = Karl Pearson's coefficient of skewness

When Mode is ill-defined

Coefficient of Skewness

$$(S_{K_p}) = \frac{3 (\text{Mean} - \text{Median})}{\sigma}$$

There is no limit to this measure in theory and this is a slight drawback. But in practice the value given by this formula is rarely very high and usually lies between  $\pm 1$ .

### (ii) Bowley's Coefficient of Skewness

It is based on quartiles ( $Q_3$  and  $Q_1$ ).

Bowley's Coeff. of Sk.

$$S_{K_B} = \frac{(Q_3 - M) - (M - Q_1)}{(Q_3 - M) + (M - Q_1)}$$

$$S_{K_B} = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

where, M = Median

This measure is called the quartile measure of skewness and values of the coefficient, thus obtained vary between  $\pm 1$ .

### (iii) Kelly's Coefficient of Skewness

Bowley's measure discussed above neglects the two extreme quarters of the data. It would be better for a measure to cover the entire data especially because in measuring skewness, we are often interested in the more extreme items. Bowley's measure can be extended by taking any two deciles equidistant from the median or any two percentiles equidistant from the median. Kelly has suggested the following formula for measuring skewness upon the 10<sup>th</sup> and the 90<sup>th</sup> percentiles (or the first and ninth deciles) :

$$Sk_K = \frac{P_{10} + P_{90} - 2 \text{ Med.}}{P_{90} - P_{10}} \text{ also}$$

$$Sk = \frac{D_1 + D_9 - 2 \text{ Med.}}{D_9 - D_1}$$

$Sk_K$  = Kelly's coefficient of skewness.

This measure of skewness has one theoretical attraction if skewness is to be based on percentiles. However, this method is not popular in practice and generally Karl Pearson's method is used.

### 29. Calculate Karl Pearson's Coefficient of Skewness and coefficient of variation from the following data.

**Mode = 33.5, Mean = 30.08,**

**Standard deviation = 13.405**

*Sol :*

Co-efficient of Skewness

$$= \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{30.08 - 33.5}{13.405}$$

$$= - 0.255$$

Co-efficient of Variation

$$= \frac{\sigma}{\bar{X}} \times 100$$

$$= \frac{13.405}{30.08} \times 100 = 44.56 .$$



30. From the following data calculate Karl Pearson's Coefficient of Skewness.

Marks	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
No. of Students	12	28	40	60	32	18	10

Sol :

Marks	No. of Students (f)	Mid (value) (X)	$d = \frac{X - A}{10}$	$d^2$	$fd$	$fd^2$
0 - 10	12	5	- 3	9	- 36	108
10 - 20	28	15	- 2	4	- 56	112
20 - 30	$f_1$ 40	25	- 1	1	- 40	40
L 30 - 40	f 60 A	35 A	0	0	0	0
40 - 50	$f_2$ 32	45	1	1	32	32
50 - 60	18	55	2	4	36	72
60 - 70	10	65	3	9	30	90
	200				- 34	454

$$\text{Karl Pearsons Coefficient of Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{S.D}}$$

$$\text{Mean} = A + \frac{\sum fd}{N} \times C$$

$$= 35 + \frac{(-34)}{200} \times 10$$

$$= 35 - \frac{340}{200}$$

$$= 35 - 1.7 = 33.3$$

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

$$\Delta_1 = f - f_1$$

$$= 60 - 40 = 20$$

$$\Delta_2 = f - f_2$$

$$= 60 - 32 = 28$$

$$= 30 + \frac{20}{20 + 28} \times 10$$

$$= 30 + \frac{200}{48}$$

$$= 30 + 4.1 = 34.1$$

$$\begin{aligned}
 S.D &= \sqrt{\left(\frac{\sum fd^2}{N}\right) - \left(\frac{\sum fd}{N}\right)^2} \times C \\
 &= \sqrt{\frac{454}{200} - \left(\frac{-34}{200}\right)^2} \times 10 = \sqrt{2.27 + 0.0289} \times 10 \\
 &= \sqrt{2.2989} \times 10 = 1.516 \times 10 = 15.16 \\
 S_{KP} &= \frac{\text{Mean} - \text{Mode}}{S.D} = \frac{33.3 - 34.1}{15.16} = \frac{-0.18}{15.16} \\
 &= -0.0531
 \end{aligned}$$

**31. Calculate the coefficient of skewness based on quartiles**

Variable	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
Frequency	5	9	14	20	25	15	8	4

*Sol:*

	Variable	Frequency	Cumulative frequency
	10-20	5	5
	20-30	9	14F
L	30-40	14 f	28 (Q1)
	40-50	20	48F
L	50-60	25 f	73F (Median)
L	60-70	15 f	88 (Q3)
	70-80	8	96
	80-90	4	100 = N

$$\text{Coefficient of skewness} = \frac{Q_3 + Q_1 - 2\text{median}}{Q_3 - Q_1}$$

**Calculation of Median**

$$\text{Median} = L + \frac{\frac{N}{2} - F}{f} \times C = \frac{N}{2} = \frac{100}{2} = 50^{\text{th}} \text{ observation}$$

Median lies in the class of 50 – 60

$$\begin{aligned}
 &= 50 + \frac{50 - 48}{25} \times 10 \\
 &= 50 + \frac{2 \times 10}{25} \\
 &= 50 + \frac{20}{25} = 50 + 0.8 = 50.8
 \end{aligned}$$

Calculation of  $Q_3$  and  $Q_1$

$$Q_3 = L + \frac{\frac{3N}{4} - F}{f} \times C$$

$$\frac{3N}{4} = \frac{3 \times 100}{4} = \frac{300}{4} = 75^{\text{th}} \text{ observation}$$

$Q_3$  lies in the class of 60 – 70

$$\begin{aligned}
 &= 60 + \frac{75 - 73}{15} \times 10 = 60 + \frac{2 \times 10}{15} \\
 &= 60 + \frac{20}{15} = 60 + 1.33 = 61.33
 \end{aligned}$$

$$Q_1 = L + \frac{\frac{N}{4} - F}{f} \times C$$

$$\frac{N}{4} = \frac{100}{4} = 25^{\text{th}} \text{ observation}$$

$Q_1$  lies in the class of 30 – 40

$$\begin{aligned}
 &= 30 + \frac{25 - 14}{14} \times 10 \\
 &= 30 + \frac{11}{14} \times 10 \\
 &= 30 + \frac{110}{14} \\
 &= 30 + 7.86 = 37.86
 \end{aligned}$$

Bowley's coefficient of skewness

$$\begin{aligned}
 &= \frac{Q_3 + Q_1 - 2 \text{ median}}{Q_3 - Q_1} = \frac{61.33 + 37.86 - 2(50.8)}{61.33 - 37.86} \\
 &= \frac{99.19 - 101.6}{23.47} = \frac{-2.41}{23.47}
 \end{aligned}$$

$$\therefore S_{K_B} = -0.1027$$

32. Calculate Coefficient of Skewness based on Quarties from the following data:

Class Interval	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80	80 - 90
f	6	10	18	30	12	10	6	2

Sol.:

C.I	F	CF
10 - 20	6	6
20 - 30	10	16 F
L 30 - 40	18 f	34 F (Q1)
L 40 - 50	30 f	64 F (Median)
L 50 - 60	12 f	76 (Q3)
60 - 70	10	86
70 - 80	6	92
80 - 90	2	94
	94	

$$\text{Bowley's Coefficient of Skewness} = \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$$

Calculation of Median

$$= L + \frac{\frac{N}{2} - F}{f} \times C$$

$$\frac{N}{2} = \frac{94}{2} = 47^{\text{th}} \text{ item}$$

Median lies in the class of 40 – 50

$$= 40 + \frac{47 - 34}{30} \times 10$$

$$= 40 + \frac{130}{30}$$

$$= 40 + 4.33 = 44.33$$

Calculation of  $Q_3$

$$Q_3 = L + \frac{\frac{3N}{4} - F}{f} \times C$$

$$\frac{3N}{4} = \frac{3 \times 94}{4} = 70.5$$

$Q_3$  lies in the class of 50 – 60

$$= 50 + \frac{70.5 - 64}{12} \times 10$$

$$= 50 + \frac{6.5}{12} \times 10$$

$$= 50 + \frac{6.5}{12}$$

$$= 50 + 5.41$$

$$= 55.41$$

Calculation of  $Q_1$

$$= L + \frac{\frac{N}{4} - F}{f} \times C$$

$$\frac{N}{4} = \frac{94}{4} = 23.5$$

$Q_1$  lies in the class of 30 – 40

$$= 30 + \frac{23.5 - 16}{18} \times 10$$

$$= 30 + \frac{7.5}{18} \times 10$$

$$= 30 + \frac{7.5}{18}$$

$$= 30 + 4.16$$

$$= 34.16$$

Coefficient of skewness

$$= \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$$

$$= \frac{55.41 + 34.16 - 2(44.33)}{55.41 - 34.16}$$

$$= \frac{0.91}{21.25}$$

$$= 0.0428.$$

**1.6 KURTOSIS****1.6.1 Mesokurtosis, Platy Kurtosis and Leptokurtosis**

**Q45. Define kurtosis, Explain briefly about measuring kurtosis.**

*Ans :*

**Meaning**

Kurtosis in Greek means "bulginess". In statistics kurtosis refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. The degree of kurtosis of a distribution is measured relative to the peakedness of normal curve.

**(i) Leptokurtosis**

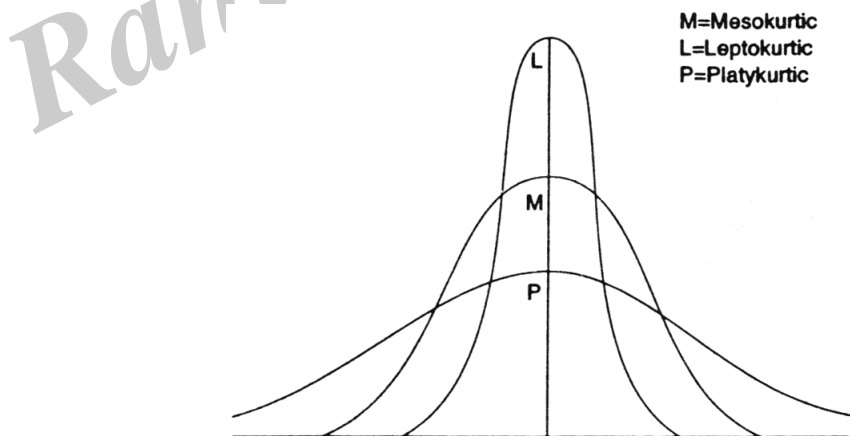
Measures of kurtosis tell us the extent to which a distribution is more peaked or flat-topped than the normal curve. If a curve is more peaked than the normal curve, it is called 'leptokurtic'. In such a case items are more closely bunched around the mode.

**(ii) Mesokurtosis**

If a curve is more flat-topped than the normal curve, it is called 'platykurtic'. The normal curve itself is known as 'mesokurtic'.

The condition of peakedness or flat-topped ness itself is known as kurtosis of excess. The concept of kurtosis is rarely used in elementary statistical analysis.

The following diagram illustrates the shape of three different curves mentioned above :



The above diagram clearly shows that these curves differ widely with regard to convexity, an attribute which Karl Pearson referred to as 'kurtosis'. Curve M is a normal one and is called 'mesokurtic'. Curve L is more peaked than M and is called 'leptokurtic'. A leptokurtic curve has a narrower central portion and higher tails than does the normal curve. Curve P is less peaked (or more flat-topped) than curve M and is called 'platykurtic'. As may be seen from the diagram, such a curve has a broader central and lower tails.

## 1.7 INTRODUCTION TO PROBABILITY

### 1.7.1 Concepts and Definitions of Probability

**Q46. Define the term probability?**

**(OR)**

**What is probability?**

*Ans :*

**(Dec.-20, July-18)**

#### Introduction

An Italian mathematician, Galileo (1564 - 1642), attempted a quantitative measure of probability while dealing with some problems related to gambling. In the middle of 17th Century, two French mathematicians, Pascal and Fermat, laid down the first foundation of the mathematical theory of probability while solving the famous 'Problem of Points' posed by Chevalier-De-Mere. Other mathematicians from several countries also contributed in no small measure to the theory of probability. Outstanding of them were two Russian mathematicians, A. Kintchine and A. Kolmogoroff, who axiomised the calculus of probability.

If an experiment is repeated under similar and homogeneous conditions, we generally come across two types of situations.

- (i) The net result, what is generally known as 'outcome' is unique or certain.
- (ii) The net result is not unique but may be one of the several possible outcomes.

The situations covered by :

- (i) are known as 'deterministic' or 'predictable' and situations covered by
- (ii) are known as 'probabilistic' or 'unpredictable'.

'Deterministic' means the result can be predicted with certainty. For example, if  $r$  is the radius of the sphere then its volume is given by  $V = \frac{4}{3}\pi r^3$  which gives uniquely the volume of the sphere.

There are some situations which do not lend themselves to the deterministic approach and they are known as 'Probabilistic'.

**For example**, by looking at the sky, one is not sure whether the rain comes or not.

In such cases we talk of chances or probability which can be taken as a quantitative measure of certainty.

#### Definitions

In a random experiment, let there be  $n$  mutually exclusive and equally likely elementary events. Let  $E$  be an event of the experiment. If  $m$  elementary events form event  $E$  (are favourable to  $E$ ), then the probability of  $E$  (Probability of happening of  $E$  or chance of  $E$ ), is defined as

$$P(E) = \frac{m}{n} = \frac{\text{Number of elementary events in } E}{\text{Total number of elementary events in the random experiment}}$$

If  $\bar{E}$  denotes the event of non-occurrence of  $E$ , then the number of elementary events in  $\bar{E}$  is  $n-m$  and hence the probability of  $\bar{E}$  (non-occurrence of  $E$ ) is

$$P(\bar{E}) = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - P(E) \Rightarrow P(E) + P(\bar{E}) = 1$$

Since  $m$  is a non-negative integer,  $n$  is a positive integer and  $m \leq n$ , we have  $m$

$$0 \leq \frac{m}{n} \leq 1.$$

Hence  $0 \leq P(E) \leq 1$  and  $0 \leq P(\bar{E}) \leq 1$ .

### 1. Statistical Probability

Suppose an experiment is repeated ' $w$ ' times under essentially identical conditions.

Let an event  $A$  happens  $w$ -times then  $\frac{m}{n}$  is defined as the relative frequency of  $A$ .

Statistical probability is also known as 'relative frequency' probability. The limit of this relative frequency as  $n \rightarrow \infty$  is defined as the probability of  $A$ .

$$\therefore P(A) = \lim_{n \rightarrow \infty} \frac{m}{n} \quad 0 \leq \frac{m}{n} \leq 1$$

### 2. Axiomatic Probability

In axiomatic probability three axioms or postulates are explained on the basis of which probability is calculated. They are,

- (i) Probability of an event ranges from, zero to one  $\Rightarrow 0 \leq P(A) \leq 1$
- (ii) Probability of entire sample space,  $P(S) = 1$
- (iii) If ' $A$ ' and ' $B$ ' are mutually exclusive events, then the probability of occurrence of either  $A$  or  $B$  is  $P(A \cup B) = P(A) + P(B)$ .

In general, there are three chances which can be expected for any events.

#### Q47. Explain the importance of probability.

*Ans :*

- (i) The probability theory is very much helpful for making prediction. Estimates and predictions form an important part of research investigation. With the help of statistical methods, we make estimates for the further analysis. Thus, statistical methods are largely dependent on the theory of probability.
- (ii) It has also immense importance in decision making.
- (iii) It is concerned with the planning and controlling and with the occurrence of accidents of all kinds.
- (iv) It is one of the inseparable tools for all types of formal studies that involve uncertainty.
- (v) The concept of probability is not only applied in business and commercial lines, rather than it is also applied to all scientific investigation and everyday life.
- (vi) Before knowing statistical decision procedures one must have to know about the theory of probability.
- (vii) The characteristics of the Normal Probability. Curve is based upon the theory of probability.



**Q48. Explain the various terms used in probability theory.**

*Ans :*

**(i) Random Experiment**

If an 'experiment' is conducted, any number of times, under essentially identical conditions, there is a set of all possible outcomes associated with it. If the result is not certain and is anyone of the several possible outcomes, the experiment is called a random trial or a random experiment. The outcomes are known as elementary events and a set of outcomes is an event. Thus an elementary event is also an event.

**(ii) Outcome**

The result of random experiment is usually referred as an outcome.

**(iii) Event**

An event is possible outcome of an experiment or a result of trial.

**(a) Simple Event:** In case of simple events we consider the probability of the happening or not happening of single events. For example, we might be interested in finding out the probability of drawing a red ball from a bag containing 10 white and 6 red balls.

**(b) Compound Events:** Compound events we consider the joint occurrence of two or more events. For example, if a bag contains 10 white and 6 red balls and if two successive draws of 3 balls are made, we shall be finding out the probability of getting 3 white balls in the first draw and 3 black balls in the second draw we are thus dealing with a compound event.

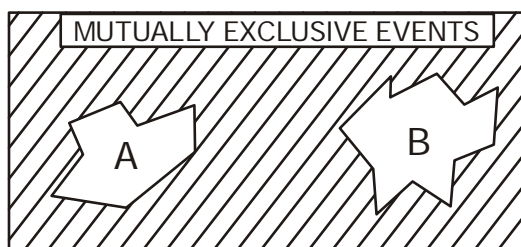
**(iv) Mutually Exclusive Events**

Two events are said to be mutually exclusive or incompatible when both cannot happen simultaneously in a single trial or, in other words, the occurrence of any one of them precludes the occurrence of the other.

**For example,** if a single coin is tossed either head can be up or tail can be up, both cannot be up at the same time. Similarly, a person may be either alive or dead at a point of time he cannot be both alive as well as dead at the same time.

To take another example, if we toss a dice and observe 3, we cannot expect 5 also in the same toss of dice. Symbolically, if A and B are mutually exclusive events,  $P(AB) = 0$ .

The following diagram will clearly illustrate the meaning of mutually exclusive events :



**Disjoint Sets**

It may be pointed out that mutually exclusive events can always be connected by the words "either ..... or". Events A, B, C are mutually exclusive only if either A or B or C can occur.

**(v) Collectively Exhaustive Events**

Events are said to be exhaustive when their totality includes all the possible outcomes of a random experiment. For example, while tossing a dice, the possible outcomes are 1, 2, 3, 4, 5 and 6 and hence the exhaustive number of cases is 6. If two dice are thrown once, the possible outcomes are :

(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)

The sample space of the experiment is  $6^2$  ordered pairs (6<sup>2</sup>). Similarly, for a throw of 3 dice exhaustive number of cases will be 216 (i.e. 6<sup>3</sup>) and for  $n$  dice they will be 6 <sup>$n$</sup> .

Similarly, black and red cards are examples of collectively exhaustive events in a draw from a pack of cards.

**(vi) Equally Likely Events**

Events are said to be equally likely when one does not occur more often than the others. For example, if an unbiased coin or dice is thrown, each face may be expected to be observed approximately the same number of times in the long run. Similarly, the cards of a pack of playing cards are so closely alike that we expect each card to appear equally often when a large number of drawings are made with replacement. However, if the coin or the dice is biased we should not expect each face to appear exactly the same number of times.

**(vii) Independent Event**

Two or more events are said to be independent when the outcome of one does not affect, and is not affected by the other.

For example, if a coin is tossed twice, the result of the second throw would in no way be affected by the result of the first throw. Similarly, the results obtained by throwing a dice are independent of the results obtained by drawing an ace from a pack of cards.

To consider two events that are not independent, let  $A$  stand for a firm's spending a large amount of money on advertisement and  $B$  for its showing an increase in sales. Of course, advertising does not guarantee higher sales, but the probability that the firm will show an increase in sales will be higher if  $A$  has taken place.

**(viii) Dependent Event**

Dependent events are those in which the occurrence or non-occurrence of one event in any one trial affects the probability of other events in other trials. For example, if a card is drawn from a pack of playing cards and is not replaced, this will alter the probability that the second card drawn is, say

an ace. Similarly, the probability of drawing a queen from a pack of 52 cards is  $\frac{4}{52}$  or  $\frac{1}{13}$ . But if the

card drawn (queen) is not replaced in the pack, the probability of drawing again a queen is  $\frac{3}{51}$  (the pack now contains only cards out of which there are 3 queens).

**(xi) Non-mutually Exclusive Events**

When two events can occur simultaneously in a single trial then such events are said to be non-mutually exclusive events.

**Example**, from a pack of cards, drawing a red card and drawing a queen are the two events. These two events can occur simultaneously while drawing a red queen.

Hence, these two events are said to be non-mutually exclusive events which can occur at the same time.

**33. What is the probability for a leap year to have 52 Mondays and 53 Sundays?**

*Sol.:*

A leap year has 366 days i.e., 52 weeks and 2 days.

These two days can be any one of the following 7 ways :

- (i) Mon & Tue
- (ii) Tues & Wed
- (iii) Wed & Thurs
- (iv) Thurs & Fri
- (v) Fri & Sat
- (vi) Sat & Sun
- (vii) Sun & Mon

Let E be the event of having 52 Mondays and 53 Sundays in the year.

Total number of possible cases is  $n = 7$

Number of favourable cases to E is  $m = 1$

(Sat & Sun is the only favourable case)

$$\therefore P(E) = \frac{m}{n} = \frac{1}{7}$$

**34. Five digit numbers are formed with 0, 1, 2, 3, 4 (not allowing a digit being repeated in any number). Find the probability of getting 2 in the ten's place and 0 in the units place always.**

*Sol.:*

Total number of 5 digit numbers using the digits 0, 1, 2, 3, 4 is

$$= n = 4 \times 4 \times 3 \times 2 \times 1 = 96 \text{ (or) } 5! - 4! = 96$$

Let E be the event of getting a number having 2 in 10's place and 0 in the units place.

So the number of numbers favourable to is  $= m = 3.2.1.1.1 = 6$

$$\therefore P(E) = \frac{m}{n} = \frac{6}{96} = \frac{1}{16}$$

**35. In a class there are 10 boys and 5 girls. A committee of 4 students is to be selected from the class. Find the probability for the committee to contain at least 3 girls.**

*Sol.:*

A committee of 4 students out of 15 can be formed in  ${}^{15}C_4$  ways i.e.,  $n = {}^{15}C_4$

Let E be the event of forming a committee with at least 3 girls.

Now the committee can have 1 boy, 3 girls or no boy, 4 girls. So the number of ways of forming the committee = The number of favourable ways to E

$$= {}^{10}C_1 \times {}^5C_3 + {}^{10}C_0 \times {}^5C_4 = 100 + 5 = 105$$

$$\therefore P(E) = \frac{m}{n} = \frac{105}{{}^{15}C_4} = 0.0769$$

36. When two dice are thrown, find the probability that the sum of the numbers is either 10 or 11.

*Sol:*

When 2 dice are thrown sample spaces  $6^2 = 36$

The no. of possible outcomes

$$(4, 6) (5, 5) (5, 6) (6, 4) (6, 5) = \frac{5}{36}$$

Let "A" be the Event that number selected would be sum is 10.

'B' be the Even that number selected would be sum is 11.

### 1.8 APPROACHES TO PROBABILITY

Q49. What are the approaches of probability.

*Ans :*

They are four different approaches on Broadly. The concept of probability. They are as follows.

1. Classical (or) priori probability
2. Relative/empirical probability
3. Subjective approach
4. Axiomatic approach

#### 1.8.1 Classical

Q50. Explain in detail about Classical approach used in probability.

*Ans :*

The probability of a given event is an expression of likelihood or chance of occurrence or an event. A probability is a number which ranges from 0 (zero) to (one) - zero for an event which cannot occur and 1 for an event certain to occur. How the number is assigned would depend on the interpretation of the term 'probability'. There is no general agreement about its interpretation and many people associate probability and chance with nebulous and mystic ideas. However, broadly speaking there are four different schools of thought on the concept of probability.

If after 'n' repetitions of an experiment, where n is very large, an event is observed to occur in K of these, then the probability of the event is K/n.

Consider an experiment which has 'n' exhaustive, mutually exclusive and equally likely events with 'K' events from the 'n' in favour of happening of the event A, then the probability 'P' of A is

$$P(A) = \frac{K}{n}$$

$$= \frac{\text{No. of events in favour of A}}{\text{No. of exhaustive equally likely and mutually exclusive events of the experiment}}$$

P(A) is also known as probability of success.

### Limitations

Classical theory does not holds good,

- (a) When all the outcomes are not equally likely
- (b) When the collectively exhaustive events of an experiment are infinite
- (c) Classical theory does not provide answers to certain question which occurs in our daily life.

**For example,**

What is the probability of occurrence of rain now? The chances of bulb getting failed etc.

### 1.8.2 Relative / Empiricals

**Q51. Explain in detail about Empirical approach used in probability.**

*Ans :*

In the 1800s, British statisticians, interested in a theoretical foundation for calculating risk of losses in life insurance and commercial insurance, began defining probabilities from statistical data collected on births and deaths. Today this approach is called relative frequency of occurrence.

This classical definition is difficult or impossible to apply as soon as we deviate from the fields of coins, dice, cards and other simple games of chance. Secondly, the classical approach may not explain actual results in certain cases.

**For example,** if a coin is tossed 10 times we may get 6 heads and 4 tails. The probability of a head is thus 0.6 and that of a tail 0.4. However, if the experiment is carried out a large number of times we should expect approximately equal number of heads and tails. As  $n$  increases, i.e., approaches  $\infty$  (infinity), we find that the probability of getting a head or tail approaches 0.5. The probability of an event can thus be defined as the relative frequency with which it occurs in an indefinitely large number of trials. If an

event occurs  $a$  times out of  $n$ , its relative frequency is  $\frac{a}{n}$  the value which is approached by  $\frac{a}{n}$  when  $n$  becomes infinity is called the limit of the relative frequency.

Symbolically 
$$P(A) = \lim_{n \rightarrow \infty} \frac{a}{n}$$

Theoretically, we can never obtain the probability of an event as given by the above limit. However, in practice we can only try to have a close estimate of  $P(A)$  based on a large number of observations. i.e.,  $n$ . For practical convenience, the estimate of  $P(A)$  can be written as if it were actually  $P(A)$  and the relative frequency definition of probability may be expressed as:

$$P(A) = \frac{a}{n}$$

In the relative frequency definition the fact that the probability is the value which is approached by  $\frac{a}{n}$  when  $n$  becomes infinity, emphasises a very important point, i.e., probability involves a long- term concept. This means that if we toss a coin only 10 times, we may not get exactly 5 heads and 5 tails. However, as the experiment is carried out larger and larger number of times, say, coin is thrown 10,000 times, we can expect heads and tails very close to 50 per cent.

The two approaches, classical and empirical, though seemingly same, differ widely. In the former,

$P(A)$  and  $\frac{a}{n}$  were practically equal when  $n$  was large whereas in the latter we say that  $P(A)$  is the limit  $\frac{a}{n}$  as  $n$  tends to infinity. In the second approach, thus, the probability itself is the limit of the relative frequency as the number of observations increases indefinitely.

### Limitations

- (i) As the experiments are repeated large number of times, it takes large amount of time.
- (ii) During the experimental time, the conditions may not be always identical and homogenous.

### 1.8.3 Subjective

**Q52. Explain in detail about Subjective approach used in probability.**

*Ans :*

The subjective approach to assigning probabilities was introduced in the year 1926 by Frank Ramsey in his book, The Foundation of Mathematics and other Logical Essays. The concept was further developed by Bernard Koopman, Richard Good and Leonard Savage. The subjective probability is defined as the probability assigned to an event by an individual based on whatever evidence is available. Hence such probabilities are based the beliefs of the person making the probability statement.

For example, if a teacher wants to find out the probability of Mr. X topping in M.Com. examination in Delhi University this year, he may assign a value between zero and one according to his degree of belief for possible occurrence. He may take into account such factors as the past academic performance, the views of his other colleagues, the attendance record, performance in periodic tests, etc., and arrive at a probability figure.

This concept emphasises the fact that since probability of an event is the degree of belief or degree of confidence placed in the occurrence of an event by a particular individual based on the evidence available to him. different individuals may differ in their degrees of confidence even when offered the same evidence. This evidence may consist of relative frequency of occurrence data and any other quantitative or non-quantitative information. Persons might arrive at different probability assignments because of differences in values, experience and attitudes, etc. If an individual believes that it is unlikely that an event will occur, he will assign a probability close to zero to its occurrence. On the other hand, if he believes that it is very likely that the event will occur, he will assign a probability close to one.

The personalistic approach is very broad and highly flexible. It permits probability assignment to events for which there may be no objective data, or for which there may be a combination of subjective and objective data. However, one has to be very careful and consistent in the assignment of these probabilities otherwise the decisions made may be misleading. Used with care the concept is extremely useful in the context of situations in business decision-making.

### 1.8.4 Axiomatic

**Q53. Explain the Axiomatic approach to probability.**

*Ans :*

The axiomatic approach to probability was introduced by the Russian mathematician A. N. Kolmogorov in the year 1933. Kolmogorov axiomised the theory of probability and his book Foundations of Probability, published in 1933, introduces probability as a set function and is considered as a classic. When this approach is followed, no precise definition of probability is given, rather we give certain axioms or postulates on which probability calculations are based. The whole field of probability theory for finite sample spaces\* is based upon the following three axioms :

1. The probability of an event ranges from zero to one. If the event cannot take place its probability shall be zero and if it is certain, i.e., bound to occur, its probability shall be one.
2. The probability of the entire sample space is 1, Le.,  $P(S) = 1$ .
3. If A and B are mutually exclusive (or disjoint) events then the probability of occurrence of either A or B denoted by  $P(A \cup B)$  shall be given by :

$$P(A \cup B) = P(A) + P(B)$$

It may be pointed out that out of the four interpretations of the concept of probability, each has its own merits and one may use whichever approach is convenient and appropriate for the problem under consideration.

The probability of a event A, denoted by  $P(A)$  is so chosen as to satisfy the following three axioms.

- i)  $P(A) \geq 0 \Rightarrow$  This axiom states that the probability of occurrence of an event A in a random experiment may be zero or any positive number and it must not be negative number.
- ii)  $P(S) = 1 \Rightarrow$  This states that the sample space, S, itself is an event and since it is the event comprising all possible outcomes, it should have the highest possible probability, i.e., one.
- iii) If  $A \cap B = \emptyset$ , Then  $P(A \cup B) = P(A) + P(B) \Rightarrow$  This axiom states that the probability of the event equal to the union of any number of mutually exclusive events is equal to the sum of the individual even probabilities.

- 37. A class consists of 6 girls and 10 boys. If a committee of 3 is chosen at random from the class, find the probability that (i) 3 boys are selected (ii) exactly 2 girls are selected.**

*Sol.:*

Total number of students = 16

$n(S) = \text{no. of ways of choosing 3 from 16} = {}^{16}C_3$

- (i) Suppose 3 boys are selected. This can be done in  ${}^{10}C_3$  ways.

Here  $n(E) = {}^{10}C_3$

$\therefore P(E) = \text{The probability that 3 boys are selected} = \frac{n(E)}{n(S)}$

$$= \frac{{}^{10}C_3}{{}^{16}C_3} = \frac{10 \times 9 \times 8}{16 \times 15 \times 14} = \frac{3}{14} = 0.2143$$

- (ii) Suppose exactly 2 girls are selected. Then

$$n(E) = {}^6C_2 \times {}^{10}C_1$$

$$\therefore P(E) = \frac{n(E)}{n(S)} = \frac{{}^6C_2 \times {}^{10}C_1}{{}^{16}C_3} = \frac{15}{56} = 0.2678$$

38. A and B throw alternately with a pair of ordinary dice. A wins if he throws 6 before B throws 7 and B wins if he throws 7 before A throws 6. If A begins, show that his chance of winning is 30/61.

*Sol.:*

When two dice are thrown, we have  $n(s) = 36$

The probability of A throwing '6' =  $\frac{5}{36}$  i.e.,  $P(A) = \frac{5}{36}$

The probability of A not throwing '6' is and is given by

$$P(\bar{A}) = 1 - P(A) = 1 - \frac{5}{36} = \frac{31}{36}$$

The probability of B throwing '7' =  $\frac{6}{36}$  i.e.,  $P(B) = \frac{6}{36} = \frac{1}{6}$

The probability of B not throwing 7 is  $P(\bar{B}) = 1 - P(B) = 1 - \frac{6}{36} = \frac{5}{6}$

$\therefore$  Chances of winning of 'A' is

$$= P(A) + P(\bar{A})P(\bar{B})P(A) + P(\bar{A})P(\bar{B})P(\bar{A})P(\bar{B})P(A) + \dots$$

$$= \left(\frac{5}{36}\right) + \left(\frac{31}{36} \times \frac{5}{6} \times \frac{5}{36}\right) + \left(\frac{31}{36} \times \frac{5}{6} \times \frac{31}{36} \times \frac{5}{6} \times \frac{5}{36}\right) + \dots$$

$$= \frac{5}{36} \left[ 1 + \left(\frac{31}{36} \times \frac{5}{6}\right) + \left(\frac{31}{36} \times \frac{5}{6}\right)^2 + \dots \right] = \frac{5}{36} \left[ \frac{1}{1 - \left(\frac{31}{36} \times \frac{5}{6}\right)} \right] = \frac{30}{61}$$

## 1.9 THEOREMS OF PROBABILITY

### 1.9.1 Addition

**Q54. Explain Addition theorem of probability.**

*Ans.:*

(Dec.-20, June-19, July-18)

Addition theorem is different for mutually exclusive non-mutually exclusive events.

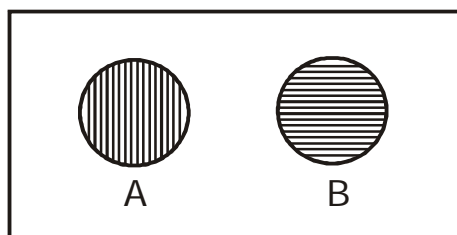
#### (i) For Mutually Exclusive Events

When 'A' and 'B' are two mutually exclusive events (i.e., both cannot occur at the same time) then the probability of occurrence of A or B is equal to the sum of their individual probabilities.

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) \\ \Rightarrow P(A \cup B) &= P(A) + P(B) \end{aligned}$$



Diagrametrically it can be represented as,



Mutually exclusive events

**Figure: Mutually Exclusive Events**

In case of 3 events A, B and C,

$$P(A \text{ or } B \text{ or } C) = P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

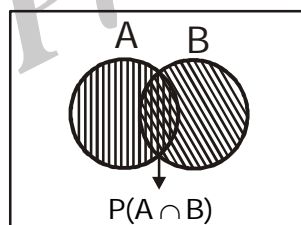
### (ii) For Non-Mutually Exclusive Events

In case of non-mutually exclusive event (i.e., if the events occur together) there is a variation in the addition theorem.

When 'A' and 'B' are non-mutually exclusive events then the probability of occurrence of A or B is the sum of their individual probability which should be deducted from the probability of A and B occurring together.

$$P(A \text{ or } B) \Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Diagrametrically it can be represented as,



**Fig.: Non-Mutually Exclusive Events**

In case of three non-mutually exclusive events.

A, B and C the probability of occurrence of A or B or C can be calculated by the following formula,

$$P(A \text{ or } B \text{ or } C) = P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

- 39. A card is drawn from a well shuffled pack of cards. What is the probability that is either a spade or an ace ?**

*Sol/ :*

Let S is the sample space of all the simple events.

$$\therefore n(s) = 52$$

Let A denote the event of getting a spade and B denote the event of getting an ace.

Then  $A \cup B$  = The event of getting a spade or an ace

$A \cap B$  = The event of getting a spade and an ace

$$\therefore P(A) = \frac{13}{52}, P(B) = \frac{4}{52}, P(A \cap B) = \frac{1}{52}$$

By Addition Theorem,

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13} \end{aligned}$$

40. Three students A, B, C are in running race. A and B have the same probability of winning and each is twice as likely to win as C. Find the probability that B or C wins.

*Sol :*

$A \cup B \cup C = S$  = Sample space of race

By data,  $P(A) = P(B)$  and  $P(A) = 2 P(C)$  .....(1)

We have  $P(A) + P(B) + P(C) = 1 \Rightarrow 2P(C) + 2 P(C) + P(C) = 1$  [by (1)]

$$\Rightarrow P(C) = \frac{1}{5} P(A) = \frac{2}{5} \text{ and } P(B) = \frac{2}{5}$$

$$\begin{aligned} \text{The probability that B or C wins} &= P(B \cup C) \\ &= P(B) + P(C) - P(B \cap C) \\ &= \frac{2}{5} + \frac{2}{5} - 0 = \frac{4}{5} \end{aligned}$$

41. From a city 3 news papers A, B, C are being published. A is read by 20%, B is read by 16%, C is read by 14% both A and B are read by 8%, both A and C are read by 5% both B and C are read by 4% and all three A, B, C are read by 2%. What is the percentage of the population that read at least one paper.

*Sol :*

Given,  $P(A) = \frac{20}{100}$ ,  $P(B) = \frac{16}{100}$ ,  $P(C) = \frac{14}{100}$  and

$$P(A \cap B) = \frac{8}{100}, P(A \cap C) = \frac{5}{100}, P(B \cap C) = \frac{4}{100} \text{ and } P(A \cap B \cap C) = \frac{2}{100}$$

$$\therefore P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) + P(A \cap B \cap C)$$

$$= \frac{20}{100} + \frac{16}{100} + \frac{14}{100} - \frac{8}{100} - \frac{5}{100} + \frac{2}{100} = \frac{35}{100}$$

$$\therefore \text{Percentage of the population that read at least one paper} = \frac{35}{100} \times 100 = 35$$

42. One card is drawn at random from a pack of 52 cards. What is the probability that it is either a king or a queen?

*Sol:*

**Probability of Drawing a King Card**

Let  $P(A)$  denoted as probability of drawing a king card from a pack of cards.

Total number of king cards = 4

1 kind card is drawn from 4 king cards =  $4c_1 = 4$ .

Let total No. of Playing cards in a pack = 52.

1 card is drawn from 52 cards =  $52c_1 = 52$

$$\therefore P(A) = \frac{4}{52}$$

**Probability of Drawing a Queen Card**

Let  $P(B)$  denoted as probability of drawing 1 Queen card from a pack of cards.

Total No. of Queen cards = 4

1 card is drawn from 4 Queen cards =  $4c_1 = 4$ .

Let total No. of playing cards in a pack = 52

1 card is drawn from 52 cards =  $52c_1 = 52$ .

$$\therefore P(B) = \frac{4}{52}$$

Probability of drawing a king or Queen is

$$P(A \cup B) = P(A) + P(B) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} \text{ (or) } \frac{2}{13}.$$

43. A bag contains 4 defective and 6 good Electronic Calculators. Two calculators are drawn at random one after the other without replacement. Find the probability that

- i) Two are good
- ii) Two are defective and
- iii) One is good and one is defective.

*Sol:*

Total Number of calculators in a bag =  $4 + 6 = 10$

**(i) Probability of Drawing 2 good Calculators**

Let  $P(A)$  denoted as drawing 1 good calculator from total calculators.

$$\therefore P(A) = \frac{6}{10}$$

Let  $P(B)$  denoted as drawing 2<sup>nd</sup> good calculator without replacing the 1<sup>st</sup> calculator.

Total good calculators after first calculator is not replaced =  $6 - 1 = 5$ .

Total Number of calculators after first calculator is drawn and not replaced =  $10 - 1 = 9$ .

$$\therefore P(B) = \frac{5}{9}$$

$\therefore$  Probability of drawing 2 good calculators .

$$P(A) \cdot P(B) = \frac{6}{10} \times \frac{5}{9} = \frac{30}{90} = \frac{1}{3}$$

**(ii) Probability of Getting Two Defectives**

Let P(A) denoted as drawing a defective calculator

Total Number of defective calculators = 4

1 Calculator is drawn from 4 =  $4C_1 = 4$ .

Total calculators in the bag =  $4 + 6 = 10$

One calculator is drawn from 10 =  $10C_1 = 10$ .

$$\therefore P(A) = \frac{4}{10}$$

Let P(B) denoted as drawing another defective calculator without replacing the first.

Total number of defective calculator after first calculator is drawn and not replaced =  $4 - 1 = 3$ .

Total number of calculators in bag after first calculator is drawn and not replaced =  $10 - 1 = 9$ .

$$\therefore P(B) = \frac{3}{9}$$

$\therefore$  Probability of drawing two are defective without replacing is  $P(A) \cdot P(B)$ .

$$= \frac{4}{10} \times \frac{3}{9} = \frac{12}{90} = \frac{2}{15}$$

**(iii) One is good and one is defective**

Let P(A) denoted as drawing 1 good calculator.

Number of good calculators = 6.

1 calculator is drawn from 6 =  $6C_1 = 6$ .

Total number 9 calculators in bag = 10

1 calculator is drawn from to =  $10C_1 = 10$ .

$$\therefore P(A) = \frac{6}{10}$$

Let P(B) denoted as drawing another calculator which is defective.

Total Number of defective calculators = 4.

1 calculator is drawn from 4 =  $4C_1 = 4$ .

Total number of calculators after first calculator is drawn and not replaced =  $10 - 1 = 9$ .

$$\therefore P(B) = \frac{4}{9}$$

Probability of drawing 1 good and 1 bad is  $P(A) \cdot P(B)$

$$= \frac{6}{10} \times \frac{4}{9} = \frac{24}{90} = \frac{4}{15}$$

### 1.9.2 Multiplication

#### Q55. Explain Multiplication theorem of probability.

*Ans :*

(Dec.-20, June-18)

If 'A' and 'B' are two independent events then the probability of occurrence of both the events is equal to the product of their individual probabilities.

For independent events,

$$P(A \cap B) = P(A) \cdot P(B)$$

Similarly,

$$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C) \text{ and so on.}$$

If 'A' and 'B' are two dependent events, in such a case multiplication theorem is altered and is given as follows. For dependent events,

$$\begin{aligned} P(A \cap B) &= P(A/B) \cdot p(B) \\ &= P(B/A) \cdot P(A) \end{aligned}$$

Where,  $P(A/B)$  is a conditional probability of A given that B has occurred (The probability of occurrence of event A when event B has already occurred is the conditional probability of A given B).

#### 44. Find the probability of drawing 2 red balls in succession from a bag containing 4 red and 5 black balls when the ball that is drawn first is not replaced (ii) replaced.

*Sol :*

Let  $E_1$  be the event of drawing a red ball in the first draw and  $E_2$  be the event of drawing a red ball in second draw also.

- (i) After the first draw the ball is not replaced. The first ball can be drawn in 9 ways and the second in 8 ways since the first ball is not replaced. Then both the balls can be drawn in  $9 \times 8$  ways.

There are 4 ways in which  $E_1$  can occur and 3 ways in which  $E_2$  can occur, so that  $E_1$  and  $E_2$  can occur in  $4 \times 3$  ways.

$$p\left(\frac{E_2}{E_1}\right) = P(E_2, \text{ given the probability of } E_1)$$

$$= P(2\text{nd ball is red, given that first ball is red}) = \frac{3}{8}$$

$$\therefore P(E_1 \cap E_2) = P(E_1) \times P\left(\frac{E_2}{E_1}\right) = \frac{4}{9} \times \frac{3}{8} = \frac{1}{6} \quad \left[ \because P(E_1) = \frac{4}{9} \right]$$

- (ii) Suppose the ball is replaced after the first draw. Then

$$P(E_1 \cap E_2) = \frac{4}{9} \cdot \frac{4}{9} = \frac{16}{81}$$

- 45. A class has 10 boys and 5 girls. Three students are selected at random one after another. Find the probability that (i) first two are boys and third is girl (ii) First and third are of same sex and the second is of opposite sex.**

*Sol :*

Total no. of students = 10 + 5 = 15

- (i) The probability that first two are boys and the third is girl is

$$P(E_1 \cap E_2 \cap E_3) = \frac{10}{15} \cdot \frac{9}{14} \cdot \frac{5}{13} = \frac{15}{91}$$

- (ii) Suppose the first and third are boys and second is a girl

$$\text{Probability of the event} = P(E_1) = \frac{10}{15} \cdot \frac{5}{14} \cdot \frac{9}{13} = \frac{15}{91} = 0.1648$$

Suppose first and third are girls and second is boy.

$$\text{Then the probability of the event} = P(E_2) = \frac{5}{15} \cdot \frac{10}{14} \cdot \frac{4}{13} = \frac{20}{273}$$

$$\therefore \text{Required probability} = P(E_1) + P(E_2)$$

$$= \frac{15}{91} + \frac{20}{273} = \frac{45 + 20}{273} = \frac{65}{273} = 0.238$$

- 46. Two marbles are drawn in succession from a box containing 10 red, 30 white 20 blue and 15 orange marbles, with replacement being made after each draw. Find the probability that**

**(i) Both are white**

**(ii) First is red and second is white.**

*Sol :*

Total no. of marbles in the box = 75

- (i) Let  $E_1$  be the event of the first drawn marble is white. Then

$$P(E_1) = \frac{30}{75}$$

Let  $E_2$  be the event of second drawn marble is also white. Then

$$P(E_2) = \frac{30}{75}$$

The probability that both marbles are white (with replacement)

$$= P(E_1 \cap E_2) = P(E_1) \cdot P(E_2|E_1) = \frac{30}{75} \cdot \frac{30}{75} = \frac{4}{25}$$

(ii) Let  $E_1$  be the event that the first drawn marble is red. Then

$$P(E_1) = \frac{10}{75} = \frac{2}{15}$$

Let  $E_2$  be the event that the drawn marble is white. Then

$$P(E_2|E_1) = \frac{30}{75} = \frac{2}{5}$$

$\therefore$  The probability that the First marble is red and Second marble is white

$$= P(E_1 \cap E_2) = P(E_1) \cdot P(E_2|E_1)$$

$$= \frac{2}{15} \cdot \frac{2}{5} = \frac{4}{75}$$

- 47. Three boxes, practically indistinguishable in appearance have two drawers each. Box 1 contains a gold coin in first and silver coin in the other drawer, Box 2 contains a gold coin in each drawer and Box 3 contains a silver coin in each drawer. One box is chosen at random and one of its drawers is opened at random and a gold coin is found. What is the probability that the other drawer contains a coin of silver.**

*Sol:*

Let  $E$  denote the event that the box is chosen,  $i = 1, 2, 3$ .

$$P(E_i) = \frac{1}{3} \text{ for } i = 1, 2, 3$$

Let  $A$  be the event that the gold coin is chosen. Then

$P(A|E_i)$  = Probability that a gold coin is chosen from the box  $i = 1, 2, 3$ .

$$\therefore P(A|E_1) = \frac{1}{2} \quad (\because \text{The total no. of coins in box 1 is 2})$$

$$P(A|E_2) = \frac{2}{2} = 1 \quad (\text{There are two gold coins in box 2})$$

$$\text{and } P(A|E_3) = \frac{0}{2} = 0 \quad (\text{There is no gold coin in box 3})$$

(i) The probability that the drawn coin is gold

$$P(A) = P(E_1) P(A|E_1) + P(E_2) P(A|E_2) + P(E_3) P(A|E_3)$$

$$= \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot 1 + 0 = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$$

- (ii) The probability that the drawn coin is silver

$$P(B) = 1 - P(A) = 1 - \frac{1}{2} = \frac{1}{2}$$

**48. Two digits are selected at random from the digits 1 through 9.**

- (i) If the sum is odd, what is the probability that 2 is one of the numbers selected ?  
 (ii) If 2 is one of the digits selected, what is the probability that the sum is odd ?

*Sol :*

The given set consists of five odd digits (1, 3, 5, 7, 9) and four even digits (2, 4, 6, 8).

We know that

even + even = even

even + odd = odd

odd + even = odd

odd + odd = even

- (i) Total number of events =  $5 \times 4 = 20$

If '2' is one of the digits, then the other digit must be odd.

$\therefore$  Number of ways = 5

So, required probability =  $\frac{5}{20} = \frac{1}{4}$

- (ii) If 2 is selected, then the remaining number of digits = 8

$\therefore$  Total events = 8

If 2 is one of the digits selected, the probability that the sum is odd

$$= \frac{\text{Favourable cases for odd}}{\text{Total events}} = \frac{5}{8}$$

### 1.10 STATISTICAL INDEPENDENCE

#### 1.10.1 Marginal, Conditional and Joint Probabilities

**Q56. Explain the concept of independent event. Discuss the different types of probabilities under Statistical Independence.**

*Ans :*

(Dec.-20, Aug.-17)

#### 1. Marginal Probability

Marginal probability is the simple probability of the occurrence of an event. It is also called as 'single probability' or 'unconditional probability'.



These two events are independent and do not overlap one another i.e., the events are statistically independent of the outcomes of next coin been tossed

The individual probabilities obtained in this case i.e.,  $P(H)$  or  $P(T)$  are called as marginal probabilities.

## 2. Joint Probability

A joint probability is the probability of occurrence of two or more simple independent events simultaneously.

In other words, the product of two marginal probabilities occurring together or in succession is called as 'joint probability'.

For example, let  $P(W)$  is the probability of a girl with while complexion and  $P(B)$  is the probability of a girl with black hair. Here,  $P(W)$  and  $P(B)$  are marginal probabilities.

Probability of a girl with white complexion and with black hair together is called as 'joint probability'. Joint probability of these two events can be represented as,  $P(W \cap B)$ .

$$\therefore P(W \cap B) = P(W) \cdot P(B)$$

Where,  $P(W \cap B)$  = Probability of a girl with white complexion and with black hair.

Similarly, joint probability for three independent events A, B and C can be represented as  $P(A \cap B \cap C)$ .

$$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$$

## 3. Conditional Probability

Conditional Probability is the probability of occurrence of second event (B) given that the first event (A) has previously occurred.

For statistically independent events, the conditional probability of event B given that event A has already occurred is simply the probability of event B. Symbolically, it can be represented as,

$$P(B/A) = P(B).$$

49. A husband and wife appear in an interview for two vacancies in the same post. The probability of husband's selection is  $1/7$  and that of wife's selection is  $1/5$ . What is the probability that only one of them will be selected ?

*Sol.:*

(Aug.-17)

$$P(A \bar{B} \cup \bar{A} B) = P(A) \cdot P(\bar{B}) + P(\bar{A}) \cdot P(B)$$

$$P(A) = \frac{1}{7} \quad P(B) = \frac{1}{5}$$

$$P(\bar{A}) = 1 - \frac{1}{7} = \frac{6}{7} \quad P(\bar{B}) = 1 - \frac{1}{5} = \frac{4}{5}$$

$$= \frac{1}{7} \times \frac{4}{5} + \frac{6}{7} \times \frac{1}{5}$$

$$\frac{4}{35} + \frac{6}{35} = \frac{10}{35} = \frac{2}{7}$$

### 1.11 BAYES' THEOREM

**Q57. State and explain Baye's probability theorem.**

*Ans :*

(Nov.-20)

$E_1, E_2, \dots, E_n$  are  $n$  mutually exclusive and exhaustive events such that  $P(E_i) > 0$  ( $i = 1, 2, \dots, n$ ) in a sample space  $S$  and  $A$  is any other event in  $S$  intersecting with every  $E_i$  (i.e.,  $A$  can only occur in combination with any one of the events  $E_1, E_2, \dots, E_n$ ) such that  $P(A) > 0$ .

If  $E_k$  is any of the events of  $E_1, E_2, \dots, E_n$  where  $P(E_1), P(E_2), \dots, P(E_n)$  and  $P(A/E_1), P(A/E_2), \dots, P(A/E_n)$  are known, then

$$P(E_k|A) = \frac{P(E_k) \cdot P(A/E_k)}{P(E_1) \cdot P(A/E_1) + P(E_2) \cdot P(A/E_2) + \dots + P(E_n) \cdot P(A/E_n)}$$

**Proof :**

$E_1, E_2, \dots, E_n$  are  $n$  events of  $S$  such that  $P(E_i) > 0$  and  $E_i \cap E_j = \phi$  for  $i \neq j$  where  $i, j = 1, 2, \dots, n$ . Also  $E_1, E_2, \dots, E_n$  are exhaustive events of  $S$  and  $A$  is any other event of  $S$  where  $P(A) > 0$ .

$$S = E_1 \cup E_2 \cup \dots \cup E_n \text{ and}$$

$$A = A \cap S = A \cap (E_1 \cup E_2 \cup \dots \cup E_n)$$

$$= (A \cap E_1) \cup (A \cap E_2) \cup \dots \cup (A \cap E_n)$$

Here  $A \cap E_1, A \cap E_2, \dots$  are mutually exclusive events. Then

$$P(E_k/A) = \frac{P(E_k \cap A)}{P(A)} = \frac{P(E_k \cap A)}{P[(A \cap E_1) \cup (A \cap E_2) \cup \dots \cup (A \cap E_n)]}$$

$$= \frac{P(E_k \cap A)}{P(A \cap E_1) + P(A \cap E_2) + \dots + P(A \cap E_n)}$$

$$= \frac{P(E_k) \cdot P(A/E_k)}{P(E_1) \cdot P(A/E_1) + P(E_2) \cdot P(A/E_2) + \dots + P(E_n) \cdot P(A/E_n)}$$

**Note :**

Baye's theorem is also known as formula for the Probability of "Causes", i.e., probability of a particular (cause)  $E$  given that event  $A$  has happened (already).

$P(E_i)$  is 'a priori probability' known even before the experiment,  $P(A/E_i)$  "Likelihoods" and  $P(E_i/A)$  'Posteriori Probabilities' determined after the result of the experiment.

#### 1.11.1 Applications

**Q58. What are the applications of Baye's Theorem?**

*Ans :*

(Imp.)

Following points highlights the application of Baye's theorem,

1. In Baye's theorem, posterior probabilities can be known by revising priori probabilities with the help of new information

2. The probability of occurrence of future events can be known by Baye's theorem.
3. Baye's theorem offers a powerful statistical tool.
4. Baye's theorem helps the business and management executives to take effective decisions in uncertain situation.

Baye's theorem is also known as 'Probability of Causes' as it helps in determining the probability which a particular effect has due to a specific cause.

**50. In a certain college, 25% of boys and 10% of girls are studying mathematics. The girls constitute 60% of the student body,**

- (a) What is the probability that mathematics is being studied ?
- (b) If a student is selected at random and is found to be studying mathematics, find the probability that the student is a girl?
- (c) Probability of maths student is a boy

*Sol:*

$$\text{Given } P(\text{Boy}) = P(B) = \frac{40}{100} = \frac{2}{5}$$

$$\text{and } P(\text{Girl}) = P(G) = \frac{60}{100} = \frac{3}{5}$$

$$\text{Probability that mathematics is studied given that the student is a boy} = P\{M / B\} = \frac{25}{100} = \frac{1}{4}$$

$$\text{Probability that mathematics is studied given that the student is a girl} = P(M / G) = \frac{10}{100} = \frac{1}{10}$$

**(a) Probability that the student studied Mathematics = P(M)**

$$= P(G) P(M/G) + P(B) P(M/B)$$

$\therefore$  By total probability theorem,

$$\begin{aligned} P(M) &= \frac{3}{5} \cdot \frac{1}{10} + \frac{2}{5} \cdot \frac{1}{4} \\ &= \frac{4}{25} \end{aligned}$$

**(b) By Baye's theorem, probability of mathematics student is a girl = P(G / M)**

$$= \frac{P(G)P(M / G)}{P(M)} = \frac{\frac{3}{5} \cdot \frac{1}{10}}{\frac{4}{25}} = \frac{3}{8}$$

**(c) Probability of maths student is a boy = P(B / M)**

$$= \frac{P(B)P(M / B)}{P(M)} = \frac{\frac{2}{5} \cdot \frac{1}{4}}{\frac{4}{25}} = \frac{5}{8}$$

51. A bag A contains 2 white and 3 red balls and a bag B contains 4 white and 5 red balls. One ball is drawn at random from one of the bags and it is found to be red. Find the probability that the red ball drawn is from bag.

*Sol:*

Let A and B denote the events of selecting bag A and bag B respectively.

$$\text{Then } P(A) = \frac{1}{2}; P(B) = \frac{1}{2}$$

Let R denote the event of drawing a red ball.

$$\text{Having selected bag A, the probability to draw a red ball from A} = P(R/A) = \frac{3}{5}$$

$$\text{Similarly } P(R/B) = \frac{5}{9}$$

One of the bags is selected at random and from it a ball is drawn at random.

It is found to be red. Then the probability that the selected bag is B

$$= P(B) \cdot P(R/B) = \frac{P(B) \cdot P(R/B)}{P(A) \cdot P(R/A) + P(B) \cdot P(R/B)}$$

$$= \frac{\frac{1}{2} \cdot \frac{5}{9}}{\frac{1}{2} \cdot \frac{3}{5} + \frac{1}{2} \cdot \frac{5}{9}} = \frac{25}{52}$$

52. A factory has two machines. Empirical evidence has established that machines I and II produce 30% and 70% of output respectively. It has also been established that 5% and 1% of the output produced by these machines respectively was defective. A defective item is drawn at random. What is the probability that the defective item was produced by either machine I or machine II ?

*Sol:*

Let  $A_1$  be the event of drawing of an item produced by machine 1.

$A_2$  is the event of drawing an item produced by machine 2

$P(A_1)$  is the probability of getting an item produced by machine 1

$P(A_2)$  is the probability of getting an item produced by machine 2.

$P(B/A_1)$  is the probability of getting defective machine 1

$P(B/A_2)$  is the probability of getting defective machine 2.

From the given data

$$P(A_1) = 0.3 \text{ (30\%)} \quad P(B/A_1) = \frac{5}{100} = 0.05$$

$$P(Q_2) = 0.7 \text{ (70\%)} \quad P(B/A_2) = 1\% = \frac{1}{100} = 0.01$$

Computation of posterior probabilities

Events	$P(A_i)$	$P(B/A_i)$	$P(A \cap B)$	$P(A_i/B)$
$A_1$	$P(A_1) = 0.3$	$P(B/A_1) = 0.05$	$0.3 \times 0.05 = 0.015$	$\frac{0.015}{0.022} = 0.682$
$A_2$	$P(A_2) = 0.7$	$P(B/A_2) = 0.01$	$0.7 \times 0.01 = 0.0007$	$\frac{0.007}{0.022} = 0.318$
	1.00	0.06	0.022	1.000

53. In a bolt factory, the Machines P,Q and R manufacture respectively 25%, 35% and 40% of the total of their outputs 5,4,2 percents respectively are defective bolts. A bolt is drawn at random from the product, and is known to be defective, What are the probabilities that it was manufactured by the machines P,Q and R.

*Sol:*

Let  $P(A)$ ,  $P(B)$ ,  $P(C)$  be the probabilities of the Events that the bolts are manufactured by the machines A, B & C respectively. Then

$$P(A) = \frac{25}{100} = 0.25, \quad P(B) = \frac{35}{100} = 0.35, \quad P(C) = \frac{40}{100} = 0.40$$

Let 'D' denotes that the bolts is defective. Then

$$P(D/A) = \frac{5}{100} = 0.05, \quad P(D/B) = \frac{4}{100}, \quad P(D/C) = \frac{2}{100}$$

- (i) If bolt is defective, then the probability that it is from machine A.

By using Baye's Theorem

$$P(A/D) = \frac{P(D/A)P(A)}{P(D/A)P(A) + P(D/B)P(B) + P(D/C)P(C)}$$

$$= \frac{0.05(0.25)}{0.05(0.25) + 0.04(0.35) + 0.02(0.4)}$$

$$P(A/D) = \frac{0.0125}{0.0125 + 0.014 + 0.008} = \frac{0.0125}{0.0345} = 0.362$$

- (ii) If bolt is defective, then the probability that it is from machine B.

$$P(B/D) = \frac{P(D/B)P(B)}{P(D/A)P(A) + P(D/B)P(B) + P(D/C)P(C)}$$

$$= \frac{0.04(0.35)}{0.05(0.25) + 0.04(0.35) + 0.02(0.40)}$$

$$P(B/D) = \frac{0.014}{0.0345} = 0.4057$$

(iii) It bolt is defective, then the probability that it is from machine C

$$P(C/D) = \frac{P(D/A)P(A) + P(D/B)P(B) + P(D/C)P(C)}{P(D/A)P(A) + P(D/B)P(B) + P(D/C)P(C)}$$

$$\frac{(0.02)(0.40)}{0.05(0.25) + 0.04(0.35) + 0.02(0.40)} = \frac{0.008}{0.0345}$$

$$P(C/D) = 0.2318.$$

**54. A company has two plants for manufacturing scooters. Plant I manufactures 80% of the Scooters and Plant II manufactures 20%. At the Plant I 85% Scooters are rated to be of standard quality and at plant II 65% Scooters are rated to be of standard quality. One Scooter was selected at random. What is the probability that**

**i) It is manufactured by Plant I**

**ii) It is manufactured by Plant II – which is of standard quality.**

*Sol:*

Let  $P(E_1)$  denoted as manufacturing scooters from plant - I.

Probability of plant-I manufacturing scooters = 80% = 0.80.

$$\therefore P(A) = 0.8$$

Let rated  $P\left(\frac{A}{E_1}\right)$  denoted as scooters are rated as standard quality.

Probability of scooters rated as standard quality in plant I = 85% = 0.85.

$$P\left(\frac{A}{E_1}\right) = 0.85$$

Let  $P(E_2)$  denoted as manufacturing scooters from plant - II.

Probability of plant - II manufacturing scooters = 20% = 0.20.

$$\therefore P(E_2) = 0.20.$$

Let  $P\left(\frac{A}{E_2}\right)$  denoted as scooters are rated standard quality in plant II.

Probability of scooters rated as standard quality in plant II is 65% = 0.65.

$$P\left(\frac{A}{E_2}\right) = 0.65$$

(i) It is manufactured by plant I

$$\begin{aligned}
 P\left(\frac{E_1}{A}\right) &= \frac{P(E_1) P\left(\frac{A}{E_1}\right)}{P(E_1) \cdot P\left(\frac{A}{E_1}\right) + P(E_2) \cdot P\left(\frac{A}{E_2}\right)} \\
 &= \frac{0.80 \times 0.85}{(0.80 \times 0.85) + (0.2) \times (0.65)} \\
 &= \frac{0.68}{0.68 + 0.13} \\
 &= \frac{0.68}{0.81}
 \end{aligned}$$

$$\therefore P\left(\frac{E_1}{A}\right) = 0.839.$$

(ii) It is manufactured by Plant - II

$$\begin{aligned}
 P\left(\frac{E_2}{A}\right) &= \frac{P(E_2) \cdot P\left(\frac{A}{E_2}\right)}{P(E_1) \cdot P\left(\frac{A}{E_1}\right) + P(E_2) \cdot P\left(\frac{A}{E_2}\right)} \\
 &= \frac{(0.20) \times (0.65)}{(0.80 \times 0.85) + (0.2)(0.65)} \\
 &= \frac{0.13}{0.68 + 0.13} = \frac{0.13}{0.81} = 0.1604.
 \end{aligned}$$

## Short Question and Answers

### 1. Joint Probability

*Ans :*

A joint probability is the probability of occurrence of two or more simple independent events simultaneously.

In other words, the product of two marginal probabilities occurring together or in succession is called as 'joint probability'.

For example, let  $P(W)$  is the probability of a girl with white complexion and  $P(B)$  is the probability of a girl with black hair. Here,  $P(W)$  and  $P(B)$  are marginal probabilities.

Probability of a girl with white complexion and with black hair together is called as 'joint probability'. Joint probability of these two events can be represented as,  $P(W \cap B)$ .

$$\therefore P(W \cap B) = P(W) \cdot P(B)$$

Where,  $P(W \cap B)$  = Probability of a girl with white complexion and with black hair.

Similarly, joint probability for three independent events A, B and C can be represented as  $P(A \cap B \cap C)$ .

$$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$$

### 2. Marginal Probability

*Ans :*

Marginal probability is the simple probability of the occurrence of an event. It is also called as 'single probability' or 'unconditional probability'.

These two events are independent and do not overlap one another i.e., the events are statistically independent of the outcomes of next coin been tossed

The individual probabilities obtained in this case i.e.,  $P(H)$  or  $P(T)$  are called as marginal probabilities.

### 3. Explain Addition theorem of probability.

*Ans :*

Addition theorem is different for mutually exclusive non-mutually exclusive events.

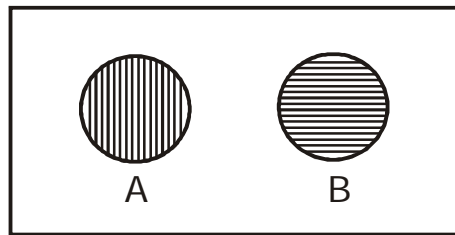
#### (i) For Mutually Exclusive Events

When 'A' and 'B' are two mutually exclusive events (i.e., both cannot occur at the same time) then the probability of occurrence of A or B is equal to the sum of their individual probabilities.

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) \\ \Rightarrow P(A \cup B) &= P(A) + P(B) \end{aligned}$$



Diagrametrically it can be represented as,



Mutually exclusive events

**Figure: Mutually Exclusive Events**

In case of 3 events A, B and C,

$$P(A \text{ or } B \text{ or } C) = P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

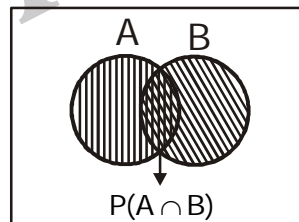
#### (ii) For Non-Mutually Exclusive Events

In case of non-mutually exclusive event (i.e., if the events occur together) there is a variation in the addition theorem.

When 'A' and 'B' are non-mutually exclusive events then the probability of occurrence of A or B is the sum of their individual probability which should be deducted from the probability of A and B occurring together.

$$P(A \text{ or } B) \Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Diagrametrically it can be represented as,



**Fig.: Non-Mutually Exclusive Events**

In case of three non-mutually exclusive events.

A, B and C the probability of occurrence of A or B or C can be calculated by the following formula,

$$P(A \text{ or } B \text{ or } C) = P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

#### 4. Define statistics in singular sense

*Ans :*

##### Statistics in Singular Sense

In singular sense, statistics acts as a helpful device used for the purpose of collecting, classifying, presenting and interpreting the data. It is otherwise known as 'Analytical statistics'.

Some of the definitions of statistics are presented as follows,

**Definitions**

- (i) **According to A.L. Bowley**, "Statistics is the science of counting".
- (ii) **According to A.L. Bowley**, "Statistics may rightly be called the science of averages".
- (iii) **According to Turtle**, "Statistics is a body of principles and techniques of collecting, classifying, presenting, comparing and interpreting the quantitative data".

Thus, the above definitions specify that statistics in singular sense, is a science which comprises different statistical methods used-for collection, organization, classification, presentation and interpretation of data.

**Statistics in Plural Sense**

In plural sense, the term "Statistics" implies systematical collection of numerical facts like statistics of population, production, price-level, national income and so on.

**Definitions**

- (i) **According to 'Dr. A.L. Bowley**, "statistics are numerical statements of facts in any department of enquiry placed in relation to each other".
- (ii) **According to 'L.R. Comor'**, "Statistics are measurements, enumeration or estimates of natural or social phenomena, systematically arranged so as to exhibit their inter-relations".
- (iii) **According to 'Prof. Yule and Prof. Kendall'**, "By statistics we mean quantitative data affected to a marked extent by multiplicity of causes".
- (iv) **According to 'Horace Secrist'**, "By statistics we mean aggregate of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other".

- (v) **According to Webster**, "Statistics are classified facts respecting the conditions of the people in a state - especially those facts which can be stated in numbers or in tables of numbers or in any other tabular or classified arrangement."

**5. Explain the various functions of statistics.**

*Ans :*

The, following are the main functions of statistics:

**1. Presents facts in numerical figures**

The first function of statistics is to present a given problem in terms of numerical figures. We know that the numerical presentation helps in having a better understanding of the nature of a problem. Facts expressed in words are not very useful because they are often vague and are likely to be understood differently by different people. For example, the statement that a large proportion of total work force of India is engaged in agriculture, is vague and uncertain. On the other hand, the statement that 70% of the total work force is engaged in agriculture is more specific and easier to grasp. Similarly, the statement that the annual rate of inflation in a country is 10% is more convincing than the statement that prices are rising.

**2. Presents complex facts in a simplified form**

Generally a problem to be investigated is represented by a large mass of numerical figures which are very difficult to understand and remember. Using various statistical methods, this large mass of data can be presented in a simplified form. This simplification is achieved by the summarization of data so that broad features of the given problem are brought into focus. Various statistical techniques such as presentation of data in the form of diagrams, graphs, frequency distributions and calculation of average, dispersion, correlation, etc., make the given data intelligible and easily understandable.

## 6. limitations of statistics.

*Ans :*

### 1. Statistics deals with numerical facts only

Broadly speaking there are two types of facts:

- (a) quantitative and
- (b) qualitative.

(a) **Quantitative:** Quantitative facts are capable of being represented in the form of numerical figures and therefore, are also known as numerical facts. These facts can be analysed and interpreted with the help of statistical methods.

(b) **Qualitative:** Qualitative facts, on the other hand, represent only the qualitative characteristics like honesty, intelligence, colour of eyes, beauty, etc. and statistical methods cannot be used to study these types of characteristics. Sometimes, however, it is possible to make an indirect study of such characteristics through their conversion into numerical figures. For example, we may assign a number 0 for a male and 1 for a female, etc.

### 2. Statistics deals only with groups and not with individuals

Statistical studies are undertaken to study the characteristics of a group rather than individuals. These studies are done to compare the general behaviour of the group at different points of time or the behaviour of different groups at a particular point of time. For example, the economic performance of a country in a year is measured by its national income in that year and by comparing national income of various years, one can know whether performance of the country is improving or not. Further, by comparing national income of different countries, one can know its relative position vis-a-vis other countries.

### 3. Statistical results are true only on the average

Statistical results give the behaviour of the group on the average and these may not hold for an individual of that very group. Thus, the statement that average wages of workers of a certain factory is ₹ 1,500 p.m. does not necessarily mean that each worker is getting this wage. In fact, some of the workers may be getting more while others less than or equal to ₹ 1,500. Further, when value of a variable is estimated by using some explanatory variable, the estimated value represents the value on the average for a particular value of the explanatory variable. In a similar way, all the laws of statistics are true only on the average.

### 4. Statistical results are only approximately true

Most of the statistical studies are based on a sample taken from the population. Under certain circumstances the estimated data are also used. Therefore, conclusions about a population based on such information are to be true only approximately.

## 7. Measures of Central Tendency.

*Ans :*

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

**Definitions**

- (i) "A measure of central tendency is a typical value around which other figures congregate".
- (ii) "An averages stands for the whole group of which forms a part yet represents the whole".
- (i) "One of the most widely used set of summary figures is known as measures of location".

**8. Arithmetic mean.***Ans :***Meaning**

Arithmetic Average (or) Mean of a series is the figure obtained by dividing the total "Values of the various items by their number. In other words it is the sum of the values divided by their number. Arithmetic means is the most widely used measure of central tendency.

**9. Define mode.***Ans :***Meaning**

Mode may be defined as the value that occurs most frequently in a statistical distribution or it is defined as that exact value in the ungrouped data if each sample which occurs most frequently.

**Definitions**

- (i) **According to Croxton and Cowden**, "The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded as the most typical of a series of values."
- (ii) **According to A.M. Tuttle**, "Mode is the value which has the greatest frequency density in its immediate neighbourhood."
- (iii) **According to Zizek**, "The mode is the value occurring most frequently in a series of items and around which the other items are distributed most densely."

**10. Define median.***Ans :***Meaning**

If a group of N observations is arranged in ascending or descending order of magnitude, then the middle value is called median of these observations and is denoted by M.

That is,  $M = \frac{N+1}{2}$  th observation.

**Definition**

**According to Croxton and Cowden**, "The median is that value which divides a series so that one half or more of the items are equal to or less than it and one half or more of the items are equal to or greater than it."

**Characteristics**

- Unlike the arithmetic mean, the median can be computed from open-ended distributions. This is because it is located in the median class-interval, which would not be an open-ended class.
- The median can also be determined graphically whereas the arithmetic mean cannot be ascertained in this manner.
- As it is not influenced by the extreme values, it is preferred in case of a distribution having extreme values.

**11. Define Dispersion.***Ans :***Meaning**

The concept of dispersion is related to the extent of scatter or variability in observations. The variability, in an observation, is often measured as its deviation from a central value. A suitable average of all such deviations is called the measure of dispersion. Since most of the measures of dispersion are based on the average of deviations of observations from an average, they are also known as the averages of second order.

### Definitions

As opposed to this, the measures of central tendency are known as the averages of first order. Some important definitions of dispersion are given below:

- (i) **According to A.L. Bowley**, "Dispersion is the measure of variation of the items."
- (ii) **According to Connor**, "Dispersion is the measure of extent to which individual items vary."
- (iii) **According to Simpson and Kafka**, "The measure of the scatteredness of the mass of figures in a series about an average is called the measure of variation or dispersion."

### 12. What is Range?

*Ans :*

Range is the simplest method of studying dispersion. It is defined as the difference between the value of the smallest item and the value of the largest item included in the distribution. Symbolically,

$$\text{Range} = L - S$$

where

L = Largest item, and

S = Smallest item.

The relative measure corresponding to range, called the coefficient of range, is obtained by applying the following formula :

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

If the averages of the two distributions are about the same, a comparison of the range indicates that the distribution with the smaller range has less dispersion, and the average of that distribution is more typical of the group.

### 13. What is Quartile Deviations?

*Ans :*

It is based on two extreme items and it fails to take account of the scatter within the range. From

this there is reason to believe that if the dispersion of the extreme items is discarded, the limited range thus established might be more instructive.

For this purpose there has been developed a measure called the interquartile range, the range which includes the middle 50 per cent of the distribution. That is, one quarter of the observations at the lower end, another quarter of the observations at the upper end of the distribution are excluded in computing the interquartile range. In other words, interquartile range represents the difference between the third quartile and the first quartile.

$$\text{Inter quartile range} = Q_3 - Q_1$$

Very often the interquartile range is reduced to the form of the Semi-interquartile range or quartile deviation by dividing it by 2.

$$\text{Quartile Deviation or Q.D.} = \frac{Q_3 - Q_1}{2}$$

### Coefficients of Quartile Deviation

Quartile deviation is an absolute measure of dispersion. Its relative measure is the coefficient of Quartile deviation.

Coefficients of Quartile Deviation

$$= \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}}$$

$$= \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

It is used to compare the degree of variation in different series.

### 14. Define mean deviation.

*Ans :*

#### Meaning

"Mean Deviation of a series is the arithmetic average of the deviations of various items from a measure of central tendency (either mean, median

or mode)". Theoretically, deviations can be taken from any of the three averages mentioned above, but in actual practice it is calculated either from mean or from Median. While Calculating deviations algebraic signs are not taken into account.

**15. What is standard deviation?**

*Ans :*

The standard deviation concept was introduced by Karl Pearson in 1823. It is by far the most important and widely used measure of studying dispersion. Its significance lies in the fact that it is free from those defects from which the earlier methods suffer and satisfies most of the properties of a good measure of dispersion. Standard deviation is also known as root mean square deviation for the reason that it is the square root of the mean of the squared deviation from the arithmetic mean. Standard deviation is denoted by the small Greek letter  $\sigma$  (read as sigma).

The standard deviation measures the absolute dispersion (or variability of distribution; the greater the amount of dispersion or variability), the greater the standard deviation, the greater will be the magnitude of the deviations of the values from their mean. A small standard deviation means a high degree of uniformity of the observation as well as homogeneity of a series; a large standard deviation means just the opposite.

**Merits**

1. It is rigidly defined
2. It takes into account every single value in the series
3. It is amenable to further algebraic or statistical treatment.
4. It is extensively used in various other statistical calculations such as correlation, regression, sampling etc

**Demerits**

1. It is relatively difficult to compute
2. It is calculated with only Arithmetic Mean as the average. Standard deviation from other averages such as Median is not an effective measure of dispersion.

**16. Define Coefficient of Variation.**

*Ans :*

Coefficient of Variation (CV) was proposed by Karl Pearson. It is used to compare the variability of two (or) more distributions. A distribution with greater C.V. is considered as more variable or less consistent, less unit form, less stable or less homogeneous distribution and the distribution with less C.V is considered as less variable or more consistent, more uniform, more stable or more homogeneous distribution.

$$\text{Coefficient of variation} = \frac{\text{Standard deviation}}{\text{Arithmetic mean}} \times 100$$

**17. Skewness.**

*Ans :*

**Introduction**

The word skewness refers to lack of symmetry. Non-normal or asymmetrical distribution is called skew distribution. Two frequency distributions may have the same mean and standard deviation and yet may differ with respect to another characteristics- the skewness or, asymmetry of the distribution. Any measure of skewness indicates the difference between the manner in which items are distributed in a particular distribution compared with a normal distribution. Lack of symmetry or skewness in frequency distributions is due to the existence of a longer tail on one side (either to the left or to the right), which has no counterpart on the other side. If the larger tail is on the right, we say that the distribution is positively skewed; whereas if the longer tail is on the left side.

**Definitions**

Some important definitions of skewness are as follows :

- **According to Corxton & Cowden** "When a series is not symmetrical it is said to be asymmetrical or skewed."
- **According to Morris Hamburg** "Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution."
- **According to Simpson & Kafka** "Measures of skewness tell us the direction and the extent of skewness. In symmetrical distribution the mean, median and mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness."
- **According to Garrett** "A distribution is said to be 'skewed' when the mean and the median fall at different points in the distribution, and the balance (or centre of gravity) is shifted to one side or the other – to left or right."

**18. Define kurtosis.**

*Ans :*

**Meaning**

Kurtosis in Greek means "bulginess". In statistics kurtosis refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. The degree of kurtosis of a distribution is measured relative to the peakedness of normal curve.

**(i) Leptokurtosis**

Measures of kurtosis tell us the extent to which a distribution is more peaked or flat-topped than the normal curve. If a curve is more peaked than the normal curve, it is called 'leptokurtic'. In such a case items are more closely bunched around the mode.

**(ii) Mesokurtosis**

If a curve is more flat-topped than the normal curve, it is called 'platykurtic'. The normal curve itself is known as 'mesokurtic'.

The condition of peakedness or flat-topped ness itself is known as kurtosis of excess. The concept of kurtosis is rarely used in elementary statistical analysis.

**19. Define the term probability?**

*Ans :*

An Italian mathematician, Galileo (1564 - 1642), attempted a quantitative measure of probability while dealing with some problems related to gambling. In the middle of 17th Century, two French mathematicians, Pascal and Fermat, laid down the first foundation of the mathematical theory of probability while solving the famous 'Problem of Points' posed by Chevalier-De-Mere. Other mathematicians from several countries also contributed in no small measure to the theory of probability. Outstanding of them were two Russian mathematicians, A. Kintchine and A. Kolmogoroff, who axiomised the calculus of probability.

If an experiment is repeated under similar and homogeneous conditions, we generally come across two types of situations.

- (i) The net result, what is generally known as 'outcome' is unique or certain.
- (ii) The net result is not unique but may be one of the several possible outcomes.

The situations covered by :

- (i) are known as 'deterministic' or 'predictable' and situations covered by
- (ii) are known as 'probabilistic' or 'unpredictable'.

'Deterministic' means the result can be predicted with certainty. For example, if  $r$  is the radius of the sphere then its volume is given by  $V = \frac{4}{3}\pi r^3$  which gives uniquely the volume of the sphere.

There are some situations which do not lend themselves to the deterministic approach and they are known as 'Probabilistic'.

**For example,** by looking at the sky, one is not sure whether the rain comes or not.

In such cases we talk of chances or probability which can be taken as a quantitative measure of certainty.



## Exercise Problems

1. Calculate the arithmetic mean from the following data :

Age (Years)	18 – 21	22 – 25	26 – 35	36 – 45	46 – 55
No. of Employees	8	32	54	36	20

**[Ans: 28.91]**

2. Calculate the arithmetic average from the following data :

Value	Frequency
Less than 10	4
Less than 20	16
Less than 30	40
Less than 40	76
Less than 50	96
Less than 60	112
Less than 70	120
Less than 80	125

**[Ans: 37.58]**

3. The mean weight of 150 students of B.Com. Class is 60 Kgs. The mean weight of boys is 70 Kgs. And that of girls is 55 Kgs. Find the number of boys and girls in the class.

**[Ans: Boys 50, Girls 100]**

4. The average Salary of Male Employees in a firm was Rs. 5,200 and that of Females was Rs. 4,200. The Average Salary of all employees was Rs. 5,000. Find out percentage of Male and Female Employees.

**[Ans: Males-80%, Females-20%]**

5. For the following Data, Compute Median.

Class Interval	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
Frequency	15	100	170	120	40

**[Ans: 26.32]**

6. Compute median from the following data,

Mid – Values	115	125	135	145	155	165	175	185	195
Frequency	6	25	48	72	116	60	38	22	3

**[Ans: 153.8]**

7. From the following data of weight of 122 persons, determine the modal weight.

Weight (in lbs)	No. of Persons
100 – 110	4
110 – 120	6
120 – 130	20
130 – 140	32
140 – 150	33
150 – 160	17
160 – 170	8
170 – 180	2

**[Ans: Modal Weight = 140.59 lbs]**

8. The following table given the marks secured by 60 students of a class.

Marks	No. of Students (f)
10 – 20	8
20 – 30	12
30 – 40	20
40 – 50	10
50 – 60	7
60 – 70	3

Calculate the arithmetic mean and geometric mean.

**[Ans: A.M = 32.50, G.H - 33.33]**

9. From the following data compute the value of harmonic mean.

Marks	10	20	30	40	50
No. of Students	20	30	50	15	5

**[Ans: H.M = 21.257]**

10. If the mode and mean of a moderately asymmetrical series are respectively 16 inches and 15.6 inches, what would be its most probable median ?

**[Ans: Median = 15.73]**

11. Calculate median and mean Deviation for the following frequency distribution.

Age (years)	No. of Persons	Age (years)	No. of persons
1 - 5	7	26 - 30	18
6 - 10	10	31 - 35	10
11 - 15	16	36 - 40	5
16 - 20	32	36 - 40	1
21 - 25	24		

**[Ans : Median = 19.95, M.D. = 7, 103]**

12. Calculate the standard deviation from the following data:

Marks in Cost Accounting	No. of Students
0 - 10	5
10 - 20	7
20 - 30	14
40 - 50	9
50 - 60	6
60 - 70	2

**[Ans :  $\sigma = 15.57$ ]**

13. Calculate coefficient of variation and Karl Pearson's Coefficient of Skewness.

Income (Rs.) Below	500	600	700	800	900	1000
No. of Employees	8	20	40	70	87	100

**[Ans : C.V. = 19.64, Co. eff. of SK = .0.13]**

14. Calculate coefficient of skewness based on quartiles from the following data.

Size	Frequency
10 - 20	6
20 - 30	10
30 - 40	18
40 - 50	30
50 - 60	12
60 - 70	10
70 - 80	6
80 - 90	2

**[Ans : 0.04]**

15. Calculate Bowley's coefficient of skewness of the following data:

Weight (lbs)	No. of Persons
Under 100	1
100 - 109	14
110 - 119	66
120 - 129	122
130 - 139	145
140 - 149	121
150 - 159	65
160 - 169	31
170 - 179	12
180 - 189	5
190 - 199	2
200 and over	2

[Ans: + 0.023]

Rahul Publications

## Choose the Correct Answers

1.  $P(E) =$  [ a ]
  - (a)  $\frac{a}{a+b}$
  - (b)  $\frac{a}{a \times b}$
  - (c)  $a \times b$
  - (d)  $a + b$
  
2. The formula for mode = [ b ]
  - (a)  $L_1 + \frac{\Delta}{\Delta + \Delta} \times i$
  - (b)  $L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$
  - (c)  $L_1 + \frac{\Delta_1}{\Delta_2} \times i$
  - (d)  $L + \frac{\Delta_1 + \Delta_2}{\Delta} \times i$
  
3. \_\_\_\_\_ events do not occur more often than the orders [ c ]
  - (a) Mutually exclusive events
  - (b) Compound events
  - (c) Equally likely events
  - (d) Dependent events
  
4. The additive laws of probability for mutually exclusive is [ b ]
  - (a)  $P(A \cap B) = P(A) + P(B)$
  - (b)  $P(A \cup B) = P(A) + P(B)$
  - (c)  $P(A) \cup P(B) = P(A) + P(B)$
  - (d)  $P(A \cup B) = P(B) + P(A \cap B)$
  
5. Baye's theorem (Rule for the inverse probability) is given by [ a ]
  - (a)  $P(A_i/E) = \frac{P(A_i) \cdot P(E/A_i)}{\sum_{i=1}^n P(A_i) P(E/A_i)}$
  - (b)  $P(A \cup E) = \frac{P(A_i) \cdot P(E/A_i)}{\sum_{i=1}^n P \cdot A}$
  - (c)  $P(E) = \frac{P(A) \cdot P(E)}{P(E/A_i)}$
  - (d)  $P(E/A) = \frac{P(A) \cdot P(A/E)}{\sum_{i=j}^n P(A_{ij}) \cdot P(E)}$
  
6. Coefficient of range is, [ d ]
  - (a)  $L - S$
  - (b)  $L + S$
  - (c)  $\frac{L + S}{L - S}$
  - (d)  $\frac{L - S}{L + S}$

7. The most important measure of kurtosis is the value of the coefficient  $\beta_2$ , which is defined as, [ c ]
- (a)  $\beta_2 = \mu^4 \cdot \mu_2$  (b)  $\beta_2 = \gamma_2 - \mu_2$
- (c)  $\beta_2 = \frac{\mu^4}{\mu_2^2}$  (d)  $\beta_2 = \frac{\mu_2^2}{\mu^4}$
8.  $SK_p =$  \_\_\_\_\_. [ a ]
- (a)  $\frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$  (b)  $3 \text{ Median} - 2 \text{ Mean}$
- (c)  $\text{Standard deviation} \times \frac{\text{Mean}}{\text{Mode}}$  (d)  $\frac{\text{Mode} - \text{Median}}{\text{Standard deviation}}$
9. Quartile Deviation = [ c ]
- (a)  $\frac{Q_3 - Q_1}{Q_3 + Q_1}$  (b)  $\frac{Q_3 + Q_1}{Q_1 + Q_2}$
- (c)  $\frac{Q_3 - Q_1}{2}$  (d)  $\frac{Q_3 - Q_1}{4}$
10. A \_\_\_\_\_ is a characteristic that takes different values at different times, places or situations. [ d ]
- (a) Attributes (b) Data
- (c) Statistics (d) Variable

## Fill in the blanks

1. \_\_\_\_\_ is the use of data which helps the decision - maker to make good decisions.
2. The formula for median = \_\_\_\_\_
3. The empirical mode = \_\_\_\_\_
4. \_\_\_\_\_ curve is used to show the degree of inequality in the distribution of income and wealth.
5. Bayes theorem is also called as \_\_\_\_\_ as it helps in determining cause of the probability of a particular effect.
6. A \_\_\_\_\_ probability is the probability of occurrence of two or more simple events.
7. The outcomes of random experiment are \_\_\_\_\_
8. \_\_\_\_\_ approach of probability assumes that all the possible outcomes of an experiment are mutually exclusive and equally likely.
9. \_\_\_\_\_ theorem states that if two events A and B are independent, the probability that they both will occur is equal to the product of their individual probabilities.
10. \_\_\_\_\_ theorem helps in calculating or revising the probabilities in the light of additional information.

### ANSWERS

1. Statistics
2.  $L + \frac{N/2 - F}{f} \times C$
3. 3 Median – 2 Mean
4. Lorenz
5. 'Probability of Causes'
6. Joint
7. Events
8. Classical approach or A prior approach
9. Multiplication theorem of probability
10. Baye's

# UNIT II

- (i) **Probability Distribution:** Random Variable (RV), Expectation and Variance of a RV. Probability distribution, function, properties, Continuous and Discrete Probability distribution functions.
- (ii) **Discrete Probability distributions:** Binomial Distribution, Properties and applications; Poisson distribution, properties and applications.
- (iii) **Continuous Probability Distributions:** Normal Distribution, Standard Normal Distribution properties, applications and importance of Normal Distribution.

## 2.1 PROBABILITY DISTRIBUTION

**Q1. Define the term Probability Distribution.**

*Ans :*

The probability distribution describes the range of possible values that a random variable can attain and the probability that the value of the random variable is within any (measurable) subset of that range.

When the random variable takes values in the set of real numbers, the probability distribution is completely described by the cumulative distribution function, whose value at each real  $x$  is the probability that the random variable is smaller than or equal to  $x$ .

A probability distribution is a graph, table, or formula that gives the probability for each value of the random variable.

If  $x$  is a random variable then denote by  $P(x)$  to be the probability that  $x$  occurs. It must be the case that  $0 \leq P(x) \leq 1$  for each value of  $x$  and  $\sum P(x) = 1$  (the sum of all the probabilities is 1).

### 2.1.1 Random Variable (RV)

**Q2. Define Random Variable. Explain different types of Random Variables.**

*Ans :*

A variable that can many real values which are determined by the outcomes of a random experiment on a real line  $(-\infty, +\infty)$ . It is a chance or Stochastic Variable.

It is a function defined on the sample space, "S" of a random experiment. A Random variable takes different values as a result of the outcomes of a random experiment. A Random variable can be Discrete or Continuous.

**For example,**

- i) A random variable can also be used to describe the process of rolling a fair die and the possible outcomes.

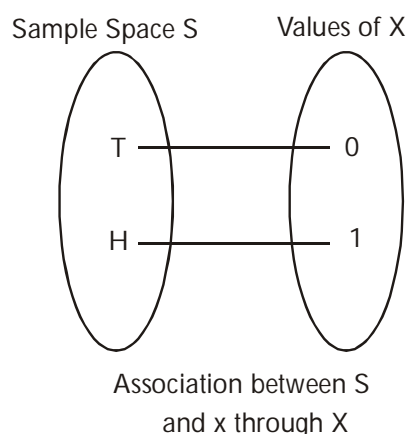
The most obvious representation is to take the set  $\{1, 2, 3, 4, 5, 6\}$  as the sample space, defining the random variable  $X$  as the number rolled. In this case,

$$X = \begin{cases} 1, & \text{if } a_1 \text{ is rolled,} \\ 2, & \text{if } a_2 \text{ is rolled,} \\ 3, & \text{if } a_3 \text{ is rolled,} \\ 4, & \text{if } a_4 \text{ is rolled,} \\ 5, & \text{if } a_5 \text{ is rolled,} \\ 6, & \text{if } a_6 \text{ is rolled,} \end{cases}$$

$$p_X(x) = \begin{cases} \frac{1}{6}, & \text{if } x = 1, 2, 3, 4, 5, 6 \\ 0, & \text{otherwise} \end{cases}$$

- ii) Consider the experiment of tossing a single coin. Sample space  $S = \{H, T\}$ . Let  $X$  represent the number of heads. Thus  $X$  can assume the values 0 and 1. Let  $x$  be these values i.e.  $x = 0, 1$ . The relationship between the sample space and the values of  $X$  is shown in figure 2.1.





**Fig.: Outcomes on Tossing a Single Coin and Values of X**

### Example

Suppose the random experiment is the toss of a coin 3 times. The sample space consists of the 8 outcomes:

$\Omega = \{(HHH) (HHT) (HTH) (HTT) (THH) (THT) (TTH) (TTT)\}$ . Find the number of heads in these three tosses.

*Sol :*

Generally, rather than being interested in these outcomes, one is more concerned with the number of heads (or tails) in the 3 tosses. Thus one is interested in the number of heads being one, not whether the outcome was (HTT), (THT) or (TTH). That is, one is interested in the values of the random variable  $X$  = number of heads in the 3 tosses, where  $X$  can take on the values 0, 1, 2, or 3; i.e., the set of outcomes of interest is the set  $\{0, 1, 2, 3\}$ .

A random variable is denoted by a capital letter (e.g.,  $X, Y, W$ ) and the values of the random variable by the corresponding lowercase letter (e.g.,  $x, y, w$ ). The set of all possible values of a random variable defines a new sample space. As before, an event is a subset of the sample space. For the random variable, the sample space is  $\Omega_x = \{0, 1, 2, 3\}$ . The event that the value of  $X$  is either 0 or 1 is the subset

$$E_1 = \{0, 1\}. \text{ This is denoted by } X \in E_1$$

### Types

Random variables can be of two types as follows:

#### 1. Discrete Random Variables

A discrete random variable is one which may take on only a countable number of distinct values such as 0, 1, 2, 3, 4, .... Discrete random variables are usually (but not necessarily) counts. If a random variable can take only a finite number of distinct values, then it must be discrete. Examples of discrete random variables include the number of children in a family, the Friday night attendance at a cinema, the number of patients in a doctor's surgery, the number of defective light bulbs in a box of ten.

The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the probability mass function.

Suppose a random variable  $X$  may take  $k$  different values, with the probability that  $X = x_i$  defined to be  $P(X = x_i) = p_i$ . The probabilities  $p_i$  must satisfy the following:

- i)  $0 \leq p_i \leq 1$  for each  $i$ ,
- ii)  $p_1 + p_2 + \dots + p_k = 1$ .

Discrete random variables take on integer values, usually the result of counting.

### Examples

- i) Suppose that one flips a coin and count the number of heads. The number of heads results from a random process - flipping a coin. And the number of heads is represented by an integer value a number between 0 and plus infinity. Therefore, the number of heads is a discrete random variable.
- ii) Suppose a variable  $X$  can take the values 1, 2, 3, or 4. The probabilities associated with each outcome are described by the following table:

Outcome	1	2	3	4
Probability	0.1	0.3	0.4	0.2

The probability that  $X$  is equal to 2 or 3 is the sum of the two probabilities:  $P(X = 2 \text{ or } X = 3) = P(X = 2) + P(X = 3) = 0.3 + 0.4 = 0.7$ . Similarly, the probability that  $X$  is greater than 1 is equal to  $1 - P(X = 1) = 1 - 0.1 = 0.9$ , by the complement rule.

## 2. Continuous Random Variables

A continuous random variable is one which takes an infinite number of possible values. Continuous random variables are usually measurements. For example height, weight, the amount of sugar in an orange, the time required to run a mile all can take infinite number of possible values.

A continuous random variable is not defined at specific values. Instead, it is defined over an interval of values, and is represented by the area under a curve (in advanced mathematics, this is known as an integral) The probability of observing any single value is equal to 0, since the number of values which may be assumed by the random variable is infinite.

Suppose a random variable  $X$  may take all values over an interval of real numbers. Then the probability that  $X$  is in the set of outcomes  $A$ ,  $P(A)$  is defined to be the area above  $A$  and under a curve.

The curve, which represents a function  $p(x)$ , must satisfy the following:

- i) The curve has no negative values ( $p(x) \geq 0$  for all  $x$ ).
- ii) The total area under the curve is equal to 1.

A curve meeting these requirements is known as a density curve.

### 2.1.2 Expectation and Variance of a RVs

#### Q3. Explain the expectations of random variables.

*Ans :*

#### Expectation of Random Variable

The arithmetic mean of a discrete random variable of the probability distribution is called as expectation of random variable' or 'expected value of the random variable.

Let 'X' be the discrete random variable with the corresponding probability distribution P(X), then the expected value is given by,

$$E(X) = \sum X.P(X)$$

$$\mu = \sum X.P(X)$$

Expected value is usually represented by a greek letter ' $\mu$ '.

In other words, the sum of the product of the random variables and its corresponding probabilities is called as 'expected value of the random variable'.

If a discrete random value 'X' has ' $x_1$ ', ' $x_2$ ', ' $x_3$ ', ..... ' $x_n$ ' as all the possible values with corresponding probabilities  $P(x_1)$ ,  $P(x_2)$ ,  $P(x_3)$ ,  $P(x_n)$ , then the expected value is given by,

$$E(X) = x_1 P(x_1) + x_2 P(x_2) + x_3 P(x_3) + \dots + x_n P(x_n) \\ = \mu$$

#### Q4. Explain the variance of Random Variable.

*Ans :*

The variance of a random variable tells us something about the spread of the possible values of the variable.

Let X be a random variable with probability distribution f(x) and mean  $\mu$ .

The variance of X is :

$$\sigma^2 = E[(X - \mu)^2] = \sum_x (x - \mu)^2 f(x), \text{ if X is discrete, and}$$

$$\sigma^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx, \text{ if X is continuous}$$

The positive square root of the variance,  $\sigma$ , is called the standard deviation of X.

For a discrete random variable X, the variance of X is written as Var (X).

$$\text{Var}(X) = E[(X - m)^2]$$

Where, m is the expected Value E(X)

This can also be written as :

$$\text{Var}(X) = E(X^2) - m^2$$

The standard deviation of X is the square root of Var (X)

The variance does not behave in the same way as expectation when we multiply and add constants to random variables. In fact :

$$\text{Var}[aX + b] = a^2 \text{Var}(X)$$

Since,

$$\begin{aligned} \text{Var}[aX + b] &= E[(aX + b)^2] - (E[aX + b])^2 \\ &= E[a^2X^2 + 2abX + b^2] - (aE(X) + b)^2 \\ &= a^2 E(X^2) + 2abE(X) + b^2 - a^2 E^2(X) - 2abE(X) - b^2 \\ &= a^2 E(X^2) - a^2 E^2(X) = a^2 \text{Var}(X) \end{aligned}$$

## 2.2 PROBABILITY DISTRIBUTION FUNCTION

**Q5. Define the term Probability Distribution Function.**

*Ans :*

Probability distribution is a set of probabilities of all the possible outcomes of a random experiment.

Probability distribution is similar to frequency distribution. Probability distribution is based on the theoretical considerations, subjective assessment or on experience.

For example, 'X' is a random variable which can take the values  $x_1, x_2, x_3, \dots$

The probabilities associated with each of the possible values of 'X',  $P(X = x_i) = P_i (i = 1, 2, 3, \dots)$

Therefore, the collection of pairs  $(x_i, P_i)$ , where  $i = 1, 2, 3, \dots$  is called probability distribution of random variable X.

The values of probability distribution are usually tabulated as given below.

$X = x_i$	$P(X = x_i)$
$x_1$	$P_1$
$x_2$	$P_2$
$x_3$	$P_3$
—	—
—	—
—	—
—	—
$x_n$	$P_n$

Probability distribution functions are usually represented as  $f(x)$ ,  $g(x)$ ,  $h(x)$  etc. Here  $f(x)$  is the function where  $f(x) = P(X = x)$  which assigns probability to each possible outcome 'x', hence called as 'probability distribution'.

### 2.2.1 Properties

**Q6. What are the Properties of probability distribution function.**

*Ans :*

#### 1. Cumulative Distributions

As it is desired to know the probability of finding x between certain limits, e.g.,  $P(x_1 \leq x \leq x_2)$ . This is given by the cumulative or integral distribution,

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} P(x) dx \quad \dots (1)$$

Where we have assumed  $P(x)$  to be continuous. If  $P(x)$  is discrete, the integral is replaced by a sum,

$$P(x_1 \leq x \leq x_2) = \sum_{i=1}^2 P(x_i) \quad \dots (2)$$

By convention, also, the probability distribution is normalized to 1, i.e.,

$$\int P(x) dx = 1 \quad \dots (3)$$

If x is continuous or,

$$\sum_i P(x_i) = 1 \quad \dots (4)$$

if x is discrete. This simply means that the probability of observing one of the possible outcomes in a given trial is defined as 1. It follows then that  $P(x_i)$  or  $\int P(x) dx$  cannot be greater than 1 or less than 0.

#### 2. Expectation Values

If x is a random variable distributed as  $P(x)$ , then,

$$E[x] = \int xP(x) dx \quad \dots (1)$$

is the expected value of x. The integration in equation (1) is over all admissible x. This, of course, is just the standard notion of an average value. For a discrete variable, equation (1) becomes a sum,

$$E[x] = \sum_i x_i P(x_i) \quad \dots (2)$$

Similarly, if  $f(x)$  is a function of  $x$ , then,

$$E[f(x)] = \int f(x)P(x)dx \quad \dots (3)$$

is the expected value of  $f(x)$ .

### 3. Distribution Moments - The Mean and Variance:

A probability distribution may be characterized by its moments. The  $r^{\text{th}}$  moment of  $x$  about some fixed point  $x_0$  is defined as the expectation value of  $(x - x_0)^r$  where  $r$  is an integer. An analogy may be drawn here with the moments of a mass distribution in mechanics. In such a case,  $P(x)$  plays the role of the mass density.

In practice, only the first two moments are of importance. And, indeed, many problems are solved with only a knowledge of these two quantities. The most important is the first moment about zero,

$$\mu = E[x] = \int xP(x)dx \quad \dots (1)$$

This can be recognized as simply the mean or average of  $x$ . If the analogy with mass moments is made, the mean thus represents the "center of mass" of the probability distribution.

It is very important here to distinguish the mean as defined in equation (1) from the mean which one calculates from a set of repeated measurements. The first refers to the theoretical mean, as calculated from the theoretical distribution, while the latter is an experimental mean taken from a sample.

The second characteristic quantity is the second moment about the mean (also known as the second central moment).

$$\sigma^2 = E[(x - \mu)^2] = \int (x - \mu)^2 P(x)dx \quad \dots (2)$$

This is commonly called the variance and is denoted as  $\sigma^2$ . The square root of the variance,  $\sigma$ , is known as the standard deviation. As can be seen from equation (2), the variance is the average squared deviation of  $x$  from the mean. The standard deviation,  $\sigma$ , thus measures the dispersion or width of the distribution and gives us an idea of how much the random variable  $x$  fluctuates about its mean. Like  $\mu$ , equation (2) is the theoretical variance and should be distinguished from the sample variance.

Further moments, of course, may also be calculated, such as the third moment about the mean. This is known as the skewness and it gives a measure of the distribution's symmetry or asymmetry. It is employed on rare occasions, but very little information is generally gained from this moment or any of the following ones.

### 4. Covariance

In the more general case, the outcomes of a process may be characterized by several random variables,  $x, y, z, \dots$ . The process is then described by a multivariate distribution  $P(x, y, z, \dots)$ . For example, a playing card which is described by two variables - its denomination and its suit.

For multivariate distributions, the mean and variance of each separate random variable  $x, y, \dots$  are defined in the same way as before (except that the integration is over all variables). In addition a third important quantity must be defined:

$$\text{cov}(x, y) = E[(x - \mu_x)(y - \mu_y)] \quad \dots (1)$$

where  $\mu_x$  and  $\mu_y$  are the means of  $x$  and  $y$  respectively.

Equation (1) is known as the covariance of  $x$  and  $y$  and it is defined for each pair of variables in the probability density. Thus, if we have a trivariate distribution  $P(x, y, z)$ , there are three covariance -  $\text{cov}(x, y)$ ,  $\text{cov}(x, z)$  and  $\text{cov}(y, z)$ .

## 2.3 TYPES OF PROBABILITY DISTRIBUTION FUNCTION

**Q7. What are the different types of probability distribution function.**

*Ans :*

There are two types of probability distribution function :

1. Discrete Probability Distribution Function
2. Continuous Probability Distribution Function.

**1. Discrete Probability Distribution Function:** The Discrete Probability Distributions are :

- i) Hyper - geometric distribution

- ii) Binomial distribution
- iii) Poisson distribution
- iv) Geometric distribution

**2. Continuous Probability Distribution Function :** The Continuous Probability distributions are :

- i) Uniform distribution
- ii) Normal distribution
- iii) Exponential distribution.
- iv) Beta distribution

### 2.4 DISCRETE PROBABILITY DISTRIBUTION FUNCTIONS

**Q8. Explain the concept of Discrete Probability Distribution Function.**

*Ans :*

Discrete probability distribution is the probability distribution of a discrete random variable 'X' which takes only a finite number of variables. Random variable 'X' as function  $f(x)$  satisfies the following conditions,

- i)  $f(x) \geq 0$
- ii)  $\sum f(x) = 1$
- iii)  $P(X = x) = f(x)$

The examples of discrete probability are binomial and Poisson distribution.

#### 2.4.1 Binomial Distribution

**Q9. What is Binomial Distribution? State the assumptions of Binomial Distribution.**

*Ans :* (June-19, Imp.)

In probability theory and statistics, the binomial distribution is the discrete probability distribution of the number of successes in a sequence of  $n$  independent yes/no experiments, each of which yields success with probability  $p$ . Such a success/failure experiment is also called a Bernoulli experiment or Bernoulli trial.

In fact, when  $n = 1$ , the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance.

This binomial distribution is also known as the Bernoulli distribution by the name of the Swiss Mathematician Jacob Bernoulli who has derived it.

The binomial probability refers to the probability that a binomial experiment results in exactly  $x$  successes.

#### Definition

If an event  $E$  has probability  $p$  of occurring in each of  $n$  independent trials and that of failure in any trial is  $q$  ( $= 1 - p$ ) then the probability that it will occur exactly  $r$  times in  $n$  trials is given by:

$$p(r) = {}^nC_r p^r q^{n-r}$$

This probability distribution is called the binomial probability distribution. The binomial distribution is a discrete distribution with parameters  $n$  and  $p$ . If  $p, q$  are equal it is symmetrical, otherwise it is nonsymmetrical.

Where

$p$  = probability of success in a single trial.

$q$  = probability of failure a single trial.

$p + q = 1$ :

$n$  = Number of trials

$r$  = Number of success in ' $n$ ' times.

#### Assumptions

1. Each trial has two mutually exclusive possible outcomes, i.e., success or failure.
2. Each trial is independent of other trials.
3. The probability of a success (say  $p$ ) remains constant from trial to trial.
4. The probability of getting a head in a toss of coins is  $\frac{1}{2}$ . This result must remain same in successive tosses.
5. The number of trials is fixed.

### 2.4.1.1 Properties

#### Q10. State the Properties of Binomial Distribution.

*Ans :*

(June-19, Imp.)

The properties of binomial distribution are as follows,

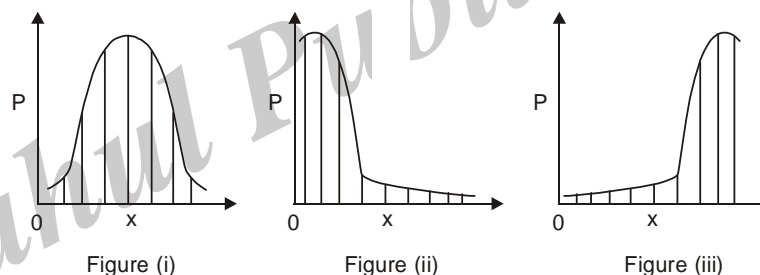
1. It describes the distribution of probabilities when there are only two mutually exclusive outcomes for each trial of an experiment for example while tossing a coin, the two possible outcomes are head and tail.
2. The process is performed under identical conditions for 'n' number of times.
3. Each trial is independent of other trials.
4. The probability of success 'p' remains same for trial to trial throughout the experiment and similarly, the probability of failure ( $q = 1 - p$ ) also remains constant overall the observations.
5. Binomial distribution is symmetrical when  $p = 0.5$  [figure (i) ] and it is skewed if  $p \neq 0.5$ , where 'n' can be any value.

When  $p > 0.5$  [figure (iii)], it is skewed to the right  $\rightarrow$  negatively skewed.

When  $p < 0.5$  (ii), ] it is skewed to the left  $\rightarrow$  positively skewed.

Hence, binomial distribution is 'Asymmetrical'

When  $p > 0.5$  and  $p < 0.5$



6. If 'n' is large and if neither 'p' nor 'q' is nearly zero, in such cases the binomial distribution is modified to normal distribution by standardizing the variable

$$Z = \frac{X - np}{\sqrt{npq}}$$

### 2.4.1.2 Applications

#### Q11. State the Applications of Binomial Distribution.

*Ans :*

Binomial distribution is applicable in case of repeated trials such as,

1. Number of applications received for a junior assistant post during a period a particular period of time.
2. Number of births taking place in a hospital.
3. Number of candidates appearing for the screening test conducted by a company.

All the trials are statistical independent and each trials has two outcomes namely, success and failure.

PROBLEMS

1. Fit a Binomial distribution to the following data.

X	0	1	2	3	4
F(x)	122	60	15	2	1

Sol :

(July-18)

X	f	f(x)
0	122	0
1	60	60
2	15	30
3	2	6
4	1	4
	200	100

$$\text{Mean} = \frac{\sum fX}{N} = \frac{100}{200} = 0.5$$

$$\text{Mean} = 0.5$$

$$np = 0.5$$

$$p = \frac{0.5}{4} = 0.13$$

$$p = 0.13$$

$$q = 1 - 0.13 = 0.87.$$

$$p(r) = {}^nC_r p^r q^{n-r}$$

$$n = 4 \quad p = 0.13 \quad r = 0, 1, 2, 3, 4 \quad q = 0.87$$

$$\begin{aligned} p(0) &= {}^4C_0 \cdot (0.13)^0 (0.87)^{4-0} \\ &= 1 \times 1 \times 0.5729 = 0.57. \end{aligned}$$

$$\begin{aligned} p(1) &= {}^4C_1 \cdot (0.13)^1 (0.87)^{4-1} \\ &= 4 \times 0.13 \times 0.66 = 0.34. \end{aligned}$$

$$\begin{aligned} p(2) &= {}^4C_2 \cdot (0.13)^2 (0.87)^{4-2} \\ &= 6 \times 0.017 \times 0.76 = 0.0775. \end{aligned}$$

$$\begin{aligned} p(3) &= {}^4C_3 \cdot (0.13)^3 (0.87)^{4-3} \\ &= 4 \times 0.002 \times 0.87 = 0.0067 \end{aligned}$$

$$\begin{aligned} p(4) &= {}^4C_4 \cdot (0.13)^4 (0.87)^0 \\ &= 1 \times 0.0003 \times 1 = 0.0003 \end{aligned}$$



## Fitting of Binomial Distribution

r	$p(r) = {}^n C_r p^r q^{n-r}$	$f(r) = N \cdot (p(r))$
0	$p(0) = {}^4 C_0 (0.13)^0 (0.87)^4 = 0.5729$	$200 \times 0.5729 = 114.58$
1	$p(1) = {}^4 C_1 (0.13)^1 (0.87)^3 = 0.3432$	$200 \times 0.3432 = 68.64$
2	$p(2) = {}^4 C_2 (0.13)^2 (0.87)^2 = 0.0775$	$200 \times 0.0775 = 15.5$
3	$p(3) = {}^4 C_3 (0.13)^3 (0.87)^1 = 0.0067$	$200 \times 0.0067 = 1.34$
4	$p(4) = {}^4 C_4 (0.13)^4 (0.87)^0 = 0.0003$	$200 \times 0.0003 = 0.06$
		200

2. Output of a production process is known to be 30% defective. What is the Probability that a sample of a 5 items would contain 0, 1, 2, 3, 4 and 5 defectives?

*Sol:*

If the appearance of a defective item is considered a success, then the probability of success in a trial,  $p = 0.3$  thus  $q = 1 - p = 1 - 0.3 = 0.7$ . With  $n = 5$ , and  $P(X = r) = {}^n C_r q^{n-r} p^r$  we have

Number of successes (r)	Probability $P(X = r)$
0 ${}^5 C_0 (0.7)^5 (0.3)^0$	0.1681
1 ${}^5 C_1 (0.7)^4 (0.3)^1$	0.3602
2 ${}^5 C_2 (0.7)^3 (0.3)^2$	0.3087
3 ${}^5 C_3 (0.7)^2 (0.3)^3$	0.1323
4 ${}^5 C_4 (0.7)^1 (0.3)^4$	0.0283
5 ${}^5 C_5 (0.7)^0 (0.3)^5$	0.0024
<b>Total</b>	<b>1.0000</b>

Binomial Distribution has two parameters:  $n$  and  $p$ . It means that a Binomial Distribution can be specified completely by its  $n$  and  $p$ .

The mean of a binomial distribution is  $np$  and its Standard Deviation equals  $\sqrt{npq}$ .

3. In an office, there are 150 employees. The pattern of their absence for duty in a particular month is recorded in the following table. Fit a Binomial Distribution to the data.

Number of days absent	X	0	1	2	3	4
Number of Absentees	f(x)	28	62	46	10	4

*Sol:*

In the usual notations we have :  $n = 4$ ;  $N = fx = 150$  ... (1)

If  $p$  is the popular of the binomial distribution,

then  $np = \text{Mean of the distribution} = \bar{X}$  ... (2)

Now  $\bar{x} = \{\Sigma fx\} / \{\Sigma f\} = (0 + 62 + 92 + 30 + 16) / 150$

$$= 200 / 150 = 4/3$$

Substituting in (2) we get

$$4 \cdot p = 4/3, \text{ or } p = 1/3 \text{ and } q = 1 - p = 2/3$$

The expected binomial probabilities are given by :

$$\begin{aligned} P(X = r) &= {}^n c_r p^r q^{n-r} \\ &= {}^4 c_r (1/3)^r (2/3)^{4-r} \quad \dots(3) \end{aligned}$$

putting  $x = 0, 1, 2, 3$ , and  $4$  in (3) we get the expected binomial probabilities as given in the following table.

#### Fitting of Binomial Distribution

X	P(x)	Frequency f(x) = N.P(x)
0	${}^4 c_0 (1/3)^0 (2/3)^4 = 16/81 = 0.1975$	$29.63 = 30$
1	${}^4 c_1 (1/3) (2/3)^3 = 4(8/81) = 0.3951$	$59.26 = 59$
2	${}^4 c_2 (1/3)^2 (2/3)^2 = 4(3/21) (4/81) = 0.2963$	$44.44 = 44$
3	${}^4 c_3 (1/3)^3 (2/3) = 4(2/81) = 0.0988$	$14.81 = 15$
4	${}^4 c_4 (1/3)^4 = 1/81 = 0.0123$	$1.85 = 2$

Hence the fitted Binomial Distribution is :

X	0	1	2	3	4	Total
f(x)	30	59	44	15	2	150

#### 2.4.2 Poisson Distribution

**Q12. Define Poisson distribution. State the assumptions of Poisson distribution.**

*Ans :*

(June-19, Imp.)

##### Meaning

In probability theory and statistics, the Poisson distribution is a discrete probability distribution that expresses probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

##### Examples

1. The number of telephone calls received at a particular switch board per minute during a certain hour of the day.
2. The number of deaths per day in a district or town in a one year by disease (but not epidemic).

3. The number of cars passing a certain point per minute.
4. The number of persons born deaf and dumb per year in a city.
5. The number of typographical errors per page.
6. The number of printing errors per page
7. The number of defective blades in a pack of 100.

### Mathematical Definition

It is defined by the probability function.

for  $x$  (no. of successes)  $= 0, 1, 2, 3,$

where  $m$  is fixed positive number.

$$e = 2.71828(\text{app.})$$

### Assumptions

- 1) The events must be statistically independent. This means that the events must not occur in clusters. For example, if a pupil is absent from school on Monday this has no effect on the probability that more absentees will occur in the same week. Or if a spelling error is found on one page, it does not mean there is a probability of finding other errors on the same page.
- 2) An event has a minimal probability of occurrence. The event, such as teacher transfer, has many opportunities to occur, but the probability that the event will occur at any opportunity is extremely small.
- 3) The probability of an event is proportional to the size of the area of probability. For example, the probability that a teacher will be transferred from a school is greater over a 10-year period than over a one - term period.

#### 2.4.2.1 Properties

#### Q13. What are the Properties of Poisson distribution?

Ans :

(June-19)

Properties of Poisson Distribution are discussed in the following section.

- i) Poisson Distribution is a Discrete Probability Distribution, since the random variable  $X$  can take only values  $0, 1, 2, \dots, \infty$ .
- ii) By putting  $r = 0, 1, 2, 3, \dots$ , in (1), we obtain the probabilities of  $0, 1, 2, 3, \dots$  successes respectively,
- iii) Total probability is 1.

$$\begin{aligned}
 \Sigma P(r) &= e^{-m} + me^{-m} + (m^2/2!) e^{-m} + (m^3/3!) e^{-m} + \dots \\
 &= e^{-m} [1 + m + m^2/2! + m^3/3! + \dots] \\
 &= e^{-m} \times e^m = e^{-m} \times m \\
 &= e^0 = 1
 \end{aligned}$$

$$E(X) = \text{Mean} = \sum_{r=0}^{\infty} rP(r) = \sum_{r=0}^{\infty} r(e^{-m}) (m^r / r!)$$

$$= me^{-m} \times e^m = m$$

$$E(x^2) - [E(x)]^2 = \text{Variance} = \sum r^2 P(r) - [\sum rP(r)]^2$$

$$= \sum r^2 P(r) - (\text{mean})^2$$

$$= [me^{-m}][e^m(1+m)] - m^2$$

**Note :** One of the special properties associated with Poisson Distribution is

Mean = Variance = m

- iv) If we know m, all the Probabilities of the Poisson Distribution can be obtained. Therefore, m is called as the parameter of the Poisson Distribution.

#### 2.4.2.2 Applications

**Q14. What are the Applications of Poisson distribution?**

*Ans :*

Some practical situations where Poisson Distribution can be used.

- i) Number of telephone calls arriving at a telephone switch board is a unit time (say, per minute)
- ii) Number of customers arriving at the super market, say, per hour.
- iii) The number of defects per unit of manufactured product (This is done for the construction of control chart for c in Statistical Quality Control)
- iv) To count the number of radio-active element per unit of time (Physics)
- v) The number of bacteria growing per unit time (Biology)
- vi) The number of defective material say, pins, blades etc. in a package manufactured by a good concern.
- vii) The number of suicides reported in a particular day.
- viii) The number of casualties (persons dying) due to rare disease such as heart attack or cancer or snake bite in a year.
- ix) Number of accidents taking place per day on a busy road.
- x) Number of typographical errors per page in a typed material or the number of printing mistakes per page in a book.

**Q15. What is meant by theoretical frequency distribution?**

*Ans :*

(June-19)

It is defined as the distributions that are drawn from the expectations based on past (or) theoretical consideration. It is also known as expected frequency distribution (or) frequency distribution.

**PROBLEMS**

4. It is known from past experience that in a certain industrial plant there are on the average 4 industrial accidents per month. Find the probability that in a given year there will be less than 4 accidents. Assume Poisson distribution ( $e^{-4} = 0.0183$ )

*Sol :*

In the usual notations we are given  $m = 4$ . If the random variable  $X$  denotes the number of accidents in the industrial plant per month, then by Poisson Probability law,

$$P(X = r) = e^{-m} m^r / r! = e^{-4} 4^r / r! \quad \dots (1)$$

The required probability that there will be less than 4 accidents is given by

$$P(X < 4) = [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$$

$$P(X < 4) = e^{-4} [1 + 4 + 4^2/2! + 4^3/3!]$$

$$= e^{-4} [1 + 4 + 8 + 10.67]$$

$$P(X < 4) = e^{-4} [23.67] = [0.0183] [23.67]$$

$$= 0.4332.$$

5. If 5% of the electric bulbs manufactured by a company are defective, using Poisson distribution find the probability than in a sample of 100 bulbs :

(i) None is defective,

(ii) 5 bulbs are defective (Given :  $e^{-5} = 0.07$ ).

*Sol :*

Here we are given  $n = 100$ ,

$P$  = Probability of a defective bulb = 5% = 0.05.

Since  $P$  is small and  $n$  is large we may approximate the given distribution by Poisson distribution. Hence the parameter  $m$  of the Poisson distribution is :

$$m = np = 100 \times 0.05 = 5$$

Let the random variable  $X$  denote the number of defective bulbs in a sample of 100. Then (by Poisson law)

$$P(X = r) = e^{-m} m^r / r! = e^{-5} 5^r / r! ; r = 0, 1, 2, \dots, \infty \quad \dots(1)$$

- i) The probability that none of the bulbs is defective is given by :

$$P(X = 0) = e^{-5} = 0.07 \text{ [from 1]}$$

- ii) The probability of 5 defective bulbs is given by :

$$P(X = 5) = e^{-5} \times 5^5 / 5! = 0.07(3125/120)$$

$$= 0.007 (625/24) = 4.375/24 = 0.1823.$$

6. In a Research Methodology Book, the following frequency mistakes per page were observed. Fit a Poisson distribution.

No. of Mistakes	0	1	2	3	4	5
No. of Pages	620	180	80	60	40	80

*Sol :*

(Dec.-20, Imp.)

Steps in the fitting of Poisson distribution are as follows.

### Step-1

Calculate the value of  $\lambda$  and probability of zero occurrence.

x	f	fx
0	620	0
1	180	180
2	80	160
3	60	180
4	40	160
5	80	400
	<b><math>\Sigma f = 1060</math></b>	<b><math>\Sigma fx = 1080</math></b>

Mean of Poisson distribution is given by ' $\lambda$ '

$$\lambda = \frac{\Sigma fx}{\Sigma f} = \frac{1080}{1060} = 1.0188$$

$$\therefore \lambda = 1.0188$$

Probability in Poisson distribution is given by,

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

$$\lambda = 1.0188$$

$$e^{-\lambda} = e^{-1.0188} = 0.3610$$

$\therefore$  Probability of zero occurrence,

$$P(0) = \frac{e^{-1.0188} \cdot \lambda^0}{0!} = e^{-1.0188} = 0.3610$$

$$\therefore P(0) = 0.3610$$

### Step-2

Calculate all the probabilities by using recurrence relation.

$$P(0) = 0.3610$$

$$P(1) = \frac{P(0) \cdot \lambda}{1} = 0.3610 \times 1.0188 = 0.3678$$

$$P(2) = \frac{P(1) \cdot \lambda}{2} = \frac{0.3678 \times 1.0188}{2} = 0.1874$$

$$P(3) = \frac{P(2) \cdot \lambda}{3} = \frac{0.1874 \times 1.0188}{3} = 0.0636$$

$$P(4) = \frac{P(3) \cdot \lambda}{4} = \frac{0.0636 \times 1.0188}{4} = 0.0162$$

$$P(5) = \frac{P(4) \cdot \lambda}{5} = \frac{0.0162 \times 1.0188}{5} = 0.0033$$

**Step-3**

Multiply each term of all probabilities with total frequency ( $\Sigma f$  or  $N$ ) to obtain expected frequencies.

<b>x</b>	<b>P(x)</b>	<b>f(x) = N P (x) = 1060 P(x)</b>
0	$P(0) = 0.3610$	$f(0) = 1060 \times 0.3610 = 382.66 \cong 383$
1	$P(1) = 0.3678$	$f(1) = 1060 \times 0.3678 = 389.868 \cong 390$
2	$P(2) = 0.1874$	$f(2) = 1060 \times 0.1874 = 198.644 \cong 199$
3	$P(3) = 0.0636$	$f(3) = 1060 \times 0.0636 = 67.416 \cong 67$
4	$P(4) = 0.0162$	$f(4) = 1060 \times 0.0162 = 17.172 \cong 17$
5	$P(5) = 0.0033$	$f(5) = 1060 \times 0.0033 = 3.50 \cong 4$
	<b><math>\Sigma P(x) = 1</math></b>	<b><math>\Sigma f(x) = 1060</math></b>

Thus, the theoretical fitted poisson distribution is as follows,

Number of Mistakes Per Page	0	1	2	3	4	5
Number of pages	383	390	199	67	17	4

7. A book containing 1000 pages has 0, 1, 2, 3 or 4 misprints per page as shown below :

Number of misprints	0	1	2	3	4
Number of pages	500	340	120	30	10

Fit Poisson distribution to the data and compare the theoretical frequencies into those given in the question

*Sol:*

### Steps in the Fitting of Poisson Distribution

#### Step 1

Calculate the variation of ' $\lambda$ ' and probability of zero occurrence,

x	f	fx
0	500	0
1	340	340
2	120	240
3	30	90
4	10	40
<b>Total</b>	<b><math>\Sigma f = 1000</math></b>	<b><math>\Sigma fx = 710</math></b>

Mean,  $\lambda = \frac{\Sigma fx}{\Sigma f} = \frac{710}{1000} = 0.71$

$$\therefore \lambda = 0.71$$

Probability of 'X' is Poisson distribution is given by,

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

$\therefore$  Probability of zero occurrence is calculated as follows.

$$P(0) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$$

$$= e^{-\lambda} \quad [\because \lambda^0 = 1 \text{ and } 0! = 1]$$

$$= e^{-0.71}$$

$$= 0.492$$

$$\therefore P(0) = 0.492$$

#### Step 2

Calculate each term of probability by using recurrence relation

$$P(0) = 0.492 \quad [\text{From step 1}]$$

$$\therefore P(1) = \frac{P(0) \times \lambda}{1} = 0.492 \times 0.71$$



$$\therefore P(2) = \frac{P(1) \times \lambda}{2}$$

$$= \frac{0.349 \times 0.71}{2}$$

$$\therefore P(2) = 0.123895 \cong 0.124$$

$$P(3) = \frac{P(2) \times \lambda}{3} = \frac{0.124 \times 0.71}{3}$$

$$\therefore P(3) = 0.0293$$

$$P(4) = \frac{P(3) \times \lambda}{4} = \frac{0.0293 \times 0.71}{4} = 0.0052.$$

### Step 3

Multiply each term of probability with total frequency ( $\Sigma f$ ) to obtain the expected frequencies is shown in table.

X	P(X)	f(x) = N.P(X) = 1000.P(X)
0	p(0) = 0.492	f(0) = 0.492 × 1000 = 492
1	p(1) = 0.349	f(1) = 0.349 × 1000 = 349
2	p(2) = 0.124	f(2) = 0.124 × 1000 = 124
3	P(3) = 0.0293	f(3) = 0.0293 × 1000 = 29.3
4	P(4) = 0.0052	f(4) = 0.0052 × 1000 = 5.2

**Table: Computation of Expected Frequencies**

The comparison of theoretical frequencies (that are fitted by Poisson distribution) with given frequencies is given below.

Number of misprints	0	1	2	3	4
Number of page (given)	500	340	120	30	10
Theoretical frequencies	492	349	124	29.3	5.2

### 8. Fit a Poisson distribution to the following data and calculate theoretical frequencies,

Deaths	0	1	2	3	4
Frequency	122	60	15	2	1

*Sol :*

(July-18)

#### Steps in the Fitting of Poisson Distribution

##### Step 1

Calculate the values of ' $\lambda$ ' and the probability of zero occurrence

x	f	fx
0	122	0
1	60	60
2	15	30
3	2	6
4	1	4
$\Sigma f = 200$		$\Sigma fx = 100$

Mean of Poisson distribution,  $\lambda = \frac{\Sigma fx}{\Sigma f}$

$$\therefore \lambda = \frac{100}{200} = 0.5 \quad \boxed{\therefore \lambda = 0.5}$$

Probability of 'X' in Poisson distribution is given by,

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

$\therefore$  Probability of zero occurrence can be computed as,

$$\begin{aligned} P(0) &= \frac{e^{-\lambda} \cdot \lambda^x}{x!} \\ &= e^{-\lambda} \quad [\because \lambda^0 = 1 \text{ and } 0! = 1] \\ &= e^{-0.5} \\ &= 0.6065 \end{aligned}$$

$$\therefore P(0) = 0.6065$$

### Step 2

Calculate each term of probability by using recurrence relation

$$P(0) = 0.6065 \quad [\text{From step 1}]$$

$$P(1) = \frac{P(0) \times \lambda}{1} = 0.6065 \times 0.5$$

$$\therefore P(1) = 0.30325$$

$$P(2) = \frac{P(1) \times \lambda}{2} = \frac{0.30325 \times 0.5}{2}$$

$$\therefore P(2) = 0.075812$$

$$P(3) = \frac{P(2) \times \lambda}{3} = \frac{0.075812 \times 0.5}{3}$$

$$\therefore P(3) = 0.0126333 \cong 0.01264$$

$$P(4) = \frac{p(3) \times \lambda}{4} = \frac{0.01264 \times 0.5}{4}$$

$$\therefore P(4) = 0.00158$$

**Step 3**

Multiply each term of probability with total frequency to obtain the values of expected frequencies.

**Computation of Expected Frequencies**

X	P(X)	f(x) = N.P(X) = 200.P(X)
0	P(0) = 0.6065	f(0) = 200 × 0.6065 = 121.3 ≈ 121
1	P(1) = 0.30325	f(1) = 200 × 0.30325 = 60.65 ≈ 61
2	P(2) = 0.075812	f(2) = 200 × 0.075812 = 15.1624 ≈ 15
3	P(3) = 0.01264	f(3) = 200 × 0.01264 = 2.528 ≈ 3
4	P(4) = 0.00158	f(4) = 200 × 0.00158 = 0.316 ≈ 0
<b>Total</b>	<b>1</b>	<b>200</b>

Thus, the theoretically fitted Poisson distribution is as follows,

Deaths	0	1	2	3	4
Theoretical frequency	121	61	51	1	0

9. One hundred car stereos are inspected as they come off the production line and the number of defects per set is recorded below.

Number of defects	0	1	2	3	4
Number of sets	79	18	2	1	0

Fit a Poisson distribution to the above data.

*Sol :*

Steps in the Fitting of Poisson Distribution

**Step 1**

Calculate the mean and probability of zero occurrence

x	f	fx
0	79	0
1	18	18
2	2	4
3	1	3
4	0	0
	Σf = 100	Σfx = 25

Mean of Poisson distribution,  $\lambda = \frac{\Sigma fx}{\Sigma f}$

$$\therefore \lambda = \frac{25}{100} = 0.25$$

Probability of 'x' in Poisson distribution is given by,

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

Where,  $\lambda = 0.25$ ,  $x = 0, 1, 2, 3, 4$

The probability of zero occurrence,

$$p(0) = \frac{e^{-0.25} \cdot \lambda^0}{0!} \quad [\because \lambda^0 = 1, 0! = 1]$$

$$p(0) = e^{-0.25} = 0.779$$

$$\therefore p(0) = 0.779$$

### Step 2

Calculate each term of the probability by using recurrence relation

$$p(0) = 0.779$$

$$p(1) = \frac{p(0) \times \lambda}{1}$$

$$= 0.779 \times 0.25$$

$$= 0.19475 \cong 0.1945$$

$$p(2) = \frac{p(1) \times \lambda}{2} = \frac{0.19475 \times 0.25}{2} = 0.0244$$

$$p(3) = \frac{p(2) \times \lambda}{3} = \frac{0.0244 \times 0.25}{3} = 0.002$$

$$p(4) = \frac{p(3) \times \lambda}{4} = \frac{0.002 \times 0.25}{4}$$

$$= 0.000125 \cong 0.0001$$

### Step 3

Multiply each term of probability with total frequency ( $\Sigma f$ ) to obtain the values of the expected frequencies.

X	P(X)	f(x) = N.p(X)
0	P(0) = 0.779	f(0) = 100 × 0.779 = 77.9 $\cong$ 78
1	P(1) = 0.1945	f(1) = 100 × 0.1945 = 19.45 $\cong$ 19.5
2	P(2) = 0.0244	f(2) = 100 × 0.0244 = 2.44 $\cong$ 2.5
3	P(3) = 0.0002	f(3) = 100 × 0.002 = 0.2 $\cong$ 0
4	P(4) = 0.001	f(4) = 100 × 0.0001 = 0.01 $\cong$ 0
	<b>Total = 1</b>	<b>100</b>

Thus, the theoretically fitted Poisson distribution is as follows,

Number of defects	0	1	2	3	4
Number of sets	78	19.5	2.5	0	0

10. An automatic machine makes paper clips from coils of wire. On the average. 1 in 400 paper clips is defective. If the paper clips are packed in boxes of 100. What is the probability that any given box of clips will contain (i) no defective, (ii) one or more defective and (iii) less than two defectives.

*Sol :*

(Aug.-17, Imp.)

Probabilities of defective paper clips  $p = \frac{1}{400}$

$$n = 100$$

$$\lambda = np$$

$$\lambda = 100 \times \frac{1}{400} = 0.25$$

$$\lambda = 0.25$$

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

- i) **No Defectives**

$$\begin{aligned} p(0) &= \frac{e^{-\lambda} \cdot \lambda^0}{0!} \\ &= \frac{e^{-0.25} \cdot 0.25^0}{0!} \quad (e^{-0.25} = 0.7788) \\ &= \frac{0.7788 \times 1}{1} = 0.7788 \end{aligned}$$

- ii) **One (or) more defectives**

$$\begin{aligned} p(x \geq 1) &= 1 - p(0) \\ &= 1 - 0.7788 \\ &= 0.2212 \end{aligned}$$

- iii) **Less than two defectives**

$$\begin{aligned} p(x \leq 2) &= p(0) + p(1) \\ &= \left( \frac{e^{-\lambda} \cdot (0.25)^0}{0!} + e^{-\lambda} \frac{(0.25)^1}{1!} \right) \\ &= (0.7788 + 0.7788(0.25)) \\ &= 0.7788 + 0.1947 = 0.9735. \end{aligned}$$

11. In a company, the average number of phone calls per minute coming into a switch board is 3.2. Find the probability that during one particular minute there will be,

- (i) Atleast 5 calls
- (ii) Exactly 4 calls
- (iii) No phone calls.

*Sol:*

(Nov.-20)

Given that,

Average rate of call (Mean),  $\lambda = 3.2$  per minute

$$P(X = x) = \frac{e^{-3.2} (3.2)^x}{x!}$$

Let X be the number of phone calls per minute

$$P(X \geq 5) = 1 - P(X < 5)$$

$$= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)]$$

$$= 1 - \left[ \frac{e^{-3.2} (3.2)^0}{0!} + \frac{e^{-3.2} (3.2)^1}{1!} + \frac{e^{-3.2} (3.2)^2}{2!} + \frac{e^{-3.2} (3.2)^3}{3!} + \frac{e^{-3.2} (3.2)^4}{4!} \right]$$

$$\therefore e^{-3.2} = 0.0408$$

$$= 1 - \left[ \frac{0.0408 \times 1}{1} + \frac{0.0408 \times 3.2}{1} + \frac{0.0408 \times 10.24}{2} + \frac{0.0408 \times 32.77}{6} + \frac{0.0408 \times 104.86}{24} \right]$$

$$= 1 - \left[ 0.0408 + 0.1306 + \frac{0.4178}{2} + \frac{1.34}{6} + \frac{4.28}{24} \right]$$

$$= 1 - [0.0408 + 0.1306 + 0.2089 + 0.2233 + 0.1783]$$

$$= 1 - 0.7819$$

$$= 0.2181$$

$$P(4) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$$

$$= \frac{e^{-3.2} (3.2)^4}{4!}$$

$$= \frac{0.0408 \times 104.86}{24}$$

$$= \frac{4.28}{24}$$

$$= 0.1783$$

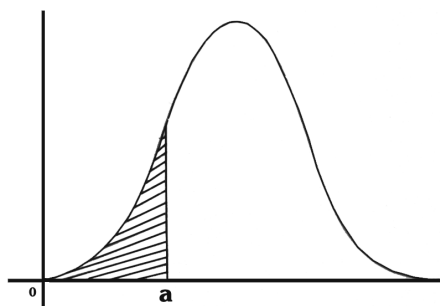
$$\begin{aligned}
 P(0) &= \frac{e^{-\lambda} \cdot \lambda^x}{x!} \\
 &= \frac{e^{-3.2} (3.2)^0}{0!} \\
 &= \frac{0.0408 \times 1}{1} \\
 &= 0.0408
 \end{aligned}$$

## 2.5 CONTINUOUS PROBABILITY DISTRIBUTIONS

**Q16. Explain the concept of Continuous Probability Distributions.**

*Ans :*

Continuous probability distribution is another type of probability distribution that defines the probability of occurrences of continuous random variables. It can not be expressed in the tabular form. Instead, it make use of equation known as probability density function inorder to define continuous probability distribution.



### Types of Continuous Distributions

Following are different types of continuous distributions,

1. Rectangular/Uniform Distribution
2. Normal Distribution
3. Exponential Distribution
4. Gamma Distribution
  - (i) Gamma Distribution with Single Parameter
  - (ii) Gamma Distribution with Two Parameters
5. Beta Distribution
6. Cauchy Distribution

### 2.5.1 Normal Distribution

**Q17. Describe briefly about Normal Distribution.**

(OR)

**What is normal distribution.**

(OR)

**Explain normal probability distribution.**

*Ans :*

(July-18)

#### Introduction

The Normal Distribution was discovered by De Moivre as the limiting case of Binomial model in 1733. It was also known to Laplace no later than 1774, but through a historical error it has been credited to Gauss who first made reference to it in 1809. Throughout the 18th and 19th centuries, various efforts were made to establish the Normal model as the underlying law ruling all continuous random variables – thus the name Normal. The Normal model has, nevertheless, become the most important probability model in statistical analysis.

The normal Distribution in approximation to Binomial Distribution, whether or not  $p$  is equal to  $q$ , the Binomial Distribution tends to the form of the continuous curve when  $n$  becomes large at least for the material part of the range. As a matter of fact, the correspondence between Binomial and the Normal curve is surprisingly close even for low values of  $n$  provide  $dp$  and  $q$  are fairly near to equality. The limiting frequency curve, obtained as  $n$ , becomes large and is called the Normal frequency curve or simply the Normal curve.

#### Probability Density Function

A random variable  $X$  is said to have a Normal Distribution with parameters  $m$  (mean) and  $s^2$  (Variance), if the density function is given by :

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

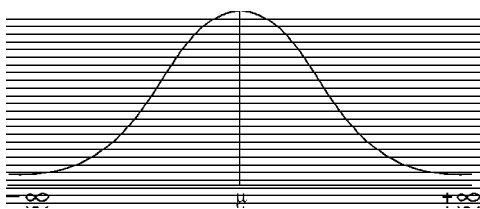
where

$X$  = Value of the continuous random variable

- $\mu$  = Mean of random variable  
 $e$  = Mathematical constant (2.7783)  
 $T_1$  = Mathematical constant (3.14)  
 $\sigma$  = Standard deviation.

### Normal Distribution\_Graph

If we draw the graph of Normal Distribution, the curve obtained will be known as Normal curve and is given below :



The Graph of  $y = f(x)$  is a famous 'bell shaped' curve. The top of the bell is directly above the mean  $m$ . For large values of  $m$ , the curve tends to flatten out and for small values of  $s^2$ , it has a sharp peak.

When we say that curve has unit area we mean that the total area under the Normal Distribution between  $(-\infty$  to  $\infty)$  is equal to 1.

### Q18. Explain the properties of normal distribution.

Ans :

(June-19)

Following are the properties of normal, distribution.

1. The normal curve is 'bell-shaped' and symmetrical about the mean (skewness = 0). If the curve is folded along its central vertical axis the curves either side of the axis would coincide.
2. The height of the normal curve is maximum at its mean. Hence, the mean and mode coincide. Thus in normal distribution, mean, mode and median are equal.
3. The height of the curve is maximum at its mean but reduces as it goes towards either of the direction but never touches the base. Hence, the curve is known as ASYMPTOTIC. The range is unlimited or infinite in both the directions.

4. As there is only one maximum point, the normal curve has only one mode and it known as 'unimodal'.
5. The points of inflection i.e., the points where the change in curvature occurs are  $\bar{x} \pm \sigma$  (or)  $\mu \pm \sigma$ .
6. The variables used in Binomial and Poisson are discrete variables whereas nonnal distribution has continuous random variable.

### 2.5.1.1 Standard Normal Distribution Properties

**Q19. What is Standard Normal Distribution. State the properties of Standard Normal Distribution.**

Ans :

A Random Variable with any mean and standard deviation can be transformed to a Standard Normal Variate (SNV) by subtracting the mean and dividing by the standard deviation. For a Normal Distribution with mean  $m$  and standard deviation  $s$ , the SNV 'Z' is obtained as

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

Where  $Z$  = the distance, expressed as a multiple of the standard deviation, that the value  $X$  lies away from the mean. The SNV  $Z$  has mean zero and variance '1'.

$x$  = Random variable,

$\mu$  = Mean

$\sigma$  = standard deviation.

In symbols, if  $X \sim N(\mu, \sigma^2)$ , then  $Z \sim N(0,1)$ .

### Properties

1. It is a continuous probability distribution having parameters  $m$  and  $\sigma$ .
2. The normal curve is perfectly symmetrical about the mean ( $m$ ) and is bell shaned. It means that if we fold the curve along the vertical line at the centre, the two halves of the curve work incide. The two tails of the curve on either side of  $x = m$  extends to infinity.
3. Mean = Median = Mode, Skewness = 0.
4. It has only one mode.



5.  $Q_1$  and  $Q_3$  are equidistant from median (or mean)  
 $Q_1 = m - 0.67 \sigma$   $Q_3 = m + 0.67 \sigma$  (app.)  
 Quartile Deviation =  $0.67 \sigma$ , mean deviation (about mean) =  $0.80 \sigma = 4/5 \sigma$ .
6. The maximum ordinate is at  $x = m$  its value is  $\frac{1}{\sigma\sqrt{2\lambda}}$ .
7. The area under the curve means the area lying between the curve and the horizontal axis and is equal to the number of frequencies. The s.d. ( $\sigma$ ) distributes this area as follows:
- 68.27% of the area lies between  $x = -\sigma$  to  $x = +\sigma$ , 34.134% area will lie on either side of the mean.
  - 95.45% of the area lies between  $x = -2\sigma$  to  $x = +2\sigma$  i.e. 47.725% area will lie on either side of the mean.
  - 99.73% of the area lies between  $x = -3\sigma$  to  $x = +3\sigma$ , 49.865% area will lie on either side of the mean.

#### 2.5.1.2 Applications

**Q20. What are the applications of normal distribution?**

*Ans :* (June-19)

#### 1. Production/Operations

- A workshop produces a known quantity of units per day. The average weight of unit and standard deviation are given. Assuming normal distribution, we can find how many units are expected to weigh less than greater than some given weight.
- A company manufacturers Electric bulbs and find that lifetime of the bulbs is normally distributed with some average life in hours and standard deviation in hours. On the basis of the information it can be estimated that the number of bulbs that is expected to burn for more than specified hours and less than specified hours.

#### 2. Finance / Accounting / Receivables

- In a business, the amount of daily collection is given for a particular period, we can estimate the average daily collection and Standard Deviation of this business. Assuming the daily collection follows a Normal Distribution.

#### 3. Health Care and Insurance Services

- In patient's medical sample information, many parameters like cholesterol, urea in blood, hemoglobin, sugar, lipid profile, blood pressure etc., are used for diagnostic testing purpose. If each parameter follows Normal Distribution, we can calculate number of patients having abnormal and normal levels to decide upon the treatment to be followed.
- In insurance industry, if insurance Premium follows normal distribution, we can calculate how many persons fall above or below a certain insured amount to plan marketing strategies by the management.

#### 4. Personnel

- Given with average and variance of a wage distribution of a group of workers one can estimate the number of workers in different wage ranges.
- Given a distribution of training hours of a category of employees, it can be planned for the number of hours required for training an employee to suit a particular work requirement.

#### 2.5.1.3 Importance of Normal Distribution

**Q21. Explain the Importance of Normal Distribution.**

(OR)

**Explain the significance of normal distribution.**

*Ans :*

The Normal Distribution has great significance in statistical data analysis, because of the following reasons :

- i) The Normal Distribution has a remarkable property stated in the central limit theorem, which asserts that if  $X_1, X_2, \dots, X_n$  are  $n$  independent and identically random samples from a Normal Distribution which mean ( $m$ ) and standard deviation ( $s$ ), then the sample mean ( $\bar{x}$ ) is also a Normal Distribution with

mean ( $m$ ) and standard error  $\left(\frac{\sigma}{\sqrt{n}}\right)$ . This result is true even if the population from which the samples are drawn is not a Normal Distribution subject to condition that  $n$ , the sample size is sufficiently large ( $n > 30$ ).

- ii) Even if a variable is not Normally distributed, it can sometimes be brought to Normal form by simple transformation of variable. For example, if Distribution of  $X$  is skewed, the Distribution of  $\sqrt{X}$  might come out to be Normal.
- iii) Many of the sampling Distributions like Student's  $t$ , Snedecor's  $F$ , etc. also tend to Normal Distribution.
- iv) The sampling theory and tests of significance are based upon the assumption that samples have been drawn from a Normal population with mean  $m$  and variance  $s^2$ .
- v) Normal Distribution find large applications in Statistical Quality Control.
- vi) As  $n$  becomes large, the Normal Distribution serves as a good approximation for many discrete Distributions (such as Binomial, Poisson, etc.).
- vii) In theoretical statistics many problems can be solved only under the assumption of a Normal population. In applied work we often find that methods developed under the normal probability law yield "satisfactory" results, even when the assumption of a normal population is not fully met, despite the fact that the problem can have a normal solution only if such a premise is hypothesized.

#### PROBLEMS

12. If the salary of workers in a factory is assumed to follow a Normal Distribution with a mean of Rs. 500 and a S.D. of Rs. 100, find Number of workers whose salary vary between Rs. 400 and Rs. 650, give

the number of workers in the factory as 15,000 ?

*Sol:*

The required area will be calculated only after finding the corresponding  $Z$  values as shown below.

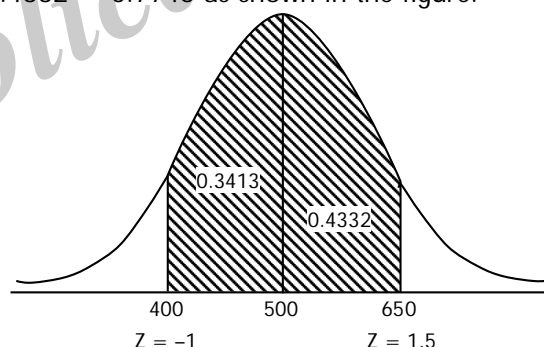
$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{(400 - 500)}{100} = -1$$

(left of the mean)

$$z_2 = \frac{x_2 - \mu}{\sigma} = \frac{(650 - 500)}{100} = +1.5$$

(right of the mean)

Now we read the area between  $Z = 0$  to  $Z = 1$  from table as 0.3413. Because of symmetry the area between  $Z = -1$  to  $Z = 0$  is same as that of the area between  $Z = 0$  to  $Z = 1$ . Again the area between  $Z = 0$  to  $Z = +1.5$  is read from table as 0.4332. Thus the desired area between  $x_1 = 400$  and  $x_2 = 650$  (i.e.,  $Z = -1$  to  $Z = +1.5$ ) is  $(0.3413 + 0.4332 = 0.7745$  as shown in the figure.



Hence, the number of workers whose salary will be between 400 and 650 is given by  $0.7745 \times 15,000 = 11,618$ .

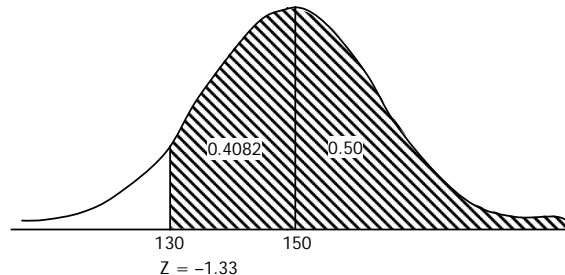
Hence, the number of workers whose salary will be between 400 and 650 is given by  $0.7745 \times 15,000 = 11,618$ .

13. A large flashlight is powered by 5 batteries. Suppose that the life of a battery is normally distributed with  $m = 150$  hours and  $s = 15$  hours. The flashlight will cease functioning if one or more of its batteries go dead. Assuming the lives of batteries are independent, what is the probability that flashlight will operate more than 130 hours ?

*Sol:*

The required area will be calculated only after finding the corresponding Z value as shown below.

$$z = \frac{x - \mu}{\sigma} = \frac{(130 - 150)}{15} = -1.33$$



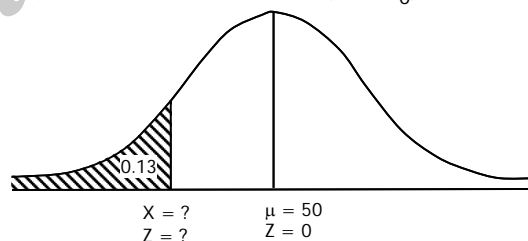
From the table we read the area from  $Z = 0$  to  $Z = 1.33$  as 0.4082. Due to symmetry, the area from  $Z = -1.33$  to  $Z = 0$  is same as 0.4082. The area to the right of mean (i.e., beyond  $Z = 0$ ) is 0.5. The required area ( $\geq 130$  hours of operating time of flash light) is  $\{0.4082 + 0.5\} = 0.9082$ . Hence the probability that the flashlight will operate for more than 130 hours is 0.9082 and is shown in the figure below.

**14. Given a normal distribution with  $m = 50$  and  $s = 10$ , find the value of  $X$  that has (i) 13% of the area to its left and (ii) 14% of the area to its right.**

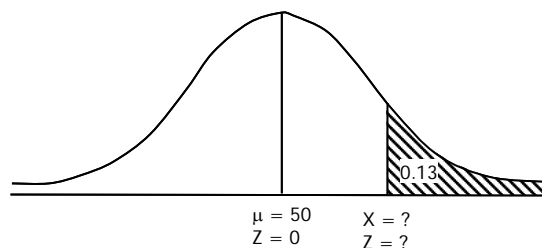
*Sol:*

In the previous examples, we solved first going from a value of  $X$  to a  $Z$  value and then computing desired area. In this example, it just reverse that we begin with a known area of probability, read  $Z$  value and then determine  $X$  by rearranging the formula given below  $z = \frac{x - \mu}{\sigma}$  to give  $X = sZ + m$

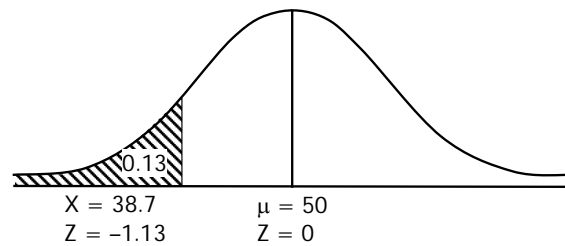
- i) An area of 0.13 to the left of the desired  $X$  value is shaded in the following figure. We require a  $Z$  value that leaves an area of 0.13 to the left i.e.,  $P(Z < z_0) = 0.13$  as shown below.



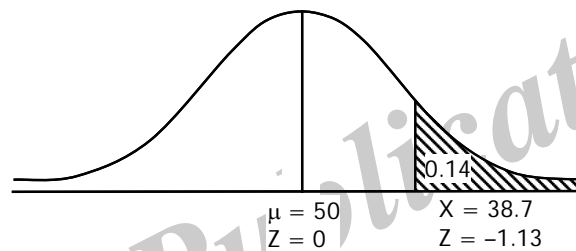
But it is not possible to read the  $Z$  value (or area) on the left side as the tables are not available for this purpose in this book. So, we use the property of symmetry and read the corresponding  $Z$  on the right side as shown below.



The table value for an area of 0.13 is 1.13. Because of symmetry this will be  $(-1.13)$  on the left side. By substituting  $Z = -1.13$  into  $X = aZ + m$  we have  $X = 10(-1, 13) + 50 = 38.7$ . Finally, values are as given below.



- ii) In this case we require a Z value that leaves 0.14 of the area to the right. This means an area of 0.46 lies between  $Z = 0$  and a Z value to be read from the table. From the tables, we can read this value as  $Z = 1.08$ . Once again we substitute the value of Z into  $X = \sigma Z + \mu$  to get  $X = 10(1.08) + 50 = 60.8$ . The final picture is as below.



## Short Question & Answers

### 1. Probability Distribution.

*Ans :*

The probability distribution describes the range of possible values that a random variable can attain and the probability that the value of the random variable is within any (measurable) subset of that range.

When the random variable takes values in the set of real numbers, the probability distribution is completely described by the cumulative distribution function, whose value at each real  $x$  is the probability that the random variable is smaller than or equal to  $x$ .

A probability distribution is a graph, table, or formula that gives the probability for each value of the random variable.

If  $x$  is a random variable then denote by  $P(x)$  to be the probability that  $x$  occurs. It must be the case that  $0 \leq P(x) \leq 1$  for each value of  $x$  and  $\sum P(x) = 1$  (the sum of all the probabilities is 1).

### 2. Define Random Variable.

*Ans :*

A variable that can many real values which are determined by the outcomes of a random experiment on a real line  $(-\infty, +\infty)$ . It is a chance or Stochastic Variable.

It is a function defined on the sample space, " $S$ " of a random experiment. A Random variable takes different values as a result of the outcomes of a random experiment. A Random variable can be Discrete or Continuous.

### 3. Discrete Random Variables

*Ans :*

A discrete random variable is one which may take on only a countable number of distinct values such as 0, 1, 2, 3, 4, ..... Discrete random variables

are usually (but not necessarily) counts. If a random variable can take only a finite number of distinct values, then it must be discrete. Examples of discrete random variables include the number of children in a family, the Friday night attendance at a cinema, the number of patients in a doctor's surgery, the number of defective light bulbs in a box of ten.

The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the probability mass function.

Suppose a random variable  $X$  may take  $k$  different values, with the probability that  $X = x_i$  defined to be  $P(X = x_i) = p_i$ . The probabilities  $p_i$  must satisfy the following:

- i)  $0 \leq p_i \leq 1$  for each  $i$ ,
- ii)  $p_1 + p_2 + \dots + p_k = 1$ .

Discrete random variables take on integer values, usually the result of counting.

### 4. Expectations of random variables.

*Ans :*

#### Expectation of Random Variable

The arithmetic mean of a discrete random variable of the probability distribution is called as 'expectation of random variable' or 'expected value of the random variable'.

Let ' $X$ ' be the discrete random variable with the corresponding probability distribution  $P(X)$ , then the expected value is given by,

$$E(X) = \sum X.P(X)$$

$$\mu = \sum X.P(X)$$

Expected value is usually represented by a greek letter ' $\mu$ '.

In other words, the sum of the product of the random variables and its corresponding probabilities is called as 'expected value of the random variable'.

If a discrete random value 'X' has ' $x_1$ ', ' $x_2$ ', ' $x_3$ '..... ' $x_n$ ' as all the possible values with corresponding probabilities  $P(x_1)$ ,  $P(x_2)$ ,  $P(x_3)$ ,  $P(x_n)$ , then the expected value is given by,

$$E(X) = x_1 P(x_1) + x_2 P(x_2) + x_3 P(x_3) + \dots + x_n P(x_n) = \mu$$

## 5. Probability Distribution Function.

*Ans :*

Probability distribution is a set of probabilities of all the possible outcomes of a random experiment.

Probability distribution is similar to frequency distribution. Probability distribution is based on the theoretical considerations, subjective assessment or on experience.

For example, 'X' is a random variable which can take the values  $x_1, x_2, x_3, \dots$

The probabilities associated with each of the possible values of 'X',  $P(X = x_i) = P_i (i = 1, 2, 3, \dots)$

Therefore, the collection of pairs  $(x_i, P_i)$ , where  $i = 1, 2, 3, \dots$  is called probability distribution of random variable X.

The values of probability distribution are usually tabulated as given below.

$X = x_i$	$P(X = x_i)$
$x_1$	$P_1$
$x_2$	$P_2$
$x_3$	$P_3$
—	—
—	—
—	—
—	—
$x_n$	$P_n$

Probability distribution functions are usually represented as  $f(x)$ ,  $g(x)$ ,  $h(x)$  etc. Here  $f(x)$  is the function where  $f(x) = P(X = x)$  which assigns probability to each possible outcome 'x', hence called as 'probability distribution'.

## 6. What is Binomial Distribution?

*Ans :*

In probability theory and statistics, the binomial distribution is the discrete probability distribution of the number of successes in a sequence of  $n$  independent yes/no experiments, each of which yields success with probability  $p$ . Such a success/failure experiment is also called a Bernoulli experiment or Bernoulli trial.

In fact, when  $n = 1$ , the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance.

This binomial distribution is also known as the Bernoulli distribution by the name of the Swiss Mathematician Jacob Bernoulli who has derived it.

The binomial probability refers to the probability that a binomial experiment results in exactly  $x$  successes.

### Mathematical Definition

If an event E has probability  $p$  of occurring in each of  $n$  independent trials and that of failure in any trial is  $q (= 1 - p)$  then the probability that it will occur exactly  $r$  times in  $n$  trials is given by:

$$p(r) = {}^nC_r p^r q^{n-r}$$

This probability distribution is called the binomial probability distribution. The binomial distribution is a discrete distribution with parameters  $n$  and  $p$ . If  $p, q$  are equal it is symmetrical, otherwise it is nonsymmetrical.

Where

$p$  = probability of success in a single trial.

$q$  = probability of failure a single trial.

$p + q = 1$ :

$n$  = Number of trials

$r$  = Number of success in ' $n$ ' times.

## 7. State the Properties of Binomial Distribution.

*Ans :*

The properties of binomial distribution are as follows,

1. It describes the distribution of probabilities when there are only two mutually exclusive outcomes for each trial of an experiment for example while tossing a coin, the two possible outcomes are head and tail.
2. The process is performed under identical conditions for 'n' number of times.
3. Each trial is independent of other trials.
4. The probability of success 'p' remains same for trial to trial throughout the experiment and similarly, the probability of failure ( $q = 1 - p$ ) also remains constant overall the observations.

### 8 State the Applications of Binomial Distribution.

*Ans :*

Binomial distribution is applicable in case of repeated trials such as,

1. Number of applications received for a junior assistant post during a period a particular period of time.
2. Number of births taking place in a hospital.
3. Number of candidates appearing for the screening test conducted by a company.

All the trials are statistical independent and each trials has two outcomes namely, success and failure.

### 9. Define Poisson distribution.

*Ans :*

#### Meaning

In probability theory and statistics, the Poisson distribution is a discrete probability distribution that expresses probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

### Examples

1. The number of telephone calls received at a particular switch board per minute during a certain hour of the day.
2. The number of deaths per day in a district or town in a one year by disease (but not epidemic).
3. The number of cars passing a certain point per minute.
4. The number of persons born deaf and dumb per year in a city.
5. The number of typographical errors per page.
6. The number of printing errors per page
7. The number of defective blades in a pack of 100.

### 10. What are the Applications of Poisson distribution?

*Ans :*

Some practical situations where Poisson Distribution can be used.

- i) Number of telephone calls arriving at a telephone switch board is a unit time (say, per minute)
- ii) Number of customers arriving at the super market, say, per hour.
- iii) The number of defects per unit of manufactured product (This is done for the construction of control chart for  $\bar{c}$  in Statistical Quality Control)
- iv) To count the number of radio-active element per unit of time (Physics)
- v) The number of bacteria growing per unit time (Biology)
- vi) The number of defective material say, pins, blades etc. in a package manufactured by a good concern.
- vii) The number of suicides reported in a particular day.

**11. What is normal distribution?***Ans :*

The Normal Distribution was discovered by De Moivre as the limiting case of Binomial model in 1733. It was also known to Laplace no later than 1774, but through a historical error it has been credited to Gauss who first made reference to it in 1809. Throughout the 18th and 19th centuries, various efforts were made to establish the Normal model as the underlying law ruling all continuous random variables – thus the name Normal. The Normal model has, nevertheless, become the most important probability model in statistical analysis.

The normal Distribution in approximation to Binomial Distribution, whether or not  $p$  is equal to  $q$ , the Binomial Distribution tends to the form of the continuous curve when  $n$  becomes large at least for the material part of the range. As a matter of fact, the correspondence between Binomial and the Normal curve is surprisingly close even for low values of  $n$  provide  $p$  and  $q$  are fairly near to equality. The limiting frequency curve, obtained as  $n$ , becomes large and is called the Normal frequency curve or simply the Normal curve.

Rahul Publications



## Exercise Problems

1. In Delhi with 100 municipal wards, each having approximately the same population, the distribution of Meningitis it is cases in 2007 was as follows:

No. of cases:	0	1	2	3	4
No. of wards:	63	28	6	2	1

Fit a Poisson Distribution for the above.

**[Ans: 60.64, 30.32, 7.58, 1.26, 0.16]**

2. The distribution of typing mistakes committed by a data entry operator is given below. Assuming a Poisson Model, find out the expected frequencies.

Mistakes per page:	0	1	2	3	4	5
No. of pages:	142	156	69	27	5	1

**[Ans: 147.16, 147.16, 73.58, 24.52, 6.13, 1.22]**

3. Fit a Poisson Distribution to the following data:

Number of mistakes per page:	0	1	2	3	4	Total
No. of pages:	109	65	22	3	1	200

**[Ans: 108.64, 66.27, 20.21, 4.11, 0.63]**

4. For a Binomial distribution, the mean and variance are respectively 4 and 3. Calculate the probability of getting a non-zero value of this distribution.

**[Ans :  $1 - \left(\frac{3}{4}\right)^{16}$ ]**

5. In a Binomial distribution consisting of 5 independent trials, probabilities of 1 and 2 successes are 0.4096 and 0.2048 respectively. Find the parameter 'p' of the distribution.

**[Ans: p = 0.2]**

6. A's chance of winning a single game against B is  $\frac{2}{3}$ . Find the chance that in a series of 5 games with B, A wins: (i) exactly three games, (ii) at least three games:

**[Ans: (i)  $\frac{80}{243}$ , (ii)  $\frac{64}{81}$ ]**

7. The customer accounts at a certain departmental store have an average balance of ₹ 480 and a standard deviation of ₹ 160. Assuming that the account balance are normally distributed.

- What proportion of the accounts is over ₹ 600?
- What proportion of the accounts is between ₹ 400 and ₹ 600?
- What proportion of the accounts is between ₹ 240 and ₹ 360?

**[Ans: (i) 22.66% (ii) 0.4649% (iii) 15.98%]**

### Choose the Correct Answer

1. Which of the following is not a condition of the binomial distribution [ c ]
  - (a) Only 2 possible outcomes
  - (b) have constant probability of success
  - (c) must have at least 3 trials
  - (d) None
2. A variable that can assume any possible value between two points is called: [ b ]
  - (a) Discrete random variable
  - (b) Continuous random variable
  - (c) Discrete sample space
  - (d) Random variable
3. A discrete probability distribution may be represented by: [ d ]
  - (a) Table
  - (b) Graph
  - (c) Mathematical equation
  - (d) All of the above
4. If C is a constant in a continuous probability distribution, then  $p(x = C)$  is always equal to: [ a ]
  - (a) Zero
  - (b) One
  - (c) Negative
  - (d) Impossible
5. If the random variable takes negative values, then the negative values will have: [ a ]
  - (a) Positive probabilities
  - (b) Negative probabilities
  - (c) Constant probabilities
  - (d) None
6. Which one is not an example of random experiment? [ d ]
  - (a) A coin is tossed and the outcome is either a head or a tail
  - (b) A six-sided die is rolled
  - (c) Some number of persons will be admitted to a hospital emergency room during any hour.
  - (d) All medical insurance claims received by a company in a given year.
7. The probability function is always [ c ]
  - (a) Negative
  - (b) Positive
  - (c) Non negative
  - (d) None
8. Area under the normal curve on either side of mean is [ a ]
  - (a) 0.5
  - (b) 1
  - (c) 2
  - (d) None
9. Normal Distribution is [ a ]
  - (a) Mesokurtic
  - (b) Leptokurtic
  - (c) Platykurtic
  - (d) None
10. Which of the following parameter control the relative flatness of normal distribution [ a ]
  - (a) Standard Deviation
  - (b) Mean
  - (c) Mode
  - (d) None

### Fill in the blanks

1. If a coin is tossed 6 times in succession, the probability of getting at least one head is \_\_\_\_\_.
2. The mean of the binomial distribution is \_\_\_\_\_.
3. If  $n$  and  $p$  are the parameters of a binomial distribution, the standard deviation of this distribution is \_\_\_\_\_.
4. The probability of having at least one tail in four throws with a coin is \_\_\_\_\_.
5. If mean of the binomial distribution is 8 and variance is 6, the mode of this distribution is \_\_\_\_\_.
6. If mean of the binomial distribution is 4 and variance is 2 then  $p =$  \_\_\_\_\_.
7. The probability of getting one boy in a family of 4 children is \_\_\_\_\_.
8. If the mean and variance of a binomial variate are 12 and 4, then the distribution is  $25$  \_\_\_\_\_.
9. In a binomial distribution the sum and product of the mean and variance are  $\frac{25}{3}$  and  $\frac{50}{3}$  respectively. The distribution is \_\_\_\_\_.
10. If the probability of a defective bolt is 0.1, the mean and standard deviation for the distribution bolts in a total of 400 are \_\_\_\_\_ and \_\_\_\_\_ respectively.

### ANSWERS

1.  $\frac{63}{64}$
2.  $np$
3.  $\sqrt{npq}$
4.  $\frac{15}{16}$
5. 8
6.  $\frac{1}{2}$
7.  $\frac{1}{4}$
8.  $\left(\frac{1}{3} + \frac{2}{3}\right)^{18}$
9.  $\left(\frac{2}{3} + \frac{1}{3}\right)^{15}$
10. 40, 6

## UNIT III

- (i) **Sampling Theory** : The basics of sampling-Sampling procedures-Random and Non-Random methods- Sample size determination-Sampling distribution, Standard Error, Central Limit Theorem.
- (ii) **Hypothesis Testing** : Statistical Estimation, Point and Interval Estimation, Properties of a Good Estimator, confidential interval.
- (iii) **Large Sample tests**: Test for one and two proportions, Test for one and two means, Test for two S.D's.

### 3.1 SAMPLING THEORY

#### 3.1.1 The Basics of Sampling

##### Q1. Define Sampling.

(OR)

##### What is Sampling?

*Ans :*

Study of entire population may not be possible to carry out and hence a part alone is selected from the given population. A portion of the population which is examined with a view to determining the population characteristics is called a sample, i.e., a sample is a subset of population and the number of objects in the sample is called the size of the sample. Size of the sample is denoted by  $n$ .

The process of selection of a sample is called sampling. It is quite often used in our day-to-day practical life.

##### Example :

- (i) To assess the quality of a bag of rice, sugar wheat or any other commodity, we examine only a portion of it by taking a handful of it from the bag and then decide to purchase it or not. The portion selected from the bag is called a sample, while the whole quantity of rice, sugar or wheat in the bag is the population.
- (ii) To estimate the proportion of defective articles in a large consignment, only a portion (i.e., a few of them) is selected and examined. The portion selected is a sample.
- (iii) Car produced in India is the population and the Nano cars is the sample.

##### Q2. What are the principles of sampling?

*Ans :*

##### 1. Law of Statistical Regularity

This law has its roots in the mathematical theory of probability, which states that a moderately large sample selected randomly from a universe is likely to represent the characteristics of population.

##### 2. Law of Inertia of Large Numbers

Law of Inertia of Large Numbers is the extension of law of Statistical Regularity. This law states that the large numbers are relatively more stable as compared to small numbers. Hence, other things remaining the same, as the sample size increases, the results or findings tend to be more reliable and accurate.

##### 3. Law of Persistence

This law states that there are certain inherent characteristics with the universe and the same will remain persisting and be reflected in the sample in same manner, even if the size of sample and that of universe is either increased or decreased.

##### 4. Law of Optimism

This law states that the size of sample should be fixed in such a manner that it produced the optimum results. This means maximum possible accuracy with minimum possible costs in terms of all kinds of resources to be used during the course of research.

##### 5. Law of Validity

This principle states that a sample should be planned and executed that the inferences

drawn from the sample remain valid for the universe or population from which samples were taken.

**Q3. Explain the Merits and Limitations of sampling.**

*Ans :* (June-19)

**Merits**

The following are the merits of sampling :

**1. Economical**

Since only a few units of population are studied in sample survey method hence, it is economical. This method saves money, resources and labour etc.

**2. Time Saving**

The process of investigation under sample survey method is time saving as only a limited number of items of population are studied.

**3. Identification of Error**

Because only a limited number of items are covered in sample survey method hence errors can be easily identified.

**4. Detailed Information**

Since numbers of items are less in sample inquiry, therefore, it is possible to obtain more detailed information from them.

**5. Flexible**

In comparison to census method, sample survey method is more flexible as it can be changed depending upon situation.

**6. More Scientific**

According to Fisher, sampling method is more scientific because the sample data can be conveniently investigated from various angles.

**Limitations**

Following are the demerits of sampling :

**1. Partial Investigation**

Since a sample is a part of the population it can not fully represent the population. Hence, it is a partial investigation of the population or universe.

**2. Difficulty in Selecting Representative Sample**

The success of sample survey method depends upon the selection of a representative sample. It is not an easy task to select a sample which would represent all the characteristics of population or universe.

**3. Specialized Knowledge**

There are several methods of selecting a sample from population. Specialized knowledge is required for choosing a representative sample from the population by applying the most appropriate technique. If a person does not have this knowledge he may commit serious mistakes.

**4. Limited Degree of Accuracy**

This technique is not suitable where a high degree of accuracy is needed. Sampling needs very careful planning, selection and execution otherwise result will be doubtful.

**5. Heterogeneous Population**

Sample survey method does not give correct result when universe or population is of heterogeneous nature.

**6. Possibility of Bias**

Investigator's business in case of certain methods of sampling, like the judgment sampling cannot be ruled out.

**Q4. What are the characteristics of a good sample ?**

*Ans :*

A good sample should possess the following characteristics :

**1. Representativeness**

A good sample must represent the population from which it is selected. If sample drawn is not a good representative of the universe, the conclusion drawn about the population will be misleading.

**2. Adequate Size**

There is no hard and fast rule regarding the size of sample but it should be adequate in

size. The size of sample should be in proportion to population size and should be determined in accordance with the purpose of statistical inquiry.

### 3. Unbiased

The sample should be unbiased. It should be free from the personal bias of the investigator.

### 4. Homogeneity

There should be homogeneity in the nature of units of population and that of the sample. It increases the accuracy of results.

### 5. Independence

The selection of units in the sample must be independent of one another. It means the selection of one unit should not affect the selection of another unit in the sample.

### 6. Random Selection

It means all the units of population have an equal chance of being selected in the sample.

#### 3.1.2 Sampling Procedures

#### Q5. Explain the process of sampling.

*Ans :*

#### Step 1: Identifying the Target Population

This is the first step of sampling process. It involves collecting the element which possesses the qualities regarding which inferences are to be made. The target population studies can be well defined in terms of units of sampling, time and extent. The element can be an object from which the information is to be obtained.

#### Example

Lakme wanted to assess the response of the customers, about a new line of compact powder and would like to take a sample of people who are above 19 years of age. In this case, it would be easier to collect a sample and the sampling unit would be the same as an element.

#### Step 2: Sampling Frame

The sampling frame represents the elements of target population. The sampling frame includes sets of direction for the identification of the target population.

#### Example

A mailing list is purchased from an organization, telephone book etc. Sometimes, it is possible that a list of population elements can be obtained but it might not include some elements of the population which would lead to sampling frame error. The researcher is required to identify and correct the sampling error.

#### Step 3: Selecting a Sampling Technique

It includes decisions of broader nature. There are several approaches available to the researcher such as the Bayesian approach, Probability sampling or Non-Probability sampling approach.

In case of Bayesian approach, the elements are selected in a sequence and are added to the sample. It also involves the process of collecting data, computation of statistics of sample and determination of its related costs. But this approach is not used widely as it does not provide complete information about elements of population and its related costs.

#### Step 4: Executing the Process of Sampling

This step involves a detailed specification of decisions related to the target population, sampling frame, techniques of sampling and sample size that are required to be implemented. This step requires the detailed information for all sampling design decisions.

### 3.2 RANDOM AND NON-RANDOM METHODS

#### Q6. Write briefly about the various sampling techniques.

(OR)

Explain various probabilistic sampling methods.

*Ans :* (Nov.-20, July-18, Aug.-17, Imp.)

The various technique of sampling can be divided into two broad categories probability sampling that is also known as random sampling and non-probability or non random sampling.

#### 1. Probability Sampling Methods

It is that method in which every item in the universe has a known chance or a probability of

being chosen for the sample. This implies that the selection of sample items is independent of the person making the study.

Probability sampling methods are discussed as follow :

### (A) Random Sampling

It is the most effective and commonly used method of selecting a sample. It is also known as 'chance sampling' (or) 'probability sampling'. This method provides an equal chance of selection to each and every item of the universe in the sample. **According to F.Yates**, "Random sample is the one in which every number of the parent population has had an equal chance of being included".

Following are some of the methods to establish randomness in framing the sample :

- (i) **Lottery Method** : In this method all the items or the units of the population are assigned different numbers which are noted on small paper slips, folded and put into a box. These are then properly reshuffled or mixed. A blindfold selection is then made from the number of slips required to constitute the desired sample size. The selection of items purely depends on chance.
- (ii) **Random Number Tables** : Some Random number tables have been developed by statisticians which can be used for selecting a sample from the given population of all these random number tables, Tippet's Table (1927) is most widely used. Tippet's random number table consists of 10,400 four-digit numbers, giving in all  $10,400 \times 4 = 41,600$  digits selected at random.

Selecting sample using Tippet's random number table involves following steps :

- First, arrange all the items of population in a serial order.
- Then by using Tippet's number table, numbers are selected on the basis of size of sample to be drawn. The numbers from the table can be selected either horizontally or vertically or diagonally.

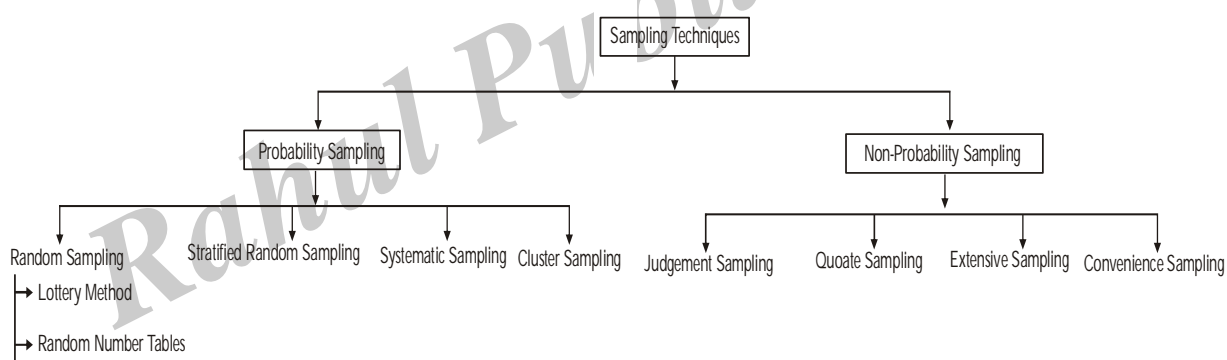


Fig.: Sampling Techniques Process

### (B) Stratified Random Sampling

Stratified sampling is based on the concept of homogeneity and heterogeneity. The process of dividing heterogeneous population into relatively homogeneous strata is termed as stratified sampling. It is a two-step process in which population is divided into subgroups or strata. The strata should be mutually exclusive and collectively exhaustive. The element within a stratum should be as homogeneous as possible, but the element in different stratum should be as heterogeneous as possible. It is pertinent to mention here that though the sampling is selected in various stages yet the last sample of the subject is studied.

When the population has different sectors with different characteristics i.e. the population is divided on heterogeneous basis, we cannot get representative sample by the random sampling technique. In this case we use stratified random sampling. In the first step, the population is divided into 'strata'

or groups on some homogeneous basis, and in the second step the selection of appropriate number of items is made from each sub-groups on random basis. The sum total of all the items taken separately from each sub-group or strata will form a stratified sample. Stratified sampling is much effectively used in market research where the division of the universe is fairly clear on the basis of occupational, economic, social or religious basis.

### (C) Systematic Sampling

Systematic sampling involves the selection of sample units at equal intervals, after all the units in the population are arranged in some systematic order such as alphabetical, chronological, geographical order etc. systematic sampling is also called 'quasi-random sampling'. In systematic sampling, the sample size is usually taken in such a way that it completely divides the population size. Let us suppose that  $N$  sampling units in the population are arranged in some systematic order and serially numbered 1 to  $N$ . Our sample size ' $n$ ' should be such that it completely divides  $N$ .  $\frac{N}{n} = K$  this  $K$  is called that sample interval. If  $K$  is in fraction then it is to be rounded off to get an integral value e.g., if we want to have a sample of size 5 from a population of size 100. Then  $K$  will be  $K = \frac{100}{5} = 20$ .

In this case the subsequent items are selected by taking every  $K^{\text{th}}$  items i.e. 20<sup>th</sup> items, refers to sample interval or sample ratio. The ratio of population size to the size of the sample.

### (D) Cluster Sampling

This sampling implies dividing population into clusters and drawing random sample either from all clusters or selected clusters. Cluster Sampling is similar to stratified sampling. In

the cluster sampling the universe is divided into number of relatively small subdivisions or clusters and then some of these clusters are randomly selected for inclusion in the overall sample. The element within cluster should be heterogeneous as possible, but cluster themselves should be as homogeneous as possible. The common form of cluster sampling is area or geographical sampling.

**For example,** if we are interested in obtaining the income or opinion data in a city, the whole city may be divided into  $N$  different blocks or localities (which determine the clusters) and a simple random sample of  $n$  blocks is drawn. The individuals in the selected blocks determine the cluster sample. The difference between cluster sampling and stratified sampling is that in case of cluster sampling only a sample of subgroups or clusters is chosen, whereas in stratified sampling all subpopulations or strata are selected for further sampling. The objectives of the both methods are also different. The objective of cluster sampling is to increase the efficiency by decreasing the costs, whereas the objective of stratified sampling is to increase the precision.

## 2. Non-Probability Sampling Methods

Non probability sample methods are those which do not provide every item in the universe with a known chance of being included in the sample. The selection process is atleast partially subjective. Non-probability sampling methods are discussed as follow :

### (A) Judgment Sampling

In judgment sampling, selection of sample units depends on the discretion or judgment of investigator. The investigator chooses the units from the universe according to his own judgment and includes all those items in the sample which he thinks best and typical of the universe.



**(B) Quota Sampling**

This method is suitable in making investigations concerning public opinion. Investigator define quotas according to some specific features of population like social classes, age groups and so on. The quotas confirm the total number of items in the sample taken as a whole. The selected sample units with the quotas depend on the personal judgments of the interviewer. For instance, in the market survey of tooth paste, the researcher may be asked to interview 100 people in certain areas of Shimla, and that out of 100 persons 40 are male, 40 are female and 20 are children under the age group of 15. With these quotas researcher is free to select the respondents to be interviewed.

**Merits**

The following are the merits of quota sampling :

**1. Reliability**

This method is more reliable as there is greater possibility of in important units of population in the sample.

**2. Public Opinion Surveys**

This method is more suitable for public opinion surveys.

**3. Organized Method**

This is more organized method than other non-probability methods. It enjoys the benefits of both stratified and purposive sampling.

**Demerits**

The following are the demerits :

**1. Prejudice**

Because of biased opinion of the investigator, it has less practical import negligence on the part of the investigator may lead to errors.

**2. Need More Skill**

The success of this method depends upon the expertise of the researcher. A little negligence on the part of researcher may lead to drastic error.

**(C) Extensive Sampling**

This method enables us to study a large universe. By this method a very large sample is collected and those items which are a sort of liability on the statistical operations or the items from which it is difficult to collect any information are dropped out.

**(D) Convenience Sampling**

In this method, the selection of sample is done on the basis of convenience of the investigator. This sampling is not representative of any definable population. A sample obtained from readily available lists such as telephone directory or trade directory is an example of convenience sampling. Sample drawn by this method is called convenience sample.

**Q7. Differentiate between random and non-random sampling methods.**

*Ans :*

(July-18, Imp.)

S.No.	Nature	Random Sampling	Non-random Sampling
1.	<b>Definition</b>	Random sampling is a sampling technique where each sample has an equal probability of getting selected	Non-random sampling is a sampling technique where the sample selected will be based on factors such as convenience, judgement and experience of the researcher and not on probability
2.	<b>Biases involved in Sampling</b>	Random sampling is unbiased in nature	Non-random sampling is biased in nature
3.	<b>Based on</b>	Based on probability	Based on other factors such as convenience, judgement and experience of researcher but, not based on probability
4.	<b>Representation of Population</b>	Random sampling is representative of the entire population	Non-random sampling lacks the representation of the entire population
5.	<b>Chances of Zero Probability</b>	Never	Zero probability can occur
6.	<b>Complexity</b>	Random sampling is the most simple sampling technique	Non-random sampling method is a somewhat complex sampling technique

### 3.3 SAMPLE SIZE DETERMINATION

**Q8. Explain about sample size determination.**

*Ans :*

The method of determining optimal sample size has been discussed in two estimation problems:

1. Determination of Sample Size 'n' when Estimating the Population Mean.
2. Determination of Sample Size 'n' when Estimating the Population Proportion.

#### 1. Determination of Sample Size (n) when Estimating the Population Mean

As we know that, for large samples, sample mean  $\bar{x}$  is an unbiased estimator of population mean  $\mu$ . The standard error of  $\bar{x}$  being. For estimating the sample size we need the following pre-assigned value :

1. The desired confidence level.
2. The Permissible sampling error (E).
3. The population standard deviation ( $\sigma$ ).

So,  $E = \bar{X} - \mu$

and standard error of  $\bar{X} = \frac{\sigma}{\sqrt{n}}$ ,

For large samples,  $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Or  $(\bar{x} - \mu) = z \frac{\sigma}{\sqrt{n}}$

Suppose the desired confidence level is 95%. Then  $z$  values defining 95% confidence level are

$$\pm 1.96, \text{ thus } \bar{x} - \mu = \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$

$$\text{or } E = \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$

$$\text{or } \sqrt{n} = \pm \frac{1.96 \times \sigma}{E}$$

$$n = \left( \frac{1.96 \times \sigma}{E} \right)^2$$

$$\text{or } n = \left( \frac{z\sigma}{E} \right)^2$$

## 2. Determining Sample Size (n) When Estimating the Population Proportion

For a sample of size  $n$ , sample proportion  $\hat{p}$  is used for estimating population proportion  $p$ .

The standard error of  $(\hat{p}) = \sqrt{\frac{pq}{n}}$  ( $p$  being known)

For large samples,

$$z = \frac{\hat{p} - p}{S.E.(\hat{p})}; \quad z = \frac{E}{\sqrt{\frac{pq}{n}}}$$

Where  $E = \hat{p} - p$  (sampling error)

### 3.4 SAMPLING DISTRIBUTION

**Q9. What is sampling distribution? Explain the properties of sampling distribution.**

*Ans :*

The symbols  $\bar{X}$  and  $S$  are used to designate the mean and standard deviation of the sample distribution. A measure characterizing a sample such as  $\bar{X}$  or  $S$ , is called a statistic. It may be noted that several sample distributions are possible from a given universe.

The sampling distribution of a statistic reveals some important features :

1. First, a sampling distribution is generated from a population distribution, known or assumed.
2. Secondly, the same population may generate an infinite number of sampling distributions for the statistic, each for special sample size  $n$ .
3. Finally, a population may generate sampling distributions for two or more different statistics.

Sampling distributions are of great importance in theory and practice of statistics. It is because of the fact that the sampling distribution of a statistic has well-defined properties and it is from these properties that we can calculate risks (errors due to chance) involved in making generalization about populations on the basis of samples.

#### Properties

1. The arithmetic mean of the sampling distribution is the same as the mean of the universe from which samples were taken. For this reason the mean of the sampling distribution may be denoted by the same symbol as that for the mean of the universe involved, namely  $\mu$ .
2. The sampling distribution of mean has a standard deviation (a standard error) equal to the population standard deviation divided by the square root of the sample size, i.e.,  $\frac{\sigma}{\sqrt{n}}$ .
3. The sampling distribution of means is normally distributed. Since the sampling distribution of the mean is normally distributed, we can develop a method in which we can use the mean of our sample to estimate the population mean. For example, if we compute an interval of  $\bar{X} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$ , the interval will include 95 per cent of all the sample means.

### 3.5 STANDARD ERROR

**Q10. Explain briefly about standard error? State the uses of standard error.**

*Ans :*

(Dec.-20)

The standard statistic is known as its Standard Error, abbreviated as S.E. The standard errors of some of the well-known statistics, for large samples, are given below, where  $n$  is the sample size,  $\sigma^2$  the population variance, and  $P$  the population proportion, and  $Q = 2 - P$ ;  $n_1$  and  $n_2$  represent the sizes of two independent random samples respectively drawn from the given population ( $s$ ).

S.No.	Statistics	Standard Error
1.	Sample mean: $\bar{x}$	$\sigma / \sqrt{n}$
2.	Observed sample proportion 'p'	$\sqrt{PQ / n}$
3.	Samples s.d. : $s$	$\sqrt{\sigma^2 / 2n}$
4.	Sample variance : $s^2$	$\sigma^2 \sqrt{2 / n}$
5.	Sample quartiles	$1.36263 \sigma / \sqrt{n}$
6.	Sample median	$1.25331 \sigma / \sqrt{n}$
7.	Sample correlation coefficient ( $r$ )	$(1 - \rho^2) / \sqrt{n}$ $\rho$ being the population correlation coefficient
8.	Sample moment: $\mu_3$	$\sigma^3 \sqrt{96 / n}$
9.	Sample moment: $\mu_4$	$\sigma^4 \sqrt{96 / n}$
10.	Sample coefficient of variation ( $v$ )	$\frac{v}{\sqrt{2n}} \sqrt{1 + \frac{2v^3}{10^4}} \approx \frac{v}{\sqrt{2n}}$
11.	Difference of two sample means : $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
12.	Difference of two sample s.d. $s$ . $(s_1 - s_2)$	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
13.	Difference of two sample proportions: $(p_1 - p_2)$	$\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}$

### 3.6 CENTRAL LIMIT THEOREM

**Q11. State the concept of Central Limit theorem.**

*Ans :* (Dec.-20, Imp.)

The central limit theorem states that as sample size is increased, the sampling distribution of the mean (and for other Sample statistics as well) approaches the normal distribution in form, regardless of the form of the population distribution from which the sample was taken.

For practical purposes, the sampling distribution of the mean can be assumed to be approximately normally distributed, even for the most non-normal populations or processes, whenever the sample size is  $n \geq 30$ . For populations that are 'only somewhat non-normal, even a smaller sample size will suffice. But a sample size of atleast 30 will take care of the most adverse population situation.

The real advantage of the central limit theorem is that sample data drawn from populations not normally distributed or from populations of unknown shape also can be analyzed by using the normal distribution, because the sample means are normally distributed for sample sizes of  $n \geq 30$ .

If a random sample of  $n$  observation is selected from any population, then, when the sample size is sufficiently large ( $n \geq 30$ ) the sampling distribution of the mean tends to approximate the normal distribution. The larger the sample size  $n$ , the better will be the normal approximation to the sampling distribution of the mean. Then, again in this case it can be shown that the mean of the sample means is same as population mean and the standard error of the mean is smaller than the population standard deviation.

Since the central limit theorem states that sample means are normally distributed regardless of the shape of the population for large samples and for any sample size with normally distributed population, thus sample means can be analyzed by using Z distribution.

When the samples are drawn from a normal population with mean  $\mu$  and standard deviation  $\sigma$ , the sampling distribution of the mean is also normal with mean  $\mu_{\bar{x}} = \mu$  and standard deviation

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ . However the sampling distribution of the mean, when the population is not normal is equally important.

The central limit theorem says that from any given population with mean  $\mu$  and standard deviation  $\sigma$ , if we draw a random sample of  $n$  observations, the sampling distribution of the mean will approach a normal distribution with a mean  $\mu$

and standard deviation  $\frac{\sigma}{\sqrt{n}}$  as the sample size increases and becomes large. In practice a sample size of 30 and above is considered to be large.

Thus the central limit theorem is a hall mark of statistical inference. It permits us to make inference about the population parameter based on random samples drawn from populations that are not necessarily normally distributed.

### 3.7 HYPOTHESIS TESTING

**Q12. Define Hypothesis. What are the characteristics of Hypothesis?**

*Ans :* (Dec.-20, June-19, Imp.)

#### Introduction

The term 'hypothesis' is derived from the ancient Greek word, 'hypothesis' that means 'to put under' (or) 'to suppose'. Hypothesis is also a combination of two words 'Hypo, Thesis' where 'Hypo' means tentative or subject to verification and 'Thesis' a statement based on concepts, theories and past experiences about the solution of the problem. The term hypothesis literally means an assumption or a supposition about the state of affairs of a certain thing or phenomena or facts or variable or situation. Thus, "hypothesis is perceived as a proposition or set of propositions set forth as an explanation for occurrence of some specified group of phenomenon either asserted merely as a provisional conjecture to guide some investigation or accepted as highly probable in the light of established facts. Quite often

a research hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent variable to some dependent variables"

### Definitions

The view point of various thinkers has been presented as under :

- (i) **According to Good, Barr and Scates** "Hypothesis is a statement temporarily accepted as true in the light of what is at the time, known about a phenomenon, and it is employed as a basis for action in the search for new truth. When hypothesis is fully established, it may take the form of facts, principles and theories".
- (ii) **According to Lundberg** "Hypotheses is a tentative generalization, the validity of which is remains to be tested. In the most elementary stage the hypothesis may be any hunch, guess, imaginative idea which become base for further investigation".
- (iii) **According to Best** "Hypothesis id a shrewd guess or inference that is formulated and provisionally adopted to explain observed facts or conditions and to guide in further investigation".
- (iv) **According to Mouly** "Hypothesis is an assumption whose testability is to be tested on the basis of the compatibility of its implications with empirical evidences and previous knowledge".
- (v) **According to Gopal** "Hypothesis is a tentative solution posed on cursory observation of known and available data and adopted provisionally to explain certain events and to guide in the investigation of others. It is in fact, a possible solution to the problem".
- (vi) **According to Theodorson and Theodorson:** "A hypotheses is a tentative statement asserting a relationship between certain facts".
- (vii) **According to Goode and Hutt:** "A proposition which can be put to a test to determine its validity".

(viii) **According to Kerlinger** "A hypothesis is a conjectural statement of the relationship between two or more variables".

(ix) **According to Black and Champion** "A hypotheses is a tentative statement about something, the validity of which is usually unknown".

(x) **According to While Bailey** "Hypothesis is a proposition that is stated in a testable form and that predicts a particular relationship between two or more variables".

(xi) **According to Grinnell** "Hypothesis is written in such a way that it can be proven or disproved by valid or reliable data it is in order to obtain these data that we perform our study".

(xii) **According to Palmar O Johnson:** "A hypothesis in statistics is simply a quantitative statement about a population".

(xiii) **According to Webster :** "Hypothesis is a tentative assumption made in order to draw out and test its logical or empirical consequences".

### Characteristics

The following the basic characteristics of a hypothesis :

#### 1. Valid

Hypothesis must be valid and related to the phenomena or situation which it is trying to explain.

#### 2. Pivot of Research

Hypothesis is backbone of all kinds of researches because the all research activities are designed to verify the hypothesis from 360 degree angle.

#### 3. Conceptual Clarity

Hypothesis must be clearly and precisely stated. There should be no ambiguity in the formulation of hypothesis. It means hypothesis should be defined lucidly, should be operationalised, should be commonly accepted and should be communicable.

#### 4. Testability

A hypothesis should be testable and not

moral judgement. It should be possible to collect empirical evidences to test the hypothesis. In the words of C. William Emory, "A hypothesis is testable if other deductions can be confirmed or disapproved by observation".

#### 5. **Specificity**

A hypothesis should be specific and explain the expected the relations between variables and the situations under which these relation will hold.

#### 6. **Consistency**

A hypothesis should be logically consistency. Two or more hypothesis logically derived from the same population must not be mutually contradictory.

#### 7. **Objectivity**

Hypothesis should be free from value judgement. In the scientific research the researcher's value system has no place in scientific enquiry.

#### 8. **Simplicity**

Hypothesis should be a simple one requiring fewer assumptions. But the simplicity does not mean vague idea.

#### 9. **Theoretical Relevance**

Hypothesis should be based upon some theoretical foundations. When a research is systematically based upon a body of existence knowledge, only then a genuine contribution is more likely to result in.

#### 10. **Availability of Technique**

Hypothesis should be related to available techniques, otherwise it will not be researchable. Hence, the researcher must ensure that statistical or mathematical techniques are available for testing the proposed hypothesis.

#### 11. **Future Oriented**

Hypothesis is forward looking concept as it is related to future verification not the past facts, information or situations.

#### **Q13. Explain the procedure for testing a hypothesis.**

(OR)

**Explain the procedure generally followed in testing of hypothesis.**

*Ans :* (Dec.-20, June-19, Imp.)

Test of Hypothesis involves the following steps :

#### **Step 1: Statement (or assumption) of hypothesis**

There are two types of hypothesis :

- (i) Null Hypothesis
- (ii) Alternative Hypothesis.

#### **(i) Null Hypothesis :**

For applying the tests of significance, we first set up a hypothesis a definite statement about the population parameter. Such a hypothesis is usually a hypothesis of no-difference, is called Null Hypothesis.

It is in the form  $H_0 : \mu = \mu_0$

$\mu_0$  is the value which is assumed or claimed for the population characteristic. It is the reference point against which the Alternative Hypothesis is set up, as explained in the next step

#### **Definition**

A null hypothesis is the hypothesis which asserts that there is no significant: difference between the statistic and the population parameter and whatever observed difference is there, is merely due to fluctuations in sampling from the same population. It is always denoted by  $H_0$  . To test whether one procedure is better than another, we assume that there is no difference between the procedures. Similarly to test whether there is a relationship between two variates , we take  $H_0$  that there is no relationship.

**For example,** in case of a single statistic,  $H_0$  will be that the sample statistic does not differ significantly from the hypothetical parameter value and in the case of two statistics ( $H_0$ ) will be that the sample statistics do not differ significantly.

**(ii) Alternative Hypothesis**

Any hypothesis which contradicts the Null Hypothesis is called an Alternative Hypothesis, usually denoted by  $H_1$ . The two hypothesis  $H_0$  and  $H_1$  are such that if one is true, the other is false and vice versa.

**For example**, if we want to test the null hypothesis that the population has a specified mean  $\mu_0$  (say) i.e.,  $H_0 : \mu = \mu_0$ , then the Alternative Hypothesis would be

$$(i) \quad H_1 : \mu \neq \mu_0 \text{ (i.e., either } \mu > \mu_0 \text{ (or) } \mu < \mu_0 \text{)}$$

(or)

$$(ii) \quad H_1 : \mu > \mu_0$$

(or)

$$(iii) \quad H_1 : \mu < \mu_0$$

**The Alternative Hypothesis**

(i) is known as a two-tailed alternative and the Alternative Hypothesis in

(ii) is known as right-tailed and in

(iii) is known as left-tailed.

The setting of alternative hypothesis is very important to decide whether we have to use a single-tailed (right or left) or two-tailed test.

Alternate Hypothesis is in one of the following forms :

$$H_1 : \mu \neq \mu_0$$

or  $H_1 : \mu > \mu_0$

or  $H_1 : \mu < \mu_0$

**Step 2 : Specification of the Level of Significance**

The level of significance denoted by  $\alpha$  is the confidence with which we reject or accept the Null hypothesis  $H_0$  i.e., it is the maximum possible probability with which we are willing to risk an error in rejecting  $H_0$  when it is true. The level of significance is generally specified before a test procedure so that the results obtained may not influence our decision. In practice, we take either 5% (i.e., 0.05) or 1% (i.e., 0.01) level of significance, although other levels such as 2%, 1/2% etc. may also be used. 5% Level of significance in a test procedure indicates that there are about 5 cases in 100 that we would reject the null hypothesis  $H_0$  when it is true i.e., we are about 95% confident that we have made the right decision. Similarly, in 1% Level of significance, there is only 1 case in 100 that the null hypothesis  $H_0$  is rejected when it is true i.e., we are about 99% confident that we have made the right decision. Level of significance is also known as the size of the test.

**Step 3 : Identification of the Test Statistic**

There are several tests of significance, viz., z, t, F etc. First we have to select the right test depending on the nature of the information given in the problem. Then we construct the test criterion and select the appropriate probability distribution.

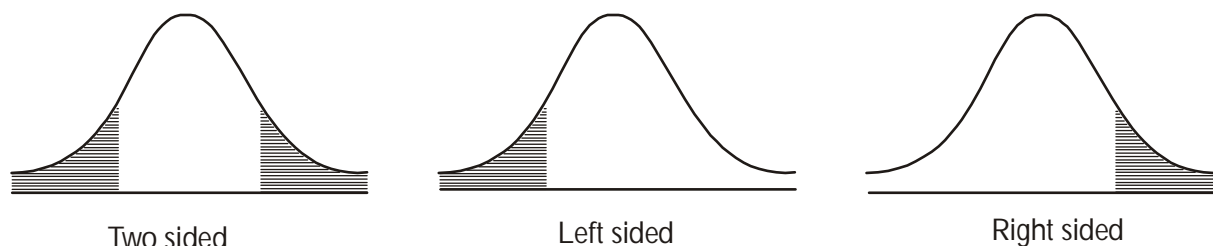
**Step 4: Critical Region**

The critical region is formed based on following factors.

(a) Distribution of the Statistic i.e., whether the statistic follows the normal, 't'  $\chi^2$  (or) 'F' distribution (will be discussed later).



- (b) **Form of Alternative Hypothesis:** If the form has  $\neq$  sign, the critical region is divided equally in the left and right tailed, sides of the distribution.



If the form of alternative hypothesis has  $<$  sign, the entire critical region is taken in the left tail of the distribution.

If the form of alternative hypothesis has  $>$  sign, the entire critical region is taken on the right side of the distribution.

### Step 5 : Making Decision

We compute the value of the appropriate statistic and ascertain whether the computed value falls in acceptance or rejection region depending on the specified Level of significance.

In finding the acceptance or rejection region we have to use critical values given in Statistical Tables. By comparing the computed value and the critical value decision is taken for accepting or rejecting  $H_0$ . If the computed value  $<$  critical value, we accept  $H_0$ , otherwise we reject  $H_0$ .

### Q14. Explain various types of errors in testing of hypothesis.

*Ans :*

When a statistical hypothesis is tested there are four possibilities :

1. The hypothesis is true but our test rejects it. (Type I error)
2. The hypothesis is false but our test accepts it. (Type II error)
3. The hypothesis is true and our test accepts it. (Correct decision)
4. The hypothesis is false and our test rejects it. (Correct decision)

Obviously, the first two possibilities lead to errors.

In a statistical hypothesis testing experiment, a Type I error is committed by rejecting the null hypothesis when it is true. The probability of committing a Type I error is denoted by  $\alpha$  (pronounced as alpha), where

$$\begin{aligned}\alpha &= \text{Prob. (Type I error)} \\ &= \text{Prob. (Rejecting } H_0/H_a \text{ is true)}\end{aligned}$$

On the other hand, a Type II error is committed by not rejecting (i.e., accepting) the null hypothesis when it is false. The probability of committing a Type II error is denoted by  $\beta$  (pronounced as beta), where

$$\begin{aligned}\beta &= \text{Probability (Type II error)} \\ &= \text{Probability (Not rejecting or accepting } H_0/H_a \text{ is false)}\end{aligned}$$

The distinction between these two types of errors can be made by an example. Assume that the difference between two population mean is actually zero. If our test of significance when applied to the sample means leads us to believe that the difference in population means is significant, we make a Type I

error. On the other hand, suppose there is true difference between the two population means. Now if our test of significance leads to the judgment "not significant", we commit a Type II error. We thus find ourselves in the situation which is described by the following table :

	Accept $H_0$	Reject $H_0$
$H_0$ is True	Correct Decision	Type I Error
$H_0$ is False	Type II Error	Correct Decision

While testing hypothesis the aim is to reduce both the types of error. Type I and Type II. But due to fixed sample size, it is not possible to control both the errors simultaneously. There is a trade-off between these types of errors: the probability of making one type of error can only be reduced if we are willing to increase the probability of making the other type of error. In order to get a low  $\alpha$ , we will have to put up with a high  $\beta$ . To deal with this trade-off in business situations, managers decide the appropriate level of significance by examining the costs or penalties attached to both types of errors.

It is more dangerous to accept a false hypothesis (Type II error) than to reject a correct one (Type I error). Hence we keep the probability of committing Type I error at a certain level, called the level of significance. The level of significance (also known as the size of the rejection region or size of the critical region or simply size of the test) is traditionally denoted by the Greek letter  $\alpha$ . In most statistical tests, the level of significance is generally fixed at 5 per cent. This means that the probability of accepting a true hypothesis is 95 per cent.

### 3.8 STATISTICAL ESTIMATION

#### 3.8.1 Point and Interval Estimation

**Q15. Define estimation. Explain different types of estimations.**

(OR)

**What is estimation?**

*Ans :*

(Nov.-20)

When data are collected by sampling from a population, the most important objective of statistical analysis is to draw inferences or generalisations about that population from the information embodied in the sample. Statistical estimation, or briefly estimation, is concerned with the methods by which population characteristics are estimated from sample information. It may be pointed out that the true value of a parameter is an unknown constant that can be correctly ascertained only by an exhaustive study of the population. However, it is ordinarily too expensive or it is infeasible to enumerate complete populations to obtain the required information. In case of finite populations, the cost of complete censuses may be prohibitive. and in case of infinite population, complete enumerations are impossible. A realistic objective may be to obtain a guess or estimate of the unknown true value or an interval of plausible values from the sample data and also to determine the accuracy of the procedure. Statistical estimation procedures provide us with the means of obtaining estimates of population parameters with desired degrees of precision. With respect to estimating a parameter\*, the following two types of estimates are possible:

1. Point estimates, and
2. Interval estimates.

#### 1. Point Estimates

Point estimate is a single number which is used as an estimate of the unknown population parameter. The procedure in point estimation is to select a random sample of  $n$  observations,  $x_1, x_2, \dots, x_n$  from a population  $f(x, \theta^*)$  and then to use some preconceived method to arrive from these observations at a number say  $\hat{\theta}$  (read theta hat) which we accept as an estimator of  $\theta$ . The estimator  $\theta$  is a single point on the real number scale and thus the name point estimation.  $\hat{\theta}$  depends on the random variables that generate the sample and hence, it too is a random variable with its own sampling distribution.

## 2. Interval Estimates

As distinguished from a point estimate which provides one single value of the parameter, an interval estimate of a population parameter is a statement of two values between which it is estimated that the parameter lies. An interval estimate would always be specified by two values, i.e., the lower one and the upper one. In more technical terms, interval estimation refers to the estimation of a parameter by a random interval, called the Confidence interval, whose end points  $L$  and  $U$  with  $L < U$ , are functions of the observed random variables such that the probability that the inequality  $L < \theta < U$  is satisfied in terms of pre-determined number,  $1 - \alpha$ .  $L$  and  $U$  are called the confidence limits and are the random end points of interval estimate. Since in an interval, estimate, we determine an interval of plausible values, hence the name interval estimation. Thus, on the basis of sample study, if we estimate the average income of the people living in a village as Rs. 875 it will be a point estimate. On the other hand, if we say that the average income could lie between Rs. 800 and Rs. 950, it will be an interval estimate.

On comparing these two methods of estimation we find that point estimation has an advantage as much as it provides an exact value for the parameter under investigation. This merit, however, is also the defect of a point estimate. Being a single point on the real number scale, a point estimate does not tell us how close the estimator is to the parameter being estimated. Moreover, in scientific investigation it is usually not necessary to know exact value of a parameter—rather it is desirable to have some degree of the confidence that the value obtained is within a certain range. The interval estimate does provide such confidence and hence interval estimate should generally be employed in practice.

## 3.8.2 Properties of a Good Estimator

### Q16. What are the Properties of a Good Estimator?

(OR)

State the properties of a good estimator.

Ans :

(Nov.-20, June-19)

A good estimator, as common sense dictates, is close to the parameter being estimated. Its quality is to be evaluated in terms of the following properties.

#### (i) Unbiasedness

An estimator is said to be unbiased if its expected value is identical with the population parameter being estimated. That is if  $\hat{\theta}$  is an unbiased estimate of  $\theta$ , then we must have  $E(\hat{\theta}) = \theta$ . Many estimators are "Asymptotically unbiased" in the sense that the biases reduce to practically insignificant values zero when  $n$  becomes sufficiently large. The estimator  $S^2$  is an example.

It should be noted that bias in estimation is not necessarily undesirable. It may turn out to be an asset in some situations. For example, it may happen that an unbiased estimator is less desirable than a biased estimator if the former has a greater variability than the latter and, as a consequence, the expected value of the latter is closer than that of the former to the parameter being estimated.

#### (ii) Consistency

If an estimator, say  $\hat{\theta}$ , approaches the parameter  $\theta$  closer and closer as the sample size  $n$  increases,  $\hat{\theta}$  is said to be a consistent estimator of  $\theta$ . Stating somewhat more rigorously, the estimator  $\hat{\theta}$  is said to be a consistent estimator of  $\theta$  if, as  $n$  approaches infinity, the probability approaches 1 that  $\hat{\theta}$  will differ from the parameter  $\theta$  by not more than an arbitrary small constant.

The sample mean is an unbiased estimator of  $\mu$  no matter what form the population distribution assumes, while the sample median is an unbiased estimate of  $\mu$  only if the population distribution is symmetrical. The sample mean is better than the sample median as an estimate of  $\mu$  in terms of both unbiasedness and consistency.

In case of large samples consistency is a desirable property for an estimator to possess. However in small samples, consistency is of little importance unless the limit of probability defining consistency is reached even with a relatively small size of the sample.

### (iii) Efficiency

The concept of efficiency refers to the sampling variability of an estimator. If two competing estimators are both unbiased, the one with the smaller variance (for a given sample size) is said to be relatively more efficient. Stated in a somewhat different language, estimator  $\hat{\theta}_1$  is said to be more efficient than another estimator  $\hat{\theta}_2$  for  $\theta$  if the variance of the first is less than the variance of the second. The smaller the variance of the estimator, the more concentrated is the distribution of the estimator around the parameter being estimated and, therefore, the better this estimator is.

If the population is symmetrically distributed, then both the sample mean and the sample median are consistent and unbiased estimators of  $\mu$ . Yet the sample mean is better than the sample median as an estimator of  $\mu$ . This claim is made in terms of efficiency.

### (iv) Sufficiency

An estimator is said to be sufficient if it conveys as much information as is possible about the parameter which is contained in the sample. The significance of sufficiency lies in the fact that if a sufficient estimator exists, it is absolutely unnecessary to consider any other estimator; a sufficient estimator ensures that all information a sample can furnish with respect to the estimation of a parameter is being utilized.

Many methods have been devised for estimating parameters that may provide estimators satisfying these properties. The two important methods are the least square method and the method of maximum likelihood.

### 3.8.3 Confidential Interval

#### Q17. Explain about confidence Limits - Population mean.

Ans :

(Aug.-17)

If we consider a large sample of size  $n$  from an infinite population with population mean  $m$  and the sample mean is then the standard normal variate (SNV) will be

$$Z = \frac{\bar{x} - E(\bar{x})}{S.E.(\bar{x})} = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

then the confidence limits for the population mean are given by

$$\bar{x} \pm z_{\alpha} \frac{\sigma}{\sqrt{n}}$$

$$\text{i.e., } Z = \left[ \bar{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}} \right] = 1 - \alpha$$

If  $\alpha = 0.05$ , then the confidence limits for population mean at 5% level of significance will be

$$\Pr \left[ \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right] = 0.95.$$

If sampling is with replacement from a finite population of size  $N$ , then confidence limits are given by

$$\bar{x} \pm z_{\alpha} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

### 1. Confidence Limits for Population Proportion

Finite and Infinite Population,

$$p \pm Z_c \sqrt{\frac{pq}{n}}$$

Where,

$p$  – Sample proportion of success

$q$  – Sample proportion of failure

$n$  – Size of the sample

$Z_c$  – Corresponding 'Z' value for given confidence level.

### 3. Confidence Limits for Difference of Means

Two samples of size each  $> 30$

$$(\bar{x}_1 - \bar{x}_2) \pm Z_c \sigma_{\bar{x}_1 - \bar{x}_2}$$

Where,  $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

$(\bar{x}_1 - \bar{x}_2)$  – Difference between two sample means

$Z_c$  – Corresponding 'Z' value for given confidence level

$n_1$  – Size of the first sample

$n_2$  – Size of the second sample

$\sigma_1$  – Standard deviation of first sample

$\sigma_2$  – Standard deviation of second sample

$\sigma_{\bar{x}_1 - \bar{x}_2}$  – Standard error of the difference of means.

### 4. Confidence Limits for Difference of Proportions

Two (infinite) Population Proportions

$$(P_1 - P_2) \pm Z_c \sigma_{p_1 - p_2}$$

Where,

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$(p_1 - p_2)$  – Difference of proportions

$\sigma_{p_1-p_2}$  – Standard error of the difference of proportions

$n_1, n_2$  – Sample sizes of 1<sup>st</sup> and 2<sup>nd</sup> sample respectively

$p_1, p_2$  – Success proportions of 1<sup>st</sup> and 2<sup>nd</sup> sample respectively

$q_1, q_2$  – Failure proportions of 1<sup>st</sup> and 2<sup>nd</sup> sample respectively.

**Note:**

Z values for various confidence levels from normal table.

Left tailed Z	–	–2.33	–2.05	–1.645	–1.28
Confidence Level	100%	99%	98%	95%	90%
Two tailed Z	3	2.58	2.33	1.96	1.645
Right tailed Z	–	2.33	2.05	1.645	1.28

**PROBLEMS**

1. The quality department of a wire manufacturing company periodically selects a sample of wire specimens in order to test for breaking strength. Past experience has shown that the breaking strengths of a certain type of wire are normally distributed with standard deviation of 200 kg. A random sample of 64 specimens gave a mean of 6,200 kg. The quality control supervisor wanted a 95 percent confidence interval for the mean breaking strength of the population.

*Sol:*

The  $Z_{\alpha}$  value corresponding to a confidence coefficient of 0.95 is 1.96. Therefore the limits are :

$$\begin{aligned} & \bar{x} \pm (1.96) \frac{\sigma}{\sqrt{n}} \\ \Rightarrow & 6200 \pm (1.96) \frac{200}{\sqrt{64}} \\ \Rightarrow & 6200 \pm 1.96 \times 25 \\ \Rightarrow & 6151 \text{ to } 6249 \end{aligned}$$

Hence, 95 percent confidence limits for mean  $m$  are 6151 to 6249.

2. A manager wants an estimate of average sales of salesman in his company. A random sample of 100 out of 500 salesman is selected and average sales is found to be Rs.750 (thousand). If population standard deviation is Rs. 150 (thousand), manager specifies a 98% level of confidence. What is the interval estimate for average sales of salesman?

*Sol:*

Here  $N = 500$ ,  $n = 100$ ,

$\bar{x} = 750$  and  $\sigma = 150$ .

The confidence limits are given by

$$\bar{x} \pm z_{\alpha} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

we have,

$$\begin{aligned}\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} &= \frac{150}{\sqrt{100}} \\ \sqrt{\frac{500-100}{500-1}} &= \frac{150}{\sqrt{100}} = \sqrt{\frac{400}{499}} \\ &= 15(0.895) = 13.429\end{aligned}$$

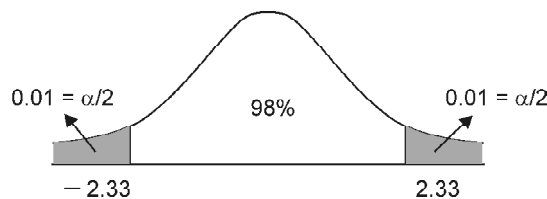


Fig. : 98% Confidence limits

the value of  $Z_{\alpha} = 2.33$  and the required confidence interval is

$$750 \pm 2.33 (13.429) \Rightarrow 718.71 \text{ to } 781.29 \text{ (in thousands)}$$

Thus it can be stated that for a 98% level of confidence, the population mean falls within the interval Rs.718.71 to Rs.781.29 (in thousands)

3. A sample of 150 items for machine A had an average life of 1,400 hours. A similar sample of 100 items from machine B had a mean life of 1,200 hours. Past records indicate that the standard deviation of the items produced by machine A is 120 hours and by machine B is 80 hours. Find 98 percent and 95 percent confidence limits for the difference of the average lifetimes for the populations of the items produced by the two machines.

Sol.:

The confidence limits for the difference of population means are given by

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha} \sigma_{\bar{x}_1 - \bar{x}_2}$$

$$\text{i.e. } (\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The  $Z_{\alpha}$  value corresponding to a confidence coefficient of  $(1-\alpha) = 0.98$  is 2.33 Therefore, the limits are

$$(\bar{x}_1 - \bar{x}_2) \pm 2.33 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

(Since the samples are from two different machines, are independent).

$$\bar{x}_1 - \bar{x}_2 = 1400 - 1200 = 200$$

and 
$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(120)^2}{150} + \frac{(80)^2}{100}} = 12.65$$

Therefore, the required 98 percent confidence limits are :

$$200 \pm 2.33 \times 12.65 = 170.5255$$

to 299.4745

The 95 percent confidence limits are :

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\bar{x}_1 - \bar{x}_2 = 1400 - 1200 = 200, z_{\alpha} = 1.96$$

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(120)^2}{150} + \frac{(80)^2}{100}} = 12.65$$

Therefore, the corresponding 95 percent confidence limits are

$$200 \pm 1.96 \times 12.65 = 175.206$$

to 224.794.

Hence, the 98 percent confidence limits are {170.5255 to 229.4745} for the average life time of the items produced by the two machines A and B. The 95 percent confidence limits are 175.206 to 224.794} for the difference average lifetime of the items produced by the two machines A and B.

4. **The Director, Human Resources of a large organization wanted to know what proportion of all persons who had ever been interviewed for a job in his organization has been hired? He was willing to settle of 95 percent confidence interval. A random sample of 500 interview records revealed that 76 of them had been hired.**

*Sol:*

The 95 percent confidence interval for the population proportion is given by

$$p \pm 1.96 \sqrt{\frac{pq}{n}}$$

Since the sample proportion =  $p = (76/500) = 0.152$

Then, the confidence interval will be

$$0.152 \pm 1.96 \sqrt{\frac{0.152 \times 0.848}{500}}$$

$$\Rightarrow 0.152 \pm 1.96(0.016)$$

$$\Rightarrow 0.152 \pm 0.031 = (0.121, 0.183)$$

Hence, the required proportion varies between 0.121 and 0.183.



5. A random sample of 160 people from a city A shown that 50 are smokers. Another sample of 250 people from another city B shown that 50 are smokers. At 95% confidence, what is difference of proportion of smokers ? ( $Z_{\alpha} = 1.96$ )

*Sol:*

Given that

$$n_1 = 160 \text{ and}$$

$$p_1 = \text{Proportion of smokers in city A}$$

$$= 50/160$$

$$= 0.3125 \text{ and}$$

$$n_2 = 250 \text{ and}$$

$$p_2 = \text{Proportion of smokers in city B}$$

$$= 50/250$$

$$= 0.20$$

$$q_1 = 1 - p_1$$

$$= 1 - 0.3125$$

$$= 0.6875$$

$$q_2 = 1 - p_2$$

$$= 1 - 0.20 = 0.80$$

Then,

$$\begin{aligned} \text{S.E. } (p_1 - p_2) &= \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \\ &= \sqrt{\frac{0.3125 \times 0.6875}{100} + \frac{0.2 \times 0.8}{250}} \\ &= 0.0445 \end{aligned}$$

(Since  $P_1, P_2$  are unknown, their unbiased sample estimates  $p_1, p_2$  are used). Therefore, 95% confidence limits are given by

$$(p_1 - p_2) \pm Z_{\alpha} \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$\Rightarrow (0.3125 - 0.2) \pm 1.96 (0.0445)$$

$$\Rightarrow 0.1125 \pm 1.96 \times 0.0445 = 0.1125 \pm 0.0873$$

$$\Rightarrow (0.0253 \text{ to } 0.1997)$$

Hence the difference of population proportion of smokers with 95% confidence limits are : **(0.0253 to 0.1997)**

### 3.9 LARGE SAMPLE TESTS

**Q18. Explain the concept of Large Sample Test.**

*Ans :*

These are basically concerned with sampling of variables such as height, weight etc., which may take any value. If the sample size is greater than 30 i.e.,  $n > 30$ , then those samples may be regarded as large samples. There is difference between large and small samples in using the test of significance, because the assumption we make for the two samples are also not the same.

#### Assumptions

The assumptions made for large samples are:

1. The random sampling distribution of statistics is approximately normal.
2. Sampling values are sufficiently close to the population value and can be used for the calculation of standard error of estimate.

In the case of large samples, when we are testing the significance of statistic, the concept of standard error is used.

#### 3.9.1 Test for One Proportion

**Q19. Explain the hypothesis testing for one proportion.**

*Ans :*

If  $X$  is the number of successes in  $n$  independent trials with constant probability  $P$  of success for each trial, then

$E(X) = nP$  and  $V(X) = nPQ$ , where  $Q = 1 - P$ , is the probability of failure.

It has been proved that for large  $n$ , the binomial distribution tends to normal distribution. Hence for large  $n$ ,  $X \sim N(nP, nPQ)$ , i.e.,

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - nP}{\sqrt{nPQ}} \sim N(0, 1)$$

and we can apply the normal test.

1. In a sample of size  $n$ , let  $X$  be the number of persons possessing the given attribute. Then

Observed proportion of successes =  $X/n = p$ , (say).

$$\therefore E(p) = E(X/n) = \frac{1}{n} E(X) = \frac{1}{n} nP = P$$

Thus the sample proportion 'p' gives an unbiased estimate of the population proportion  $P$ .

$$\text{Also } V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} V(X) = \frac{1}{n^2} nPQ = \frac{PQ}{n} \Rightarrow \text{S.E.}(p) = \sqrt{\frac{PQ}{n}} \quad -044b)$$

Since  $X$  and consequently  $X/n$  is asymptotically normal for large  $n$ , the normal test for the proportion of successes becomes :

$$Z = \frac{p - E(p)}{S.E(p)}$$

2. If we have sampling from a finite population of size  $N$ , then
3. Since the probable limits for a normal variate  $X$  are  $E(X) \pm 3\sqrt{V(X)}$ , the probable limits for the observed proportion of successes are :

$$E(p) \pm 3S.E.(p), \text{ i.e., } \pm 3\sqrt{PQ/n}.$$

If  $P$  is not known then taking  $p$  (the sample proportion) as an estimate of  $P$ , the probable limits for the proportion in the population are :  $p \pm 3\sqrt{pq/n}$

However, the limits for  $P$  at level of significance  $\alpha$  are given by :  $p \pm Z_{\alpha}\sqrt{pq/n}$ ,

where  $z_{\alpha}$  is the significant value of  $Z$  at level of significance  $\alpha$ .

In particular : 95% confidence limits for  $P$  are given by :  $p \pm 1.96\sqrt{pq/n}$ ,

and 99% confidence limits for  $P$  are given by :  $p \pm 2.58\sqrt{pq/n}$ .

6. **A die is thrown 9,000 times and a throw of 3 or 4 is observed 3,240 times. Show that the die cannot be regarded as an unbiased one and find the limits between which the probability of a throw of 3 or 4 lies.**

*Sol :*

If the coming of 3 or 4 is called a success, then in usual notations :

$n = 9,000$ ;  $X = \text{Number of successes} = 3,240$

Under the null hypothesis ( $H_0$ ) that the die is an unbiased one, we get

$$P = \text{Probability of success} = \text{Probability of getting a 3 or 4} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Alternative hypothesis,  $H_1$ :  $p \neq \frac{1}{3}$ , (i.e., die is biased).

We have  $Z = \frac{X - nP}{\sqrt{nQP}} \sim N(0,1)$ , since  $n$  is large.

$$\text{Now } Z = \frac{3240 - 9000 \times (1/3)}{\sqrt{9000 \times (1/3) \times (2/3)}} = \frac{240}{\sqrt{2000}} = \frac{240}{44.73} = 5.36$$

Since  $|Z| > 3$ ,  $H_0$  is rejected and we conclude that the die is almost certainly biased.

Since die is not unbiased,  $P \neq \frac{1}{3}$ . The probable limits for 'P' are given by :

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = p \pm 3\sqrt{pq/n}, \text{ where } \hat{p} = p = \frac{3,240}{9,000} = 0.36 \text{ and } \hat{Q} = q = 1 - p = 0.64.$$

Probable limits for population proportion of successes may be taken as :

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = 0.36 \pm 3\sqrt{\frac{0.36 \times 0.64}{9000}} = 0.36 \pm 3 \times \frac{0.6 \times 0.8}{30\sqrt{10}} = 0.345 \text{ and } 0.375.$$

Hence the probability of getting 3 or 4 almost certainly lies between 0.345 and 0.375.

- 7. A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that the S.E. of the proportion of bad ones in a sample of this size is 0.015 and deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8.5 and 17.5.**

*Sol.:*

Here we are given :  $n = 500$

$X =$  Number of bad pineapples in the sample = 65

$$p = \text{Proportion of bad pineapples in the sample} = \frac{65}{500} = 0.13 \Rightarrow q = 1 - p = 0.87$$

Since  $P$ , the proportion of bad pineapples in the consignment is not known, we may take (as in the last example):  $\hat{P} = p = 0.13$ ,  $\hat{Q} = q = 0.87$ .

$$\text{S.E. of proportion} = \sqrt{\hat{P}\hat{Q}/n} = \sqrt{0.13 \times 0.87 / 500} = 0.015$$

Thus, the limits for the proportion of bad pineapples in the consignment are :  $A/T_A$

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = 0.130 \pm 3 \times 0.015 = 0.130 \pm 0.045 = (0.085, 0.175)$$

Hence the percentage of bad pineapples in the consignment lies almost certainly between 8.5 and 17.5.

- 8. A random sample of 500 apples was taken from a large consignment and 60 were found to be bad. Obtain the 98% confidence limits for the percentage of bad apples in the consignment.**

*Sol.:*

We have:

$$p = \text{Proportion of bad apples in the sample} = \frac{60}{500} = 0.12$$

Since significant value of  $Z$  at 98% confidence coefficient (level of significance 2%) is 2.33, [from Normal Tables], 98% confidence limits for population proportion are :

$$\begin{aligned} p \pm 2.33\sqrt{pq/n} &= 0.12 \pm 2.33\sqrt{0.12 \times 0.88 / 500} = 0.12 \pm 2.33 \times \sqrt{0.0002112} \\ &= 0.12 \pm 2.33 \times 0.01453 = (0.08615, 0.15385) \end{aligned}$$

Hence 98% confidence limits for percentage of bad apples in the consignment are (8.61, 15.38).

9. In a sample of 1,000 people in Maharashtra, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this State at 1% level of significance?

*Sol :*

In the usual notations, we are given :  $n = 1,000$

$X =$  Number of rice eaters  $= 540$

$$\therefore p = \text{Sample proportion of rice eaters} = \frac{X}{n} = \frac{540}{1000} = 0.54$$

Null Hypothesis,  $H_0$  : Both rice and wheat are equally popular in the State so that

$P =$  Population proportion of rice eaters in Maharashtra  $= 0.5 \Rightarrow Q = 1 - P = 0.5$ .

Alternative Hypothesis,  $H_1$  :  $P \neq 0.5$  (two-tailed alternative)

Test Statistic. Under  $H_0$ , the test statistic is :

$$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0,1), \text{ (since } n \text{ is large).}$$

$$\text{Now } Z = \frac{0.54 - 0.50}{\sqrt{0.5 \times 0.5 / 1000}} = \frac{0.04}{0.0138} = 2.532$$

### Conclusion

The significant or critical value of  $Z$  at 1% level of significance for two-tailed test is 2.58. Since computed  $Z = 2.532$  is less than 2.58, it is not significant at 1% level of significance. Hence the null hypothesis is accepted and we may conclude that rice and wheat are equally popular in Maharashtra State.

10. Twenty people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survival rate, if attacked by this disease, is 85% in favour of the hypothesis that it is more, at 5% level. (Use Large Sample Test.)

*Sol :*

In the usual notations, we are given :  $n = 20$ .

$X =$  Number of persons who survived after attack by a disease  $= 18$

$$p = \text{Proportion of persons survived in the sample} = \frac{18}{20} = 0.90$$

Null Hypothesis,  $H_0$  :  $P = 0.85$ , i.e., the proportion of persons survived after attack by a disease in the lot is 85%.

Alternative Hypothesis,  $H_1$  :  $P > 0.85$  (Right-tailed alternative).

Test Statistic. Under  $H_0$ , the test statistic is :  $p - P$

$$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0,1), \text{ (since sample is large).}$$

$$\text{Now } Z = \frac{0.90 - 0.85}{\sqrt{0.85 \times 0.15 / 20}} = \frac{0.05}{0.079} = 0.633$$

### Conclusion

Since the alternative hypothesis is one-sided (right-tailed), we shall apply right-tailed test for testing significance of Z. The significant value of Z at 5% level of significance for right-tailed test is + 1.645. Since computed value of Z = 0.633 is less than 1.645, it is not significant and we may accept the null hypothesis at 5% level of significance.

**11. Work sampling studies are conducted to find the utilization of a machine. Out of 200 observations made, only 40 observations indicated the machine to be idle. Find the number of observations to be made in order to satisfy 95% confidence to state the utilization of machine with expected accuracy of  $\pm 5\%$ .**

*Sol :*

The sample size for n estimating the population proportion P with confidence coefficient  $(1 - \alpha) = 0.95$  is given by the equation :

$$P_r[1p - P1 \leq 1.96\sqrt{PQ/n}] = 0.95 \quad \dots(*)$$

$$\text{We want } P_r[1p - P1 < 0.05] = 0.95. \quad \dots(**)$$

$$\text{From (*) and (**), we obtain } 1.96\sqrt{PQ/n} = 0.05 \Rightarrow n = \frac{PQ(1.96)^2}{(0.05)^2}$$

$$\text{Since P is not known, we may use its sample estimate } \hat{p} = p = \frac{40}{200} = 0.2.$$

$$\therefore n = \frac{0.2 \times (1 - 0.2) \times 3.8416}{0.0025} = \frac{0.6147}{0.0025} = 245.88 \approx 246.$$

### 3.9.2 Test for Two Proportions

**Q20. Explain the hypothesis testing for two proportions.**

*Ans :*

Test of Significance for Difference of Proportions. Suppose we want to compare two distinct populations with respect to the prevalence of a certain attribute, say A, among their members. Let  $X_1, X_2$  be the number of persons possessing the given attribute A in random samples of sizes  $n_1$  and  $n_2$  from the two populations respectively. Then sample proportions are given by :  $p_1 = X_1/n_1$  and  $p_2 = X_2/n_2$ .

If  $P_1$  and  $P_2$  are population proportions, then

$$E(p_1) = P_1, E(p_2) = P_2$$

$$\text{and } V(p_1) = \frac{P_1 Q_1}{n_1} \text{ and } V(p_2) = \frac{P_2 Q_2}{n_2}$$

Since for large samples,  $p_1$  and  $p_2$  are independently and asymptotically normally distributed,  $(p_1 - p_2)$  is also normally distributed. Then the standard variable corresponding to the difference  $(p_1 - p_2)$  is given by :

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0, 1)$$

Under the null hypothesis,  $H_0 : P_1 = P_2$ , i.e., there is no significant difference between the sample proportions, we have

$$E(p_1 - p_2) = E(P_1) - E(P_2) = P_1 - P_2 = 0 \quad (\text{Under } H_0)$$

Also  $V(p_1 - p_2) = V(p_1) + V(p_2)$ ,

the covariance term  $\text{Cov}(p_1, p_2)$  vanishes, since sample proportions are independent.

$$\Rightarrow V(p_1 - p_2) = \frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2} = PQ \left( \frac{1}{n_1} + \frac{1}{n_2} \right),$$

$$[\because \text{under } H_0 : P_1 = P_2 = P \text{ (say), and } Q_1 = Q_2 = Q].$$

Hence, under  $H_0 : P_1 = P_2$ , the test statistic for the difference of proportions becomes :

$$Z = \frac{p_1 - p_2}{\sqrt{PQ \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0, 1)$$

In general, we do not have any information as to the proportion of A's in the populations from which the samples have been taken. Under  $H_0 : P_1 = P_2 = P$  (say), an unbiased estimate of the population proportion  $P$ , based on both the samples is

given by : 
$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

The estimate is unbiased, since

$$\begin{aligned} E(\hat{P}) &= \frac{1}{n_1 + n_2} [n_1 p_1 + n_2 p_2] = \frac{1}{n_1 + n_2} [n_1 E(p_1) + n_2 E(p_2)] \\ &= \frac{1}{n_1 + n_2} (n_1 P_1 + n_2 P_2) = P \quad [\because P_1 = P_2 = P, \text{ under } H_0] \end{aligned}$$

Thus, along with gives the required test statistic.

### Remarks

1. Suppose we want to test the significance of the difference between  $p_1$  and  $p$ , where  $p = \frac{(n_1 p_1 + n_2 p_2)}{(n_1 + n_2)}$

gives a pooled estimate of the population proportion on the basis of both the samples.

We have  $V(p_1 - p) = V(p_1) + V(p) - 2 \text{Cov}(p_1, p)$ .

Since  $p_1$  and  $p$  are not independent,  $\text{Cov}(p_1, p) \neq 0$ . ... (\*)

$$\text{Cov}(p_1, p) = E[\{p_1 - E(p_1)\} \{p - E(p)\}]$$

$$\begin{aligned}
&= E \left[ \{p_1 - E(p_1)\} \left\{ \frac{1}{n_1 + n_2} \{n_1 p_1 + n_2 p_2 - E(n_1 p_1 + n_2 p_2)\} \right\} \right] \\
&= \frac{1}{n_1 + n_2} E[\{p_1 - E(p_1)\} \{n_1(p_1 - E(p_1)) + n_2(p_2 - E(p_2))\}] \\
&= \frac{1}{n_1 + n_2} [n_1 E\{p_1 - E(p_1)\}^2 + n_2 E\{(p_1 - E(p_1))(p_2 - E(p_2))\}] \\
&= \frac{1}{n_1 + n_2} [n_1 V(p_1) + n_2 \text{Cov}(p_1, p_2)] = \frac{1}{n_1 + n_2} n_1 V(p_1), \quad [\because \text{Cov}(p_1, p_2) = 0] \\
&= \frac{n_1}{n_1 + n_2} \cdot \frac{pq}{n_1} = \frac{pq}{n_1 + n_2}
\end{aligned}$$

$$\begin{aligned}
\text{Var}(p) &= \text{Var} \left[ \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \right] = \frac{1}{(n_1 + n_2)^2} \text{Var}(n_1 p_1 + n_2 p_2) \\
&= \frac{1}{(n_1 + n_2)^2} [n_1^2 \text{Var}(p_1) + n_2^2 \text{Var}(p_2)]
\end{aligned}$$

covariance term vanishes since  $p_1$  and  $p_2$  are independent.

$$\therefore \text{Var}(p) = \frac{1}{(n_1 + n_2)^2} \left( n_1^2 \cdot \frac{pq}{n_1} + n_2^2 \cdot \frac{pq}{n_2} \right) = \frac{pq}{n_1 + n_2}$$

Substituting (\*) and simplifying, we shall get

$$V(p_1 - p) = \frac{pq}{n_1} + \frac{pq}{n_1 + n_2} - 2 \frac{pq}{n_1 + n_2} = pq \left[ \frac{n_2}{n_1(n_1 + n_2)} \right]$$

Also  $E(P_1 - P) = E(P_1) - E(P) = P - P = 0$

Thus, the test statistic in this case becomes :

$$Z = \frac{(p_1 - p) - E(p_1 - p)}{\text{S.E.}(p_1 - p)} = \frac{p_1 - p}{\sqrt{\left\{ \frac{n_2}{(n_1 + n_2)} \cdot \frac{pq}{n_1} \right\}}} \sim N(0, 1)$$

- Suppose the population proportions  $P_1$  and  $P_2$  are given to be distinctly different, i.e.,  $P_1 \neq P_2$  and we want to test if the difference  $(P_1 - P_2)$  in population proportions is likely to be hidden in simple samples of sizes  $n_1$  and  $n_2$  from the two populations respectively.

We have seen that in the usual notations,



$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{S.E.(p_1 - p_2)} = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\left(\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}\right)}} \sim N(0, 1)$$

Here sample proportions are not given. If we set up the null hypothesis  $H_0 : p_1 = p_2$ , i.e., the samples will not reveal the difference in the population proportions or in other words the difference in population proportions is likely to be hidden in sampling, the test statistic becomes:

- 12. Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women severe in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal, are same against that they are not, at 5% level.**

*Sol :*

Null Hypothesis  $H_0 : P_1 = P_2 = P$  (say), i.e., there is no significant difference between the opinions of men and women as far as proposal of flyover is concerned.

Alternative Hypothesis,  $H_1 : P_1 \neq P_2$  (two-tailed).

We are given :

$n_1 = 400$ ,  $X_1 =$  Number of men favouring the proposal = 200

$n_2 = 600$ ,  $X_2 =$  Number of women favouring the proposal = 325

$\therefore p_1 =$  Proportion of men favouring the proposal in the sample  $= \frac{X_1}{n_1} = \frac{200}{400} = 0.5$

$p_2 =$  Proportion of women favouring the proposal in the sample  $= \frac{X_2}{n_2} = \frac{325}{600} = 0.541$

Test Statistic. Since samples are large, the test statistic under the Null Hypothesis,  $H_0$  is:

$$Z = \frac{p_1 - p_2}{\sqrt{pQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1), \text{ where}$$

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{200 + 325}{400 + 600} = 0.525 \Rightarrow Q = 1 - P = 1 - 0.525 = 0.475$$

$$\therefore Z = \frac{0.500 - 0.541}{\sqrt{0.525 \times 0.475 \times \left(\frac{1}{400} + \frac{1}{600}\right)}} = \frac{-0.041}{\sqrt{0.001039}} = \frac{-0.041}{0.0323} = -1.269$$

### Conclusion

Since  $|Z| = 1.269$  which is less than 1.96, it is not significant at 5% level of significance. Hence  $H_0$  may be accepted at 5% level of significance and we may conclude that men and women do not differ significantly as regards proposal of flyover is concerned.

13. In a large city A, 20 per cent of a random sample of 900 school children had defective eye-sight. In other large city B, 15 per cent of random sample of 1,600 children had the same defect. Is this difference between the two proportions significant? Obtain 95% confidence limits for the difference in the population proportions.

*Sol :*

In usual notations :  $n_1 = 900$ ,  $p_1 = 20\% = 0.20$ ,  $n_2 = 1600$ ,  $p_2 = 15\% = 0.15$ .

Null hypothesis,  $H_0 : P_1 = P_2$

Alternative hypothesis,  $H_1 : P_1 \neq P_2$  (Two-tailed).

Test Statistic. Under  $H_0$ , the test statistic is :

$$Z = \frac{p_1 - p_2}{\text{S.E.}(p_1 - p_2)} \sim N(0,1), \text{ (since the samples are large.)}$$

where 
$$\hat{p} = \frac{900 \times 0.20 + 1600 \times 0.15}{900 + 1600} = 0.168 \Rightarrow \hat{Q} = 1 - \hat{p} = 0.832$$

$$\text{S.E.}(p_1 - p_2) = \sqrt{\left\{0.168 \times 0.832 \left( \frac{1}{900} + \frac{1}{1600} \right)\right\}} = \sqrt{0.0002427} = 0.0156.$$

$$\therefore Z = \frac{0.20 - 0.15}{0.0156} = 3.21$$

### Conclusion

Since the calculated value of Z is greater than 1.96, it is significant at 5% level. We, therefore, reject the null hypothesis  $H_0$  and conclude that the difference between the two proportions is significant.

The 95% confidence limits for the difference  $P_1 - P_2$  are :

$$(p_1 - p_2) \pm 1.96 \text{ S.E. of } (p_1 - p_2),$$

where 
$$\text{S.E. of } (p_1 - p_2) = \sqrt{\left( \frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2} \right)} = \sqrt{\left( \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} \right)}$$

$$= \sqrt{\left( \frac{0.20 \times 0.80}{900} + \frac{0.15 \times 0.85}{1600} \right)} = 0.016$$

Hence, the 95% confidence limits for  $P_1 - P_2$  are :

$$(0.20 - 0.15) \pm 1.96 (0.016) = 0.05 \pm 0.031 = (0.019 \text{ and } 0.081).$$

14. A company has the head office at Kolkata and a branch at Mumbai. The personnel director wanted to know if the workers at the two places would like the introduction of a new plan of work and a survey was conducted for this purpose. Out of a sample of 500 workers at Kolkata, 62% favoured the new plan. At Mumbai out of x, sample of 400 workers, 41% were against the new plan. Is there any significant difference between the two groups in their attitude towards the new plan at 5% level ?

*Sol:*

In the usual notations, we are given :

$$n_1 = 500, p_1 = 0.62 \text{ and } n_2 = 400, p_2 = 1 - 0.41 = 0.59$$

Null hypothesis,  $H_0: P_1 = P_2$ , i.e., there is no significant difference between the two groups in their attitude towards the new plan.

Alternative Hypothesis,  $H_1: P_1 \neq P_2$  (Two-tailed).

Test Statistic. Under  $H_0$ , the test statistic for large samples is :

$$Z = \frac{p_1 - p_2}{\text{S.E.}(p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{\hat{p}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \text{ where}$$

$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{500 \times 0.62 + 400 \times 0.59}{500 + 400} = 0.607 \Rightarrow \hat{Q} = 1 - \hat{p} = 0.393$$

$$\therefore Z = \frac{0.62 - 0.59}{\sqrt{0.607 \times 0.393 \times \left(\frac{1}{500} + \frac{1}{400}\right)}} = \frac{0.03}{\sqrt{0.00107}} = \frac{0.03}{0.0327} = 0.917$$

Critical region. At 5% level of significance, the critical value of Z for a two-tailed test is 1.96. Thus the critical region consists of all values of  $Z \geq 1.96$  or  $Z \leq -1.96$ .

### Conclusion

Since the calculated value of  $|Z| = 0.917$  is less than the critical value of Z (1.96), it is not significant at 5% level of significance. Hence the data do not provide us any evidence against the null hypothesis which may be accepted, and we may conclude that there is no significant difference between the two groups in their attitude towards the new plan.

### 3.9.3 Test for One Mean

#### Q21. Explain the hypothesis testing for single mean.

*Ans:*

if  $x_i$ , ( $i = 1, 2, \dots, n$ ) is a random sample of size  $n$  from a normal population with mean  $\mu$  and variance  $\sigma^2$ , then the sample mean is distributed normally with mean  $\mu$  and variance  $\sigma^2/n$ , i.e.,  $\bar{x} \sim N(\mu, \sigma^2/n)$ . However, this result holds, i.e.,  $\bar{x} \sim N(\mu, \sigma^2/n)$ , even in random sampling from non-normal population provided the sample size  $n$  is large [c.f. Central Limit Theorem]. Thus for large samples, the standard normal variate corresponding to  $\bar{x}$  is :

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Under the null hypothesis  $H_0$ , that the sample has been drawn from a population with mean  $\mu$  and variance  $\sigma^2$ , i.e., there is no significant difference between the sample mean ( $\bar{x}$ ) and population mean ( $\mu$ ), the test statistic (for large samples), is :

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

15. A sample of 900 members has a mean 3.4 cms. and s.d. 2.61 cms. Is the sample from a large population of mean 3.25 cms. and s.d. 2.61 cms.?

If the population is normal and its mean is unknown, find the 95% and 98% fiducial limits of true mean.

*Sol:*

Null Hypothesis, ( $H_0$ ): The sample has been drawn from the population with mean  $\mu = 3.25$  cms. and S.D.  $\sigma = 2.61$  cms.

Alternative Hypothesis,  $H_1$  :  $\mu \neq 3.25$  (Two-tailed).

Test Statistic. Under  $H_0$ , the test statistic is :  $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$ , (Since  $n$  is large.)

Here, we are given :  $\bar{x} = 3.4$  cms.,  $n = 900$  cms.,  $\mu = 3.25$  cms. and  $\sigma = 2.61$  cms.

$$\therefore Z = \frac{3.40 - 3.25}{2.61 / \sqrt{900}} = \frac{0.15 \times 30}{2.61} = 1.73$$

Since  $|Z| < 1.96$ , we conclude that the data don't provide us any evidence against the null hypothesis ( $H_0$ ) which may, therefore, be accepted at 5% level of significance.

95% fiducial limits for the population mean  $\mu$  are :

$$\bar{x} \pm 1.96(\sigma\sqrt{n}) = 3.40 \pm 1.96(2.61 / \sqrt{900}) = 3.40 \pm 0.1705, \text{ i.e., } 3.5705 \text{ and } 3.2295$$

98% fiducial limits for  $\mu$  are given by :

$$\bar{x} \pm 2.33 \frac{\sigma}{\sqrt{n}} = 3.40 \pm 2.33 \times \frac{2.61}{30} = 3.40 \pm 0.2027, \text{ i.e., } 3.6027 \text{ and } 3.1973$$

**Remark.**

2.33 is the value  $z_1$  of  $Z$  from standard normal probability integrals, such that

$$P(|Z| > z_1) = 0.98 \Rightarrow P(Z > z_1) = 0.49$$

16. The mean muscular endurance score of a random sample of 60 subjects was found to be 145 with a s.d. of 40. Construct a 95% confidence interval for the true mean. Assume the sample size to be large enough for normal approximation. What size of sample is required to estimate the mean within 5 of the true mean with a 95% confidence?

*Sol:*

In usual notations, we are given :  $n = 60$ ,  $\bar{x} = 145$  and  $s = 40$ .

95% confidence limits for true mean ( $\mu$ ) are :

$$\begin{aligned} & \bar{x} \pm 1.96 s / \sqrt{n} \quad (\sigma^2 = s^2, \text{ since sample is large}) \\ & = 145 \pm \frac{1.96 \times 40}{\sqrt{60}} = 145 \pm \frac{78.4}{7.75} = 145 \pm 10.12 = 134.88 \text{ and } 155.12 \end{aligned}$$

Hence 95% confidence interval for  $\mu$  is (134.88, 155.12). In the notations of Example 14.19, we have

$$n = \left( \frac{z_{\alpha} \cdot \sigma}{E} \right)^2 = \left( \frac{1.96 \times 40}{5} \right)^2 = (15.68)^2 = 245.86 \approx 246.$$

$$[\because z_{0.05} = 1.96, \hat{\sigma} = s = 40 \text{ and } |\bar{x} - \mu| < 5 = E]$$

### 3.9.4 Test for Two Means

**Q22. Explain the hypothesis testing for two means.**

*Ans :*

Let  $\bar{x}_1$  be the mean of a sample of size  $n_1$  from a population with mean  $\mu_1$  and variance  $\sigma_1^2$  and let  $\bar{x}_2$  be the mean of an independent random sample of size  $n_2$  from another population with mean  $\mu_2$  and variance  $\sigma_2^2$ . Then, since sample sizes are large,

$$\bar{x}_1 \sim N(\mu_1, \sigma_1^2/n_1) \text{ and } \bar{x}_2 \sim N(\mu_2, \sigma_2^2/n_2)$$

Also  $\bar{x}_1 - \bar{x}_2$ , being the difference of two independent normal variates is also a normal variate. The value of  $Z$  (S.N.V.) corresponding to  $\bar{x}_1 - \bar{x}_2$  is given by :

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{\text{S.E.}(\bar{x}_1 - \bar{x}_2)} \sim N(0, 1)$$

Under the null hypothesis,  $H_0 : \mu_1 = \mu_2$ , i.e., there is no significant difference between the sample means, we get

$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2 = 0$ ;  $V(\bar{x}_1 - \bar{x}_2) = V(\bar{x}_1) + V(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ , the covariance term vanishes, since the sample means  $\bar{x}_1$  and  $\bar{x}_2$  are independent.

Thus under  $H_0 : \mu_1 = \mu_2$ , the test statistic becomes (for large samples),

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\{1/n_1 + 1/n_2\}}} \sim N(0, 1)$$

**17. The means of two single large samples of 1,000 and 2,000 members are 67.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from the same population of standard deviation 2.5 inches? (Test at 5% level of significance.)**

*Sol :*

In usual notations, we are given :

$$n_1 = 1,000, n_2 = 2,000, \bar{x}_1 = 67.5 \text{ inches, } \bar{x}_2 = 68.0 \text{ inches.}$$

Null hypothesis,  $H_0 : \mu_1 = \mu_2$  and  $\sigma = 2.5$  inches, i.e., the samples have been drawn from the same population of standard deviation 2.5 inches.

Alternative Hypothesis,  $H_1 : \mu_1 \neq \mu_2$  (Two-tailed)

Test Statistic. Under  $H_0$ , the test statistic is :

$$Z = \frac{\bar{x} - \bar{x}_2}{\sqrt{\left\{ \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right\}}} \sim N(0,1) \quad (\text{since samples are large})$$

$$\text{Now } Z = \frac{67.5 - 68.0}{2.5 \times \sqrt{\left( \frac{1}{1000} + \frac{1}{2000} \right)}} = \frac{-0.5}{2.5 \times 0.0387} = -5.1.$$

### Conclusion

Since  $|Z| > 3$ , the value is highly significant and we reject the null hypothesis and conclude that samples are certainly not from the same population with standard deviation 2.5.

- 18. In a survey of buying habits, 400 women shoppers are chosen at random in super market 'A' located in a certain section of the city. Their average weekly food expenditure is Rs. 250 with a standard deviation of Rs.40. For 400 women shoppers chosen at random in super market 'B' in another section of the city, the average weekly food expenditure is Rs. 220 with a standard deviation of Rs. 55. Test at 1% level of significance whether the average weekly food expenditure of the two populations of shoppers are equal.**

*Sol :*

In the usual notations, we are given that

$$n_1 = 400, \quad \bar{x}_1 = \text{Rs. } 250, \quad s_1 = \text{Rs. } 40$$

$$n_2 = 400, \quad \bar{x}_2 = \text{Rs. } 220, \quad s_2 = \text{Rs. } 55$$

Null hypothesis,  $H_0 : \mu_1 = \mu_2$  i.e., the average weekly food expenditures of the two populations of shoppers are equal.

Alternative Hypothesis, if :  $\mu_1 \neq \mu_2$  (Two-tailed)

Test Statistic. Since samples are large, under  $H_0$ , the test statistic is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left( \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)}} \sim N(0, 1)$$

Since  $\sigma_1$  and  $\sigma_2$ , the population standard deviations are not known, we can take for large samples  $\sigma_1^2 = s_1^2$  and  $\sigma_2^2 = s_2^2$  and then Z is given by :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}} = \frac{250 - 220}{\sqrt{\frac{(40)^2}{400} + \frac{(55)^2}{400}}} = 8.82 \text{ (approx.)}$$

### Conclusion

Since  $|Z|$  is much greater than 2.58, the null hypothesis ( $\mu_1 = \mu_2$ ) is rejected at 1% level of significance and we conclude that the average weekly expenditures of two populations of shoppers in markets A and B differ significantly.

19. The average hourly wage of a sample of 150 workers in a plant 'A' was Rs. 2.56 with a standard deviation of Rs. 1.08. The average hourly wage of a sample of 200 workers in plant 'B' was Rs. 2.87 with a standard deviation of Rs. 1.28. Can an applicant safely assume that the hourly wages paid by plant 'B' are higher than those paid by plant 'A'?

*Sol:*

Let  $X_1$  and  $X_2$  denote the hourly wages (in Rs.) of workers in plant A and plant B respectively. Then, in usual notations we are given :

$$\left. \begin{aligned} n_1 &= 150, & \bar{x}_1 &= 2.56, & s_1 &= 1.80 = \hat{\sigma}_1 \\ n_2 &= 200, & \bar{x}_2 &= 2.87, & s_2 &= 1.28 = \hat{\sigma} \end{aligned} \right\} \text{(Since samples are large.)}$$

Null Hypothesis,  $H_0 : \mu_1 = \mu_2$ , i.e., there is no significant difference between the mean level of wages of workers in plant A and plant B.

Alternative Hypothesis, :  $H_1 : \mu_2 > \mu_1$  i.e.,  $\mu_1 = \mu_2$  (Left-tailed test)

Test Statistic. Under  $H_0$ , the test statistic (for large samples) is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim N(0, 1)$$

$$\therefore Z = \frac{2.56 - 2.87}{\sqrt{\left\{\frac{(1.08)^2}{150} + \frac{(1.28)^2}{200}\right\}}} = \frac{-0.31}{\sqrt{0.016}} = \frac{-0.31}{0.126} = -2.46$$

Critical region. For a one-tailed test, the critical value of Z at 5% level of significance is 1.645. The critical region for left-tailed test thus consists of all values of  $Z \leq -1.645$ .

### Conclusion

Since calculated value of Z (-2.46) is less than critical value (-1.645), it is significant at 5% level of significance. Hence the null hypothesis is rejected at 5% level of significance and we conclude that the average hourly wages paid by plant 'B' are certainly higher than those paid by plant 'A'.

### 3.9.5 Test for two S.D's

**Q23. Explain the hypothesis testing for the difference of Standard Deviation.**

*Ans :*

Let  $S_1$  and  $S_2$  be the standard deviations of two independent random samples of sizes  $n_1$  and  $n_2$  for two populations with standard deviations  $\sigma_1$  and  $\sigma_2$  respectively.

Null Hypothesis,  $H_0 : \sigma_1 = \sigma_2$  i.e., the sample standard deviations don't differ significantly.

$$\text{Test statistic, } Z = \frac{s_1 - s_2}{\text{S.E.}(s_1 - s_2)} \sim N(0, 1) \text{ (for large samples)}$$

In sampling from a large population.

Sampling distribution of s is,

$$\text{Var}(S) = \frac{\sigma^2}{2n}$$

$$\text{or } S.E(S) = \frac{\sigma}{\sqrt{2n}}$$

$$\therefore \text{Var}(s_1 - s_2) = \text{Var}(S_1) + \text{Var}(S_2) = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$$

As samples are independent, covariance term is discarded.

$$\therefore Z = \frac{s_1 - s_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}}$$

$\sigma_1^2$  and  $\sigma_2^2$  are usually unknown and for large samples, we use their estimates given by their corresponding sample variances.

Hence, the test statistic reduces to,

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} \sim N(0,1) \text{ (for large samples).}$$

20. Random samples drawn from two countries gave the following data relating to the heights of adult males :

	Country A	Country B
Mean height (in inches)	67.42	67.25
Standard deviation (in inches)	2.58	2.50
Number in samples	1,000	1,200

- (i) Is the difference between the means significant?  
 (ii) Is the difference between the standard deviations significant?

*Sol:*

In usual notations, we are given :

$$n_1 = 1,000, \quad \bar{x}_1 = 67.42 \text{ inches}, \quad s_1 = 2.58 \text{ inches}$$

$$n_2 = 1,200, \quad \bar{x}_2 = 67.25 \text{ inches}, \quad s_2 = 2.50 \text{ inches}$$

As in the previous examples (since sample sizes are large), we can take

$$\hat{\sigma}_1 = s_1 = 2.58, \quad \hat{\sigma}_2 = s_2 = 2.50$$



- (i)  $H_0 : \mu_1 = \mu_2$ , i.e., the sample means do not differ significantly.

$$H_1 : \mu_1 \neq \mu_2 \text{ (Two tailed).}$$

Under the Null hypothesis,  $H_0$ , the test statistic is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2 / n_1) + (s_2^2 / n_2)}} \sim N(0,1), \text{ since samples are large.}$$

$$\text{Now } Z = \frac{67.42 - 67.25}{\sqrt{\left\{ \frac{(2.58)^2}{1000} + \frac{(2.50)^2}{1200} \right\}}} = \frac{0.17}{\sqrt{\left\{ \frac{6.66}{1000} + \frac{6.25}{1200} \right\}}} = 1.56$$

### Conclusion 1

Since  $|Z| < 1.96$ , null hypothesis may be accepted at 5% level of significance and we may conclude that there is no significant difference between the sample means.

- (ii) Under  $H_0$  : that there is no significant difference between sample standard deviations,

$$Z = \frac{s_1 - s_2}{\text{S.E.}(s_1 - s_2)} \sim N(0, 1), \text{ since samples are large.}$$

$$\text{Now } \text{S.E.}(s_1 - s_2) = \sqrt{\left\{ \frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2} \right\}} = \sqrt{\left\{ \frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2} \right\}}$$

if  $\sigma_1$  and  $\sigma_2$  are not known and  $\hat{\sigma}_1 = s_1$ ,  $\hat{\sigma}_2 = s_2$ .

$$\therefore \text{S.E.}(s_1 - s_2) = \sqrt{\left\{ \frac{(2.58)^2}{2 \times 1000} + \frac{(2.50)^2}{2 \times 1200} \right\}} = 0.07746$$

$$\text{Hence } Z = \frac{2.58 - 2.50}{0.07746} = \frac{0.08}{0.07746} = 1.03$$

### Conclusion 2

Since  $|Z| < 1.96$ , the data don't provide us any evidence against the null hypothesis which may be accepted at 5% level of significance. Hence the sample standard deviations do not differ significantly.

## Short Question and Answers

### 1. Central limit theorem.

*Ans :*

The central limit theorem states that as sample size is increased, the sampling distribution of the mean (and for other Sample statistics as well) approaches the normal distribution in form, regardless of the form of the population distribution from which the sample was taken.

For practical purposes, the sampling distribution of the mean can be assumed to be approximately normally distributed, even for the most non-normal populations or processes, whenever the sample size is  $n \geq 30$ . For populations that are 'only somewhat non-normal, even a smaller sample size will suffice. But a sample size of at least 30 will take care of the most adverse population situation.

The real advantage of the central limit theorem is that sample data drawn from populations not normally distributed or from populations of unknown shape also can be analyzed by using the normal distribution, because the sample means are normally distributed for sample sizes of  $n \geq 30$ .

### 2. Merits of Sampling.

#### 1. Economical

Since only a few units of population are studied in sample survey method hence, it is economical. This method saves money, resources and labour etc.

#### 2. Time Saving

The process of investigation under sample survey method is time saving as only a limited number of items of population are studied.

#### 3. Identification of Error

Because only a limited number of items are covered in sample survey method hence errors can be easily identified.

#### 4. Detailed Information

Since numbers of items are less in sample

inquiry, therefore, it is possible to obtain more detailed information from them.

### 5. Flexible

In comparison to census method, sample survey method is more flexible as it can be changed depending upon situation.

### 3. What are the Properties of a Good Estimator?

*Ans :*

A good estimator, as common sense dictates, is close to the parameter being estimated. Its quality is to be evaluated in terms of the following properties.

#### (i) Unbiasedness

An estimator is said to be unbiased if its expected value is identical with the population parameter being estimated. That is if  $\hat{\theta}$  is an unbiased estimate of  $\theta$ , then we must have  $E(\hat{\theta}) = \theta$ . Many estimators are "Asymptotically unbiased" in the sense that the biases reduce to practically insignificant values zero when  $n$  becomes sufficiently large. The estimator  $S^2$  is an example.

It should be noted that bias in estimation is not necessarily undesirable. It may turn out to be an asset in some situations. For example, it may happen that an unbiased estimator is less desirable than a biased estimator if the former has a greater variability than the latter and, as a consequence, the expected value of the latter is closer than that of the former to the parameter being estimated.

#### (ii) Consistency

If an estimator, say  $\hat{\theta}$ , approaches the parameter  $\theta$  closer and closer as the sample size  $n$  increases,  $\hat{\theta}$  is said to be a consistent estimator of  $\theta$ . Stating somewhat more rigorously, the estimator  $\hat{\theta}$  is said to be a consistent estimator of  $\theta$  if, as  $n$  approaches infinity, the probability approaches 1 that  $\hat{\theta}$  will differ from the parameter  $\theta$  by not more than an arbitrary small constant.

The sample mean is an unbiased estimator of  $\mu$  no matter what form the population distribution assumes, while the sample median is an unbiased estimate of  $\mu$  only if the population distribution is symmetrical. The sample mean is better than the sample median as an estimate of  $\mu$  in terms of both unbiasedness and consistency.

In case of large samples consistency is a desirable property for an estimator to possess. However in small samples, consistency is of little importance unless the limit of probability defining consistency is reached even with a relatively small size of the sample.

### (iii) Efficiency

The concept of efficiency refers to the sampling variability of an estimator. If two competing estimators are both unbiased, the one with the smaller variance (for a given sample size) is said to be relatively more efficient. Stated in a somewhat different language, estimator  $\hat{\theta}_1$  is said to be more efficient than another estimator  $\hat{\theta}_2$  for  $\theta$  if the variance of the first is less than the variance of the second. The smaller the variance of the estimator, the more concentrated is the distribution of the estimator around the parameter being estimated and, therefore, the better this estimator is.

If the population is symmetrically distributed, then both the sample mean and the sample median are consistent and unbiased estimators of  $\mu$ . Yet the sample mean is better than the sample median as an estimator of  $\mu$ . This claim is made in terms of efficiency.

### 4. What is standard error?

Ans :

The standard statistic is known as its Standard Error, abbreviated as S.E. The standard errors of some of the well-known statistics, for large samples, are given below, where  $n$  is the sample size,  $\sigma^2$  the population variance, and  $P$  the population proportion, and  $Q = 1 - P$ ;  $n_1$  and  $n_2$  represent the sizes of two independent random samples respectively drawn from the given population ( $\hat{s}$ ).

S.No.	Statistics	Standard Error
1.	Sample mean: $\bar{x}$	$\sigma / \sqrt{n}$
2.	Observed sample proportion 'p'	$\sqrt{PQ / n}$
3.	Samples s.d. : $s$	$\sqrt{\sigma^2 / 2n}$
4.	Sample variance : $s^2$	$\sigma^2 \sqrt{2 / n}$
5.	Sample quartiles	$1.36263 \sigma / \sqrt{n}$
6.	Sample median	$1.25331 \sigma / \sqrt{n}$
7.	Sample correlation coefficient ( $r$ )	$(1 - \rho^2) / \sqrt{n}$ p being the population correlation coefficient
8.	Sample moment: $\mu_3$	$\sigma^3 \sqrt{96 / n}$

9.	Sample moment: $\mu_4$	$\sigma^4 \sqrt{96/n}$
10.	Sample coefficient of variation (v)	$\frac{v}{\sqrt{2n}} \sqrt{1 + \frac{2v^3}{10^4}} \approx \frac{v}{\sqrt{2n}}$
11.	Difference of two sample means : $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
12.	Difference of two sample s.d. s. $(s_1 - s_2)$	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
13.	Difference of two sample proportions: $(p_1 - p_2)$	$\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}$

### 5. Define Sampling.

*Ans :*

Study of entire population may not be possible to carry out and hence a part alone is selected from the given population. A portion of the population which is examined with a view to determining the population characteristics is called a sample, i.e., a sample is a subset of population and the number of objects in the sample is called the size of the sample. Size of the sample is denoted by n.

The process of selection of a sample is called sampling. It is quite often used in our day-to-day practical life.

#### Example :

- To assess the quality of a bag of rice, sugar wheat or any other commodity, we examine only a portion of it by taking a handful of it from the bag and then decide to purchase it or not. The portion selected from the bag is called a sample, while the whole quantity of rice, sugar or wheat in the bag is the population.
- To estimate the proportion of defective articles in a large consignment, only a portion (i.e., a few of them) is selected and examined. The portion selected is a sample.
- Car produced in India is the population and the Nano cars is the sample.

### 6. Principles of Sampling.

*Ans :*

#### 1. Law of Statistical Regularity

This law has its roots in the mathematical theory of probability, which states that a moderately large sample selected randomly from a universe is likely to represent the characteristics of population.

#### 2. Law of Inertia of Large Numbers

Law of Inertia of Large Numbers is the extension of law of Statistical Regularity. This law states that the large numbers are relatively more stable as compared to small numbers. Hence, other things remaining the same, as the sample size increases, the results or findings tend to be more reliable and accurate.

**3. Law of Persistence**

This law states that there are certain inherent characteristics with the universe and the same will remain persisting and be reflected in the sample in same manner, even if the size of sample and that of universe is either increased or decreased.

**4. Law of Optimism**

This law states that the size of sample should be fixed in such a manner that it produced the optimum results. This means maximum possible accuracy with minimum possible costs in terms of all kinds of resources to be used during the course of research.

**7. What are the characteristics of a good sample?**

*Ans :*

A good sample should possess the following characteristics :

**1. Representativeness**

A good sample must represent the population from which it is selected. If sample drawn is not a good representative of the universe, the conclusion drawn about the population will be misleading.

**2. Adequate Size**

There is no hard and fast rule regarding the size of sample but it should be adequate in size. The size of sample should be in proportion to population size and should be determined in accordance with the purpose of statistical inquiry.

**3. Unbiased**

The sample should be unbiased. It should be free from the personal bias of the investigator.

**4. Homogeneity**

There should be homogeneity in the nature of units of population and that of the sample. It increases the accuracy of results.

**5. Independence**

The selection of units in the sample must be independent of one another. It means the

selection of one unit should not affect the selection of another unit in the sample.

**6. Random Selection**

It means all the units of population have an equal chance of being selected in the sample.

**8. Stratified Random Sampling**

*Ans :*

Stratified sampling is based on the concept of homogeneity and heterogeneity. The process of dividing heterogeneous population into relatively homogeneous strata is termed as stratified sampling. It is a two-step process in which population is divided into subgroups or strata. The strata should be mutually exclusive and collectively exhaustive. The element within a stratum should be as homogeneous as possible, but the element in different stratum should be as heterogeneous as possible. It is pertinent to mention here that though the sampling is selected in various stages yet the last sample of the subject is studied.

When the population has different sectors with different characteristics i.e. the population is divided on heterogeneous basis, we cannot get representative sample by the random sampling technique. In this case we use stratified random sampling. In the first step, the population is divided into 'strata' or groups on some homogeneous basis, and in the second step the selection of appropriate number of items is made from each sub-groups on random basis. The sum total of all the items taken separately from each sub-group or strata will form a stratified sample. Stratified sampling is much effectively used in market research where the division of the universe is fairly clear on the basis of occupational, economic, social or religious basis.

**9. Systematic Sampling.**

*Ans :*

Systematic sampling involves the selection of sample units at equal intervals, after all the units in the population are arranged in some systematic order such as alphabetical, chronological, geographical order etc. systematic sampling is also called 'quasi-random sampling'. In systematic sampling, the sample size is usually taken in such a way that it completely divides the population size.

Let us suppose that  $N$  sampling units in the population are arranged in some systematic order and serially numbered 1 to  $N$ . Our sample size ' $n$ '

should be such that it completely divides  $N$ .  $\frac{N}{n} = K$

this  $K$  is called that sample interval. If  $K$  is in fraction then it is to be rounded off to get an integral value e.g., if we want to have a sample of size 5 from a

population of size 100. Then  $K$  will be  $K = \frac{100}{5} = 20$ .

In this case the subsequent items are selected by taking every  $K^{\text{th}}$  items i.e.  $20^{\text{th}}$  items, refers to sample interval or sample ratio. The ratio of population size to the size of the sample.

### 10. Cluster Sampling.

*Ans :*

This sampling implies dividing population into clusters and drawing random sample either from all clusters or selected clusters. Cluster Sampling is similar to stratified sampling. In the cluster sampling the universe is divided into number of relatively small subdivisions or clusters and then some of these clusters are randomly selected for inclusion in the overall sample. The element within cluster should be heterogeneous as possible, but cluster themselves should be as homogeneous as possible. The common form of cluster sampling is area or geographical sampling.

**For example,** if we are interested in obtaining the income or opinion data in a city, the whole city may be divided into  $N$  different blocks or localities (which determine the clusters) and a simple random sample of  $n$  blocks is drawn. The individuals in the selected blocks determine the cluster sample. The difference between cluster sampling and stratified sampling is that in case of cluster sampling only a sample of subgroups or clusters is chosen, whereas in stratified sampling all subpopulations or strata are selected for further sampling. The objectives of the both methods are also different. The objective of cluster sampling is to increase the efficiency by decreasing the costs, whereas the objective of stratified sampling is to increase the precision.

### 11. Non-Probability Sampling Methods.

*Ans :*

Non probability sample methods are those which do not provide every item in the universe with a known chance of being included in the sample. The selection process is atleast partially subjective. Non-probability sampling methods are discussed as follow :

#### (A) Judgment Sampling

In judgment sampling, selection of sample units depends on the discretion or judgment of investigator. The investigator chooses the units from the universe according to his own judgment and includes all those items in the sample which he thinks best and typical of the universe.

#### (B) Quota Sampling

This method is suitable in making investigations concerning public opinion. Investigator define quotas according to some specific features of population like social classes, age groups and so on. The quotas confirm the total number of items in the sample taken as a whole. The selected sample units with the quotas depend on the personal judgments of the interviewer. For instance, in the market survey of tooth paste, the researcher may be asked to interview 100 people in certain areas of Shimla, and that out of 100 persons 40 are male, 40 are female and 20 are children under the age group of 15. With these quotas researcher is free to select the respondents to be interviewed.

### 12. Type I error.

*Ans :*

In a statistical hypothesis testing experiment, a Type I error is committed by rejecting the null hypothesis when it is true. The probability of committing a Type I error is denoted by a  $\alpha$  (pronounced as alpha), where

$$\alpha = \text{Prob. (Type I error)}$$

$$= \text{Prob. (Rejecting } H_0/H_a \text{ is true)}$$

## *Exercise Problems*

1. A random sample of 16 values from a normal population showed a mean of 48.5 and sum of squares of deviations from the mean equal to 135. Can it be assumed that the mean of the population is 43-5? (Use 5 per cent level of significance.)

**[Ans:  $t = 2.67$ ]**

2. A random sample of 12 pairs of observations from a normal population gives a coefficient of correlation of 0.54. Is this value significant of correlation in the population?

**[Ans:  $t = 1.597$ ]**

3. In an examination in Psychology, 12 students in one class had a mean grade of 78 with a standard deviation of 6, while 15 students in another class had a mean grade of 74 with a standard deviation of 8. Is there a significant difference between the means of the two groups?

**[Ans:  $t = 1.44$ ]**

4. In a random sample of 600 males in Jaipur, 400 were found to be smokers while in another random sample of 900 females in Delhi, 450 were found to be smokers. Discuss the question whether the data reveal a significant difference in Jaipur to Delhi so far as the proportion of smokers is concerned.

**[Ans: Diff./S.E. = 6.42]**

### Choose the Correct Answers

1. A hypothesis is true, but is rejected, this is an error of type [ a ]  
(a) I (b) II  
(c) I and II (d) None
2. A hypothesis is false, but accepted, this is an error of type [ b ]  
(a) I (b) II  
(c) I and II (d) None
3. A random sample of 400 products contains 52 defective items. Standard error of proportion is [ b ]  
(a) 0.168 (b) 0.0168  
(c) 0.0016 (d) 1.68
4. 500 eggs are taken from a large consignment and 50 are found to be bad. Standard error of proportion is [ c ]  
(a) 1.3 (b) 0.13  
(c) 0.013 (d) None
5. Among 900 people in a state 90 are found to be chapati eaters. The 99% confidence interval for the true proportion is [ c ]  
(a) (0.08,0.12) (b) (0.8,1.2)  
(c) (0.07,0.13) (d) None
6. If  $n = 40, Z = 2.06$  then the maximum error with 99% confidence is [ b ]  
(a) 0.7377 (b) 0.8387  
(c) 0.6387 (d) 0.536
7. If  $n = 100, Z = 5$  then the maximum error with 95% confidence is [ a ]  
(a) 0.98 (b) 1.2875  
(c) 3.92 (d) 1.16



8. If the size of the sample is 25 and the maximum error with 95% confidence is 0.1 then the standard deviation of the sample is [ c ]
- (a) 2.55 (b) 2.12
- (c) 0.255 (d) 0.025
9. If the maximum error with probability 0.95 is 1.2 and the standard deviation of population is 10, then sample size is [ c ]
- (a) 26 (b) 266
- (c) 267 (d) 269
10. A sample of size  $n$  is taken from a population whose variance is 9. The maximum error of estimate for  $p$  with 95% confidence is 0.5. Then  $n =$  [ d ]
- (a) 12 (b) 68
- (c) 128 (d) 139

### *Fill in the blanks*

1. \_\_\_\_\_ is the process of gathering information about the entire population, by taking samples.
2. The \_\_\_\_\_ method involves conducting the investigation of the entire population.
3. \_\_\_\_\_ states that other things being equal, with increase in sample size the results tends to be more reliable and accurate.
4. \_\_\_\_\_ explains the relationship between shape of the population distribution and the sampling distribution of the mean.
5. \_\_\_\_\_ error means acceptance of hypothesis which should have been rejected.
6. The standard deviation of a sampling distribution of a statistic is known as its \_\_\_\_\_.
7. \_\_\_\_\_ is a sample which is selected on the basis of laws of chance or probability.
8. Any hypothesis which is complementary to the null hypothesis is called an \_\_\_\_\_.
9. \_\_\_\_\_ is the procedure of using a sample statistic to estimate a population parameter.
10. \_\_\_\_\_ is the probability of rejecting  $H_0$  given that a specific alternative is true.
11. The null hypothesis asserts that there is no true difference in the \_\_\_\_\_ and the \_\_\_\_\_ in the particular matter under consideration.
12. Type I error is committed when the hypothesis is true but our test \_\_\_\_\_ it.
13. Type II errors are made when we accept a null hypothesis which is \_\_\_\_\_.
14. In \_\_\_\_\_ tail test rejection region is located in one tail.
15. The standard deviation of sampling distribution is called \_\_\_\_\_.
16. The distribution formed of all possible values of a statistics is called the \_\_\_\_\_.
17. The mean of sampling distribution of means is equal to the \_\_\_\_\_.
18. Standard error provides an idea about the \_\_\_\_\_ of sample.
19. An estimator is said to be \_\_\_\_\_ if it covers as much information as possible about the parameter which is contained in the sample.
20. An \_\_\_\_\_ estimate of a population parameter provides two values, between which it is estimated that the parameter lies.

### ANSWERS

1. Sampling
2. Census
3. Principle of inertia of large numbers
4. Central limit theorem

5. Type II
6. Standard error
7. Random sample
8. Alternative hypothesis
9. Statistical estimation
10. Power of a test.
11. Sample, population
12. Rejects
13. Not true
14. One
15. Standard error
16. Sampling distribution
17. Population mean
18. Unreliability
19. Sufficient
20. Interval

Rahul Publications

## UNIT IV

- (i) **Small Sample Tests- t-Distribution:** properties and applications, testing for one and two means, paired t-test.
- (ii) **Analysis of Variance:** One Way and Two ANOVA (with and without Interaction).
- (iii) **Chi-square distribution:** Test for a specified Population variance, Test for Goodness of fit, Test for Independence of Attributes.

### 4.1 SMALL SAMPLE TESTS

**Q1. What is Small sample test ?**

*Ans :*

#### Meaning

Small sample size referred to size of sample which is less than 30. In case of small sample size the z-test is not appropriate test statistic as the assumptions on which it is based do not hold good in case of small sample. The theoretical work on t-distribution was done by W.S. Gosset (1876-1937) under the pen name "student" as he was the employee of the company Gunnies & Sons, a Dublin bravery, Ireland, which did not allowed it employees to publish research findings under their own names. The t-distribution is used when sample size is less than 30 and the population standard deviation is not known.

#### 4.1.1 t-Distribution

**Q2. Explain briefly about t-test. State the assumptions of t-test.**

*Ans :*

When the size of the sample is small i.e., less than 30, the Z-tests using normal distribution are not applicable because the assumptions on which they are based generally do not hold good in case of small samples. The sampling distribution of small samples follow student's t-distribution. The student's t-distribution has a greater dispersion than the standard normal distribution. As 'n' gets larger, the t-distribution approaches the normal form.

#### Degree of Freedom

Degree of freedom is used to see the table value for testing the hypothesis as  $V = n - 1$ . If hypothesis is to be tested 5% level of significance under one tail, then the value is to be seen below the 0.025 level. If the value of table is two tails, then the value is to be seen below the 0.05.

#### Assumptions

The following are the pre-requisites for the application of t-test :

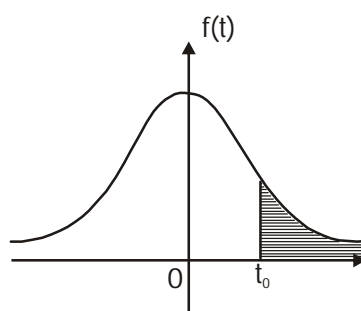
1. The population from which a sample is drawn is normal.
2. The samples have been drawn at random.
3. The population standard deviation is not known.
4. Sample size should small i.e., less than 30.

### 4.1.2 Properties

**Q3. What are the properties of t-distribution ?**

*Ans :*

1. The shape of t-distribution is bell-shaped, which is similar to that of a normal distribution and is symmetrical about the mean.
2. The t-distribution curve is also asymptotic to the t-axis, i.e., the two tails of the curve on both sides of  $t = 0$  extends to infinity.



**Fig.: t-distribution**

3. It is symmetrical about the line  $t = 0$ .
4. The form of the probability curve varies with degrees of freedom i.e., with sample size.
5. It is unimodal with Mean = Median = Mode.
6. The mean of standard normal distribution and as well as t-distribution zero but the variance of t-distribution depends upon the parameter  $v$  which is called the degrees of freedom.

### 4.1.3 Applications

**Q4. State the applications of the t-distribution.**

*Ans :*

The t - distribution has a wide number of applications in Statistics, some of them are given below :

1. To test the significance of the sample mean, when population variance is not given
2. To test the significance of the mean of the sample i.e., to test if the sample mean differs significantly from the population mean.
3. To test the significance of the difference between two sample means or to compare two samples.
4. To test the significance of an observed sample correlation coefficient and sample regression coefficient.

### 4.1.4 Testing for One and Two Means

**Q5. Explain the test concerning the significance of single and two mean**

*Ans :*

#### 1. T-test for Single Mean

- (i) If a random sample  $x$ . of size  $n$  has been drawn from a normal population with a specified mean  $p$ .
- (ii) If the sample mean differs significantly from the hypothetical value  $p$  the population mean.

In this case the statistic is given by  $t = \frac{\bar{x} - \mu}{S / \sqrt{n}} \sim t_{(n-1)}$  where  $\bar{x}$ ,  $\mu$ ,  $S$ ,  $n$  have usual meanings.

Let a random sample of size  $n$  ( $n < 30$ ) has a sample mean  $\bar{x}$ . To test the hypothesis that the population mean  $\mu$  has a specified value  $\mu_0$  when population S. D.  $\sigma$  is not known.

Let the Null Hypothesis be  $H_0: \mu = \mu_0$

Then the Alternative Hypothesis is  $H_1: \mu \neq \mu_0$

Assuming that  $H_0$  is true, the test statistic given by  $t = \frac{\bar{x} - \mu}{s / \sqrt{n-1}}$ , where  $s$  is the sample S. D. follows t-distribution with  $v = (n - 1)$  d.f.

We calculate the value of  $t$  and compare this value with the table value of  $t_{\alpha}$  level of significance. If the calculated value of  $t >$  the table value of  $t$ , we reject  $H_0$  at  $\alpha$  level. Otherwise we accept  $H_0$ .

In this case, 95% confidence limits for the population mean  $\mu$  are  $\bar{x} \pm t_{\alpha} \cdot \frac{s}{\sqrt{n-1}}$  where  $\alpha = 0.025$

for two-tailed test and  $s$  = sample S. D. and 99% confidence limits  $\mu$  are  $\bar{x} \pm t_{\alpha} \cdot \frac{s}{\sqrt{n-1}}$  where  $\alpha = 0.05$ .

## 2. Test for Concerning Two Means

For a two-tailed test at  $\alpha$  level of significance, value of  $\alpha / 2$  is taken for  $\alpha$ .

The more common situation involving tests on two means are those in which variances are unknown. If we assume that distributions are normal and that  $\sigma_1 = \sigma_2 = \sigma$ . The pooled t-test (often called the two-sample t-test) may be used. The test statistic is given by the following test procedure,

$$t = \frac{\bar{X} - \bar{Y}}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (\text{or}) \quad t = \frac{\bar{X} - \bar{Y}}{S^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

With  $n_1 + n_2 - 2$  degrees of freedom.

Where,

$$S^2 = \frac{\sum(X_i - \bar{X}) + \sum(Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

Where  $\bar{x}_1$ ,  $\bar{x}_2$  are the means of two samples of size  $n_1$  and  $n_2$ .

The critical region with this t-distribution can be obtained in a similar way.

**For example**, when A.H is  $\mu_1 - \mu_2 \neq \delta$ , the null hypothesis ( $H_0$ ) is not rejected when,

$$-t_{\frac{\alpha}{2}, n_1+n_2-2} < t < t_{\frac{\alpha}{2}, n_1+n_2-2} \quad \text{and the critical region is } t < -t_{\frac{\alpha}{2}, n_1+n_2-2} \quad (\text{or}) \quad t_{\frac{\alpha}{2}, n_1+n_2-2}$$

Critical region for testing  $H_0 : \mu_1 - \mu_2 = \delta$

Alternate hypothesis; Reject null hypothesis if ( ),

$$(i) \quad \mu_1 - \mu_2 \neq \delta \quad t < -t_{\alpha/2} \text{ (or) } t > t_{\alpha/2}$$

$$(ii) \quad \mu_1 - \mu_2 > \delta \quad t > t_{\alpha}$$

$$(iii) \quad \mu_1 - \mu_2 < \delta \quad t < -t_{\alpha}$$

**Note :**

1. The two-sample t-test can not be used if  $\sigma_1 \neq \sigma_2$ .
2. The two-sample t-test can not be used for 'before and after' kind of data, where the is naturally paired.

In other words the samples must be 'independent' for two sample t-test.

### PROBLEMS

1. **A mechanist making engine parts with axle diameters of 0.700 inch. A random sample of 10 parts shows a mean diameter of 0.742 inch with a S.D. of 0.040 inch. Compute the statistic you would use to test whether the work is meeting the specification at level of significance.**

*Sol :*

Here the sample size  $n = 10 < 30$

Hence the sample is small sample.

Also sample mean  $\bar{x} = 0.742$  inches , the population mean  $\mu = 0.700$  inches and S.D. = 0.040 inches are given.

$\therefore$  We use student's t-Test

(i) **Null Hypothesis  $H_0$  :** The product is confirming to specification.

(ii) **Alternative Hypothesis  $H_1$  :**  $\mu \neq 0.700$

(iii) **Level of significance is,  $\alpha = 0.05$**

(iv) **The test statistic is,  $t = \frac{\bar{x} - \mu}{s / \sqrt{n-1}}$**

Here  $\bar{x} = 0.742$  inches,  $\mu = 0.700$  inches, S.D. = 0.040 inches and  $n = 10$ . Degrees of freedom (d.f) =  $w-1 = 10-1 = 9$

$$\therefore t = \frac{0.742 - 0.700}{\frac{0.040}{\sqrt{10-1}}} = 3.15$$

$\therefore$  The calculated value of  $t = 3.15$

The tabulated value of  $t$  at 5% level with 9 degrees of freedom is  $t_{0.05} = 2.26$

Since calculated value of  $t >$  tabulated value of  $t$ , therefore,  $H_0$  is rejected.

$\therefore$  The product is not meeting the specification.

2. A sample 26 bulbs gives a mean life of 990 hours with a S.D. of 20 hours. The manufacturer claims that the mean life of bulbs is 1000 hours. Is the sample not upto the standard.

*Sol :*

Here sample size,  $n = 26 < 30$

$\therefore$  The sample is small sample.

Also given, sample mean,  $\bar{x} = 990$

Population mean,  $\mu = 1000$  and S.D.,  $s = 20$

Degrees of freedom =  $n - 1 = 26 - 1 = 25$

Here we know  $\bar{x}$ ,  $\mu$ , S.D. and  $n$ .

$\therefore$  We use students 't' test.

- (i) Null Hypothesis  $H_0$  : The sample is upto the standard.

- (ii) Alternative Hypothesis  $H_1$  :  $\mu < 1000$

(The sample is below standard) (left-tail test)

- (iii) Level of significance :  $\alpha = 0.05$

- (iv) The test statistic is  $t = \frac{\bar{x} - \mu}{s / \sqrt{n-1}} = \frac{990 - 1000}{20 / \sqrt{25}} = -2.5$

$\therefore |t| = 2.5$

i.e., Calculated value of  $t = 2.5$

Tabulated value of 't' at 5% level with 25 degrees of freedom for left-tailed test is 1.708.

Since calculated  $t >$  tabulated  $t$ , we reject the null hypothesis  $H_0$  and conclude: that the sample is not upto the standard.

3. A machine is designed to produce insulating washers for electrical device of average thickness of 0.025 cm. A random sample of 10 washers was found to have a thickness of 0.024 cm with a S.D of 0.002 cm. Test the significance of the deviation. Value of  $t$  for 9 degrees of freedom at 5% level is 2.262.

*Sol :*

Here the sample size is  $10 < 30$

$\therefore$  The sample is small

Also given Sample mean,  $\bar{x} = 0.024$  cm

Population mean,  $\mu = 0.025$  cm

S.D. = 0.002 cm

Degrees of freedom (d.f) =  $n - 1 = 10 - 1 = 9$



(i) **Null Hypothesis  $H_0$**  : The difference between  $\bar{x}$  and  $\mu$  is not significant.

(ii) **Alternative Hypothesis  $H_1$**  :  $\mu_1 \neq 0.025$

(iii) **Level of significance** :  $\alpha = 0.05$

(iv) **The test statistic is 't'** 
$$= \frac{\bar{x} - \mu}{s / \sqrt{n-1}}$$

$$= \frac{0.024 - 0.025}{\frac{0.002}{\sqrt{10-1}}}$$

$$= -1.5$$

$$\Rightarrow |t| = 1.5$$

$\therefore$  Calculated value of  $t = 1.5$  for two tailed test.

Tabulated value of  $t$  for 9 degrees of freedom at 5% level = 2.262

Since calculated  $t <$  tabulated  $t$ , we accept the null hypothesis and conclude that the difference between  $x$  and  $p$  is not significant.

4. **Two different types of drugs A and B were tried on certain patients for increasing weight, 5 persons were given drug A and 7 persons were given drug B. The increase in weight (in pounds) is given below**

Drug A	8	12	13	9	3		
Drug B	10	8	12	15	6	8	11

**Do the drugs differ significantly with regard to their effect in increasing weight ?**

*Sol:*

Let the weights (in kgs) of the patients treated with drugs A and B be denoted by suitable variances  $X$  and  $Y$  respectively.

We set up the null hypothesis,  $H_0 : \mu_x = \mu_y$  i.e., there is no significant difference between the drugs A and B with regard to their effect on increase in patients weight.

Alternative hypothesis,  $H_1 : \mu_x \neq \mu_y$

Under  $H_0$ , the appropriate test statistic is,

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Degree of freedom (d.f) =  $t_{n_1+n_2} - 2$

## Computation of Sample Means and Standard Deviations

X	(X - $\bar{X}$ ) (X - 9)	(X - $\bar{X}$ ) <sup>2</sup>	Y	(Y - $\bar{Y}$ ) (Y - 10)	(Y - $\bar{Y}$ ) <sup>2</sup>
8	-1	1	10	0	0
12	3	9	8	-2	4
13	4	16	12	2	4
9	0	0	15	5	25
3	-6	36	6	-4	16
			8	-2	4
			11	1	1
$\Sigma X = 45$	$\Sigma(X - \bar{X}) = 0$	$\Sigma(X - \bar{X})^2 = 62$	$\Sigma Y = 70$	0	54

Here,  $n_1 = 5$ ,  $\Sigma X = 45$ ,  $\Sigma(X - \bar{X})^2 = 62$

$$\bar{X} = \frac{\Sigma X}{n_1} = \frac{45}{5} = 9$$

$n_2 = 7$ ,  $\Sigma Y = 70$ ,  $\Sigma(Y - \bar{Y})^2 = 54$

$$\bar{Y} = \frac{\Sigma Y}{n_2} = \frac{70}{7} = 10$$

and  $S^2 = \frac{1}{n_1 + n_2 - 2} [\Sigma(X - \bar{X})^2 + \Sigma(Y - \bar{Y})^2]$

$$= \frac{1}{5 + 7 - 2} [62 + 54]$$

$$= \frac{1}{10} [116]$$

$$S^2 = \frac{116}{10} = 11.6$$

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{9 - 10}{\sqrt{11.6 \times \frac{12}{35}}}$$

$$= \frac{-1}{\sqrt{3.98}}$$

$$= \frac{-1}{1.99}$$

$$\text{Degree of freedom (df)} = t_{n_1+n_2-2}$$

$$= t_{5+7-2}$$

$$= t_{10}$$

Hence, tabulate value of t for 10 df at 5% level of significance for the two tailed test is 2.228. Thus, calculated  $t = -0.50$ , is less than tabulated value of t (i.e., 2.228).

Therefore, null hypothesis  $H_0$  is accepted at 5% level of significance and we may conclude that the drugs A and B do not differ significantly with regard to their effect on increase in patients weights.

5. A sample of sales in similar shops in two towns are taken for a new product with the following results:

Town	Mean Sales	Variance	Sample Size
A	5	5	5
B	7	3	7

Is there any evidence of difference in sales in the two towns? Use 5 per cent level of significance to test this difference between the means of two samples.

*Sol.:*

(Dec.-20)

Given that,

$$\bar{x}_1 = 5 \quad \bar{x}_2 = 7$$

$$\sigma S_1^2 = 5 \quad \sigma S_2^2 = 3$$

$$n_1 = 5 \quad n_2 = 7$$

**Null Hypothesis ( $H_0$ )** = There is no significant differences between in sales in two towns.

**Alternative Hypothesis ( $H_1$ )** = There is a significant differences between in sales in two towns.

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$S^2 = \left( \frac{n_1 \sigma_1^2 + n_2 \sigma_2^2}{n_1 + n_2 - 2} \right)$$

$$S^2 = \left( \frac{5 \times 5 + (7 \times 3)}{5 + 7 - 2} \right)$$

$$= \frac{25 + 21}{10}$$

$$= \frac{46}{10} = 4.6$$

$$t = \frac{5 - 7}{\sqrt{4.6 \left( \frac{1}{5} + \frac{1}{7} \right)}}$$

$$= \frac{-2}{\sqrt{1.564}} = -1.6$$

$$\text{Degree of freedom} = n_1 + n_2 - 2$$

$$= 5 + 7 - 2 = 10$$

Tabulated value of t for 10 d.f at 5% level of significance of for the two tailed test is  $2 < 2.228$ .

As calculated value  $-1.6$  is less than tabulated value  $(-2.2288)$  therefore  $H_0$  is accepted.

#### 4.1.5 Paired T-test

##### Q6. Discuss in detail about Paired t-test.

*Ans :* (July-18, Imp.)

Paired observations arise in many practical situations where each homogeneous experimental unit receives both population conditions. As a result, each experimental unit has a pair of observations, one for each population.

For instance, to test the effectiveness of "drug" some 11 persons blood pressure is measured "before" and "after" the intake of certain drug. Here the individual person is the experimental unit and the two populations are blood pressure "before" and "after" the drug is given. Thus for each observation in one sample, there is a corresponding observation in the other sample pertaining to the same character. Hence the two samples are not independent.

Consider another example. Suppose a business concern is interested to know whether a particular media of promoting sales of a product is really effective or not. In this case we have to test whether the average sales before and after the sales promotion are equal.

If  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , be the pairs of sales data before and after the sales production in a business concern, we apply paired t - test to examine the significance of the difference of the two situations.

Let  $d_i = x_i - y_i$  (or)  $y_i - x_i$  for  $i = 1, 2, 3, \dots, n$

Let the Null Hypothesis be  $H_0 : \mu_1 = \mu_2$  (i.e.,  $\mu = 0$ ), there is no significant difference between the means in two situations.

Then the Alternative Hypothesis is  $H_1 : \mu_1 \neq \mu_2$

Assuming that  $H_0$  is true, the test statistic for n paired observations (which are dependent) by taking the differences  $d_1, d_2, \dots, d_n$  of the paired data.

$$t = \frac{\bar{d} - \mu}{S / \sqrt{n}} = \frac{\bar{d}}{S / \sqrt{n}} (\because \mu = 0)$$

where  $t = \frac{\bar{d}}{S} = \frac{1}{n} \sum d_i$  and  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$

or  $S^2 = \frac{\sum d^2 - n(\bar{d})^2}{n-1}$  or  $\frac{1}{n-1} \left[ \sum d^2 - \frac{(\sum d)^2}{n} \right]$

are the mean and variance of the differences  $d_1, d_2, \dots, d_n$  respectively and  $p$  is the area of the population of differences.

The above statistic follows student's t-distribution with  $(n - 1)$  degrees of freedom.

### PROBLEMS

6. Ten workers were given a training programme with a view to study then assembly time for a certain mechanism. The results of the time and motion studies before and after the training programme are given below.

Workers	1	2	3	4	5	6	7	8	9	10
$X_1$	15	18	20	17	16	14	21	19	13	22
$Y_1$	14	16	21	10	15	18	17	16	14	20

$X_1$  = Time taken for assembling before training,

$Y_1$  = Time taken for assembling after training.

Test whether there is significant difference in assembly times before and after training

*Sol:*

From the given paired data, we see that we are to use paired t-test. Let  $\mu$  be the mean of population of differences.

- (i) **Null Hypothesis  $H_0$ :**  $\mu_1 = \mu_2$  or  $\mu = 0$  i.e., training is not useful.
- (ii) **Alternative Hypothesis  $H_1$ :**  $\mu_1 \neq \mu_2$  i.e., training is useful in assembly time.
- (iii) **Level of significance  $\alpha$**  = 0.05
- (iv) **Computation :** Differences  $d_i$ 's (before and after training) are  
1, 2, -1, 7, 1, -4, 4, 3, -1, 2

$$\bar{d} = \text{mean of differences of sample data} = \frac{\sum d}{n} = \frac{14}{10} = 1.4$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$$

$$= \frac{1}{9} [(1 - 1.4)^2 + (2 - 1.4)^2 + (-1 - 1.4)^2 + (7 - 1.4)^2 + (1 - 1.4)^2 + (-4 - 1.4)^2 + (4 - 1.4)^2 + (3 - 1.4)^2 + (-1 - 1.4)^2 + (2 - 1.4)^2]$$

$$= \frac{1}{9} [0.16 + 0.36 + 5.76 + 31.36 + 0.16 + 29.16 + 6.76 + 2.56 + 5.76 + 0.36]$$

$$= \frac{82.4}{9} = 9.1555$$

$$\therefore S = 3.026$$

(v) The test statistic is  $t = \frac{\bar{d} - \mu}{S/\sqrt{n}} = \frac{\bar{d}}{S/\sqrt{n}} = \frac{1.4}{3.026/\sqrt{10}} = \frac{(1.4)(3.163)}{3.026} = 1.46$

$$\therefore \text{Calculated } |t| = 1.46$$

Tabulated  $t_{0.05}$  with  $10 - 1 = 9$  degrees of freedom is 1.833

Since calculated  $t < t_{0.05}$ , we accept the Null Hypothesis  $H_0$  and conclude that there is no significant difference in assembly times before and after training.

7. Scores obtained in a shooting competition by 10 soldiers before and after intensive training are given below :

Before	67	24	57	55	63	54	56	68	33	43
After	70	38	58	58	56	67	68	75	42	38

Test whether the intensive training is useful at 0.05 level of significance.

*Sol:*

Let us apply paired test.

Let  $\mu$  be the mean of population of differences.

(i) **Null Hypothesis  $H_0$**  :  $\mu_1 = \mu_2$  i.e.,  $\mu = 0$  there is no significant effect of the training.

(ii) **Alternative Hypothesis  $H_1$**  :  $\mu_1 \neq \mu_2$  intensive training is useful.

(iii) **Level of significance**,  $\alpha = 0.05$

(iv) **Computation :**

Differences  $d_i$ 's (before and after training) are -3, -14, -1, -3, 7, -13, -12, -7, -9, 5

$$\bar{d} = \frac{1}{10} \sum_{i=1}^{10} d_i = \frac{-50}{10} = -5$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{10} (d_i - \bar{d})^2$$

$$= \frac{1}{9} [(2)^2 + (-9)^2 + (4)^2 + (2)^2 + (12)^2 + (-8)^2 + (-7)^2 + (-2)^2 + (-4)^2 + (10)^2]$$

$$= \frac{1}{9} [4 + 81 + 16 + 4 + 144 + 64 + 49 + 4 + 16 + 100]$$

$$= \frac{482}{9} = 53.5555$$

$$\therefore S = 7.32$$

(v) The test statistic is  $t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{-5-0}{7.32/\sqrt{10}} = 2.16$

Tabulated  $t_{0.05}$  with  $10 - 1 = 9$  degrees of freedom is 1.83.

Since calculated  $t > t$  tabulated  $t$ , we reject the Null Hypothesis and conclude that the intensive training is useful.

8. The Blood Pressure of 5 women before and after intake of a certain drug are given below :

Before	110	120	125	132	125
After	120	118	125	136	121

Test whether there is significant change in Blood Pressure at 1 % level of significance.

*Sol :*

Let the Null Hypothesis be  $H_0 : \mu_1 = \mu_2$  i. e., there is no significant difference in blood pressure before and after intake of a drug.

Then the Alternative Hypothesis is  $H_1 : \mu_1 < \mu_2$

Assuming that  $H_0$  is true, the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}} \text{ where } \bar{d} = \frac{\sum d}{n}, d = y - x$$

$$\text{and } S^2 = \frac{\sum (d - \bar{d})^2}{n-1} = \frac{\sum d^2 - (\bar{d})^2 \times n}{n-1}$$

#### Calculations For d and S

Womens	B.P. before intake of drug (x)	B.P. after intake of drug (y)	d = y - x	d <sup>2</sup>
1	110	120	10	100
2	120	118	- 2	4
3	123	125	2	4
4	132	136	4	16
5	125	121	- 4	16
<b>Total</b>			<b>Σd = 10</b>	<b>Σd<sup>2</sup> = 140</b>

$$\therefore \bar{d} = \frac{\sum d}{n} = \frac{10}{5} = 2 \text{ and } S^2 = \frac{140 - (2)^2 \times 5}{4} = \frac{140 - 20}{4} = \frac{120}{4} = 30$$

$$\therefore S = \sqrt{30}$$

$$\therefore t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{2}{\sqrt{30}/\sqrt{5}} = \frac{2}{\sqrt{6}} = 0.82$$

Degrees of freedom =  $n - 1 = 5 - 1 = 4$

Thus  $1 = 0.82 < 4.6$  at 1% level with 4 d.f. .

Since the calculated value of  $t <$  the tabulated value with 4 d.f. at 1% level, we accepted  $H_0$  at 1% level and conclude that there is no significant change in Blood Pressure aft intake of a certain drug.

## 4.2 ANALYSIS OF VARIANCE

**Q7. What is ANOVA? State the assumptions and applications of ANOVA.**

*Ans :*

**(Imp.)**

### Meaning

The variance test is also known as ANOVA. ANOVA is the acronym for Analysis of Variance. Analysis of variance is a statistical technique specially designed to test whether the means of more than two quantitative populations are equal i.e., to make inferences about whether those samples are drawn from the populations having the same mean.

The test is called 'F' test as it was developed by R.A Fisher in 1920's. The test is conducted in situations where we have three or more to consider, at a time an alternative procedure (to t-test) needed for testing the hypothesis that all samples could likely be drawn from the same population.

### Example

Five fertilizers are applied to four plots, each of wheat and yield of wheat on these plots is given. We are interested in finding out whether the effects of these fertilizers on the yields are significantly different or, in other words, whether the samples have come from the same population. ANOVA answers this question.

### Assumptions

Analysis of variance test is based on the test statistic F (or variance ratio).

It is based on the following assumptions,

- (i) Observations are independent.
- (ii) Each sample is drawn randomly from a normal population as the sample statistics reflect the characteristic of the population.
- (iii) Variance and means are identical for those population from which samples have been drawn.

### Applications

The applications of ANOVA are as follows,

1. Anova is used in education, industry, business, psychology fields mainly in their experiment design.
2. Anova helps to save time and money as several population means can be compared simultaneously.
3. Anova is used to test the linearity of the fitted regression line and correlation ratio, significance test statistic of anova

$$= F(r - 1, n - r).$$



### 4.2.1 One Way ANOVA

**Q8. Explain briefly about One Way ANOVA.**

*Ans :*

**(Imp.)**

In these classification the data is classified according to only one criteria i.e., It includes only one factor.

#### Steps involved in ANOVA One Way

- (i) Null Hypothesis  $H_0 = \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$  (where means are equal)
- (ii) Alternative Hypothesis  $H_1 = \mu_1 \neq \mu_2 \dots \neq \mu_k$  (when means are not equal)

Arithmetic mean and drawn from the means of population from which "K" samples are drawn which are equal to another.

#### Step 1

- (i) Calculation variance between samples
- (ii) Calculation of grand average  $\bar{\bar{X}}$

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{N}$$

- (iii) Take the difference between means of variance samples of grand average.
- (iv) Square the deviations and obtain total which will give some of squares between samples.
- (v) Divide the total by degrees of freedom  $K = \text{No. of samples}$   $V = K - 1$

#### Step 2

- (i) Calculate variance within samples.
- (ii) Calculation of mean value of samples  $\bar{X}_1, \bar{X}_2, \dots$
- (iii) Take the deviation of variance items in  $\alpha$  samples from mean values.
- (iv) Square the derivations and obtain total which give sum of squares within samples.
- (v) Divide the total by degrees of freedom

$$V = N - K$$

$N = \text{No. of observations}$

$K = \text{Refers to the no. of samples}$

#### Step 3

Calculation of "F" Ratio

$$"F" = \frac{S_1^2}{S_2^2}$$

#### Step 4

Compare the calculated value of "F" with 5% level of significance.

- If  $F_{cal} > F_{tab} \rightarrow$  Difference between sample means are significant.
- If  $F_{cal} < F_{tab} \rightarrow$  Difference between sample means are not significant.

## Step 5

Sources of Variations	SS (Sum of Squares)	V=Degress of Freedom	MS Mean Squares	"F" Ration
Between Samples	SSC	$V_1 = C - 1$	$MSC = \frac{SSC}{C - 1}$	$\frac{MSC}{MSE}$
Within Samples	SSE	$V_2 = M - C$	$MSE = \frac{SSE}{M - C}$	
Total	SST	$n - 1$		

SSC = Sum of Squares between samples (columns)

SSE = Sum of Squares within samples (rows)

SST = Total Sum & Squares of Variations

MSC = Mean sum of squares between samples

MSE = Mean sum of squares within samples.

**PROBLEMS**

9. Test whether the significance of possible variation in performance in a certain test between the grammar schools of a City A common test is given to number of students taken of random from the senior Vth Class of four schools.

Schools				
	A	B	C	D
1	8	12	18	13
2	10	11	12	9
3	12	9	16	12
4	8	14	6	16
5	7	4	8	15

*Sol:*

Given

No. of samples (k) = 4 samples (A, B, C, D)

**Hypothesis**

**Null Hypothesis :** There is no significant difference between schools

$$H_0 = \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

**Alternative Hypothesis :** There is significant difference between schools

$$H_1 = \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

**Step 1: Calculation of Variance between Samples**Calculation of  $\bar{X}_1, \bar{X}_2, \bar{X}_3, \bar{X}_4$ 

$X_1$	$X_2$	$X_3$	$X_4$
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15
$\bar{X}_1 = 9$	$\bar{X}_2 = 10$	$\bar{X}_3 = 12$	$\bar{X}_4 = 13$

Calculation of grand mean  $\bar{\bar{X}}$ 

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4}{N}$$

$$\bar{\bar{X}} = \frac{9 + 10 + 12 + 13}{4}$$

$$\boxed{\bar{\bar{X}} = 11}$$

Calculation of variance between samples

Sample (A) $(\bar{X}_1 - \bar{\bar{X}})^2$	Sample (B) $(\bar{X}_2 - \bar{\bar{X}})^2$	Sample (C) $(\bar{X}_3 - \bar{\bar{X}})^2$	Sample (D) $(\bar{X}_4 - \bar{\bar{X}})^2$
$(9 - 11)^2 = 4$	$(10 - 11)^2 = 1$	$(12 - 11)^2 = 1$	$(13 - 11)^2 = 4$
$(9 - 11)^2 = 4$	$(10 - 11)^2 = 1$	$(12 - 11)^2 = 1$	$(13 - 11)^2 = 4$
$(9 - 11)^2 = 4$	$(10 - 11)^2 = 1$	$(12 - 11)^2 = 1$	$(13 - 11)^2 = 4$
$(9 - 11)^2 = 4$	$(10 - 11)^2 = 1$	$(12 - 11)^2 = 1$	$(13 - 11)^2 = 4$
$(9 - 11)^2 = 4$	$(10 - 11)^2 = 1$	$(12 - 11)^2 = 1$	$(13 - 11)^2 = 4$
20	5	5	20

Sum of the squares of between samples

$$= 20 + 5 + 5 + 20$$

$$= 50$$

Mean sum of the squares of between samples

$$= \frac{50}{K-1}$$

$$= \frac{50}{4-1}$$

$$= \frac{50}{3}$$

$$= 16.6$$

**Step 2 : Calculation Variance with in Samples**

Sample (A)		Sample (B)		Sample (C)		Sample (D)	
$X_1$	$(X_1 - \bar{X}_1)^2$	$X_2$	$(X_2 - \bar{X}_2)^2$	$X_3$	$(X_3 - \bar{X}_3)^2$	$X_4$	$(X_4 - \bar{X}_4)^2$
8	1	12	4	18	36	13	0
10	1	11	1	12	0	9	16
12	9	9	1	16	16	12	1
8	1	14	16	6	36	16	9
7	4	4	36	8	16	15	4
	<u>16</u>		<u>58</u>		<u>104</u>		<u>30</u>

Sum of the square within samples

$$= 16 + 58 + 104 + 30$$

$$= 208$$

Mean sum of the squares of within samples

$$= \frac{208}{M-K}$$

M = Total no. of observations

K = No. of samples

$$= \frac{208}{20-4}$$

$$= \frac{208}{16}$$

$$= 13$$

**Step 3 : Calculation of "F" Ratio**

Sources of Variations	Sum of Squares	Degrees of Freedom	MS	"F" Ration
Between Samples	50	$V_1 = C - 1$ $V_1 = 4 - 1$ $V_1 = 3$	$MSC = \frac{SSC}{C - 1}$ 16.6	$F = \frac{MSC}{MSE}$ $F = 1.27$
Within Samples	208	$V_2 = M - C$ $V_2 = 20 - 4$ $V_2 = 16$	$MSE = \frac{SSC}{M - C}$ 13	
Total	258	19		

**Step 4 : Acceptance & Reject**

$$(V_1 = 3, V_2 = 16 \quad t_{0.05} = 3.24)$$

$$F_{cal} < F_{tab}$$

$$1.24 < 3.24$$

$\therefore$  Accept  $H_0$

$\therefore$  There is no significance difference between schools.

10. Four machines A, B, C and D are used to produce a certain kind of cotton fabrics. Samples of size 4 with each unit as 100 square metres are selected from the outputs of the machines at random and the number of flaws in each 100 square metres show the following result.

A	B	C	D
8	6	14	20
9	8	12	22
11	10	18	25
12	4	9	23

Do you think that there is a significant difference in the performance of the four machines?

*Sol :*

Let the null hypothesis be that there is no significant difference in the performance of the four machines i.e.,  $\mu_0 = \mu_1 = \mu_{11} = \mu_{111}$ .

A	B	C	D	Total
8	6	14	20	48
9	8	12	22	51
11	10	18	25	64
12	4	9	23	48
40	28	53	90	GT = 211

$$\text{Correction Factor (C.F.)} = \frac{(GT)^2}{N} = N = 16$$

$$\text{C.F.} = \frac{(211)^2}{16} = 2782.56$$

Total Sum of Squares (TSS)

$$\begin{aligned} &= \sum_i \sum_j X_{ij}^2 - \text{C.F.} \\ &= [(8)^2 + (6)^2 + (14)^2 + (20)^2 + (9)^2 + (8)^2 + (12)^2 + (22)^2 + (11)^2 + (10)^2 + (18)^2 \\ &\quad + (25)^2 + (12)^2 + (4)^2 + (9)^2 + (23)^2] - 2782.56 \\ &= [64 + 36 + 196 + 400 + 81 + 64 + 144 + 484 + 121 + 100 + 324 + 625 + 144 \\ &\quad + 16 + 81 + 529] - 2782.56 \\ &= 3409 - 2782.56 = 626.44 \end{aligned}$$

Sum of squares between the samples,

$$\begin{aligned} (\text{SSB}) &= \sum_j \frac{T_j^2}{n_j} - \frac{(GT)^2}{N} \\ &= \left[ \frac{(40)^2}{4} + \frac{(28)^2}{4} + \frac{(53)^2}{4} + \frac{(90)^2}{4} \right] - 2782.56 \\ &= [400 + 196 + 702.25 + 2025] - 2782.56 \\ &= 3323.25 - 2782.56 \\ &= 540.69 \end{aligned}$$

Sum of squares within samples,

$$\begin{aligned} \text{SSW} &= \text{TSS} - \text{SSB} \\ &= 626.44 - 540.69 \\ &= 85.75 \end{aligned}$$

**ANOVA Table**

Sources of Variation	Degree of Freedom	Sum of Squares	Mean Square	F-Ratio
Between Samples	$(k-1) = (4-1) = 3$	540.69	$\frac{540.69}{3} = 180.23$	$\frac{180.23}{7.14} = 25.22$
Within samples	$(n-k) = 16-4 = 12$	85.75	$\frac{85.75}{12} = 7.14$	
Total	$(n-1) = 16-1 = 15$			

$$\text{F-ratio}_{(3, 12)} = \text{calculated} = 25.22$$

$$\text{F-ratio from table } V_1 = 3 \text{ and } V_2 = 12 \text{ at 5\% level of significance} = 3.49$$

Since  $F_{(3, 12)} \text{ calculated} > F_{(3, 12)} \text{ table value}$  we reject  $H_0$  which means that there is a significant difference between the performance of four machines.

11. A manufacturing company wishes to test the average life of the four brands of electric bulbs. The company uses all brands in a randomly selected production plants. The records showing the lives (in "00" hours) of bulbs are as given in the table below:

Brand 1	Brand 2	Brand 3	Brand 4
22	21	23	17
25	17	21	19
20	19	22	18
19	22	19	20
	18	18	

Test the hypothesis that the average life for each brand of bulbs is the same. Assume alpha 1%.

*Sol.:*

(July-18, Imp.)

**Null Hypothesis:** There is no significant differences between average life of four brands of bulbs.

**Alternative Hypothesis:** There is a significant differences between average life of four brands of bulbs.

Brand 1		Brand 2		Brand 3		Brand 4	
$x_1$	$x_1^2$	$x_2$	$x_2^2$	$x_3$	$x_3^2$	$x_4$	$x_4^2$
22	484	21	441	23	529	17	289
25	625	17	289	21	441	19	361
20	400	19	361	22	484	18	324
19	361	22	484	19	361	20	400
		18	324	18	324	–	–
86	1870	97	1899	103	2139	74	1374

Calculation of Grand Total

$$86 + 97 + 103 + 75 = 360$$

Calculation of correction factor

$$= \frac{(GT)^2}{N}$$

$$= \frac{(360)^2}{18} = 7,200$$

Calculation of total sum of squares (TSS)

$$\begin{aligned}
 TSS &= \sum_i \sum_j \sum_{ij}^2 - CF \\
 &= (1870 + 1899 + 2139 + 1374) - 7200 \\
 &= 7,282 - 7,200 \\
 &= 82
 \end{aligned}$$

Calculation of sum of squares between samples (SSB)

$$\begin{aligned}
 SSB &= \frac{\sum T_j^2}{n_j} - CF \\
 &= \left( \frac{(86)^2}{4} + \frac{(97)^2}{5} + \frac{(103)^2}{5} + \frac{(74)^2}{5} \right) - 7,200 \\
 &= (1849 + 1881.8 + 2121.8 + 1369) - 7,200 \\
 &= 7221.6 - 7,200 = 21.6
 \end{aligned}$$

Calculation of sum of squares within samples

$$\begin{aligned}
 SSW &= TSS - SSB \\
 &= 82 - 21.6 = 60.4
 \end{aligned}$$

#### Anova

Sources of variation	Sum of squares	Degree of freedom	Mean squares	Frantic
Between samples	21.6	$(4-1) = 3$	$\frac{21.6}{3} = 7.2$	$\frac{7.2}{4.31} = 1.67$
Within Samples	$\frac{60.4}{82}$	$(18 - 4) = \frac{14}{17}$	$\frac{604}{3} = 4.31$	
Total	82	17		

$$F_{cal} = 1.67$$

Value of  $F_{v_1 = 3 \text{ and } v_2 = 14}$  at 1% level of significance 5.56 from

$$F_{cal} < F_{tab} \quad \text{accept } H_0.$$

i.e., there is a no significant differences in the average life of 4 brands of bulbs.

#### 4.2.2 Two way ANOVA

**Q9. Explain briefly about two way ANOVA with and without interaction?**

*Ans :*

Two way classification/two factor ANOVA is defined where two independent factors have an effect on the response variable of interest.



**Example :** Yield of crop affected by type of seed as well as type of fertilizer.

In Two way classification (or two factor ANOVA), two independent factors are considered and their effect on the response variable of interest is observed.

**Example :** Yield of crops is affected by type of seed as well as type of fertilizer.

### Steps in Conducting Two-Way ANOVA Test

The steps involved in conducting two-way ANOVA test using short-cut method are as follows,

#### Step 1:

Set null hypothesis and alternative hypothesis for the two factors,

Say A and B

$$H_{0A} : \mu_1 = \mu_2 \dots \mu_k$$

$$H_{0B} : \mu_1 = \mu_2 \dots \mu_k$$

Null hypothesis states that arithmetic mean of populations from which samples are taken are equal.

$$H_{1A} : \mu_1 \neq \mu_2 \dots \mu_k$$

$$H_{1B} : \mu_1 \neq \mu_2 \dots \mu_k$$

Alternative hypothesis states that arithmetic means of populations from which samples are taken are not equal.

#### Step 2 :

Calculate the sum of all the items of all.

$$GT = \sum \sum x_{ij} = \sum_i R_i = \sum_j C_j$$

Where,

R = Row

C = Column

#### Step 3 :

Calculate Correlation Factor (C.F)

$$C.F = \frac{(GT)^2}{N}$$

#### Step 4 :

Calculate total sum of squares (TSS)

$$TSS = RSS - C.F$$

$$\Rightarrow RSS = \sum_i \sum_j x_{ij}^2$$

Where

R = Raw sum of squares

**Step 5 :**

Calculate Sum of Squares between Rows (SSR)

$$SSR = \frac{R_1^2}{n} + \frac{R_2^2}{n} + \frac{R_3^2}{n} + \dots + \frac{R_k^2}{n} - C.F$$

Where

n = Number of samples

**Step 6 :**

Calculate Sum of Squares between Columns (SSC).

$$SSC = \frac{C_1^2}{n} + \frac{C_2^2}{n} + \frac{C_3^2}{n} + \dots + \frac{C_k^2}{n} - C.F$$

**Step 7 :**

Calculate Sum of Squares due to Error (SSE).

$$SSE = TSS - SSR - SSC$$

**ANOVA Table for Two-way Classification**

Source of Variation	Sum of Squares	d.f.	Mean Sum of Squares	Ratio of F
Between Samples	SSC	(C - 1)	$MSC = \frac{SSC}{(C - 1)}$	F = MSC / MSE
Between Rows	SSR	(r - 1)	$MSR = \frac{SSR}{(r - 1)}$	F = MSR / MSE
Residual or Error	SSE	(C-1) (r-1)	$MSE = \frac{SSE}{(r - 1)(C - 1)}$	
Total	SST	N - 1		

Where,

MSC = Mean sum of squares of columns

MSR = Mean sum of squares of rows

MSE = Mean sum of squares of errors

**Step 8 :**

Compare the calculated value of F ratio with table value of F for degrees of freedom for between and within samples.

If calculated value of  $F < \text{table value of } F$ , the difference is taken as insignificant and we accept null hypothesis. If calculated value of  $F > \text{table value of } F$ , the difference is taken as significant and we reject null hypothesis.

### Hypothesis Tests in Two-way ANOVA

- **Factor A Test** : Hypothesis is designed to determine whether there are any factor A main effects. Null Hypothesis true if and only if there are no differences in means due to different treatments (population) of factor A.
- **Factor B Test** : Hypothesis is test designed to detect factor B main effects. Null hypothesis is true if and only if there are no differences in means due to different treatments (populations) of factor B.
- **Test for AB Interactions** : Test for existence of interactions between levels of the two factors Null hypothesis is true if and only if there are no two way interactions between levels of factor A and levels of factor B, means factor effects are additive for two way ANOVA.

### PROBLEMS

12. Four technicians analyzed three samples each of the moisture content in the sample. The results are given below :

Samples	Technicians			
	A	B	C	D
X	9	12	10	11
Y	12	11	15	12
Z	9	10	12	14

Analyze the data and comment. Use 5% significance level.

*Sol :*

(Aug.-17, Imp.)

A two-way ANOVA technique will enable us to analyses the data and comment.

Here, the two factors are, Technicians (T) and Samples (S).

#### Step-1: Null Hypothesis

$H_{or} = \mu_A = \mu_B = \mu_C = \mu_D$  i.e., the mean moisture content is same according to all technicians.

$H_{os} = \mu_x = \mu_y = \mu_z$  i.e., the mean moisture content is same in all the samples.

#### Alternative Hypothesis

$H_{IT} = \text{Atleast two of } \mu_A, \mu_B, \mu_C \text{ and } \mu_D \text{ are different.}$

$H_{IS} = \text{Atleast two of } \mu_x, \mu_y, \mu_z \text{ are different.}$

Samples	Technicians					
	A	B	C	D	Row Totals (R)	R <sup>2</sup>
X	9	12	10	11	R <sub>1</sub> = 42	R <sub>1</sub> <sup>2</sup> = 1764
Y	12	11	15	12	R <sub>2</sub> = 50	R <sub>2</sub> <sup>2</sup> = 2500
Z	9	10	12	14	R <sub>3</sub> = 45	R <sub>3</sub> <sup>2</sup> = 2025
Column Totals (C)	C <sub>1</sub> = 30	C <sub>2</sub> = 33	C <sub>3</sub> = 37	C <sub>4</sub> = 37	G = 137	
C <sup>2</sup>	C <sub>1</sub> <sup>2</sup> = 900	C <sub>2</sub> <sup>2</sup> = 1089	C <sub>3</sub> <sup>2</sup> = 1369	C <sub>4</sub> <sup>2</sup> = 1369		

**Step-2: Calculation Sum of Items of all Samples**

$$GT = \sum \sum x_{ij} = \sum_i R_i = \sum_j C_j$$

$$GT = 30 + 33 + 37 + 37 = 137$$

or

$$GT = 42 + 50 + 45 = 137$$

**Step-3: Calculation of Correction Factor (C.F)**

Here, GT = 137, N = 12,

$$C.F = \frac{(GT)^2}{N} = \frac{(137)^2}{12} = \frac{18769}{12} = 1564.08$$

**Step-4: Calculation of Total Sum of Squares (TSS)**

$$TSS = \sum_i \sum_j X_{ij}^2 - C.F$$

$$TSS = (9^2 + 12^2 + 10^2 + 11^2 + 12^2 + 15^2 + 12^2 + 9^2 + 10^2 + 12^2 + 14^2) - 1564.08$$

$$TSS = (81 + 144 + 100 + 121 + 144 + 225 + 144 + 81 + 100 + 144 + 196) - 1564.08$$

$$TSS = 1601 - 1564.08$$

$$= 36.92$$

**Step-5: Calculation of Sum of Squares Between Rows (i.e., Between Samples)**

$$SSR = \frac{R_1^2}{n} + \frac{R_2^2}{n} + \frac{R_3^2}{n} + \dots + \frac{R_k^2}{n} - C.F$$

$$SSR = \left( \frac{(42)^2}{4} + \frac{(50)^2}{4} + \frac{(45)^2}{4} \right) - 1564.08$$

$$SSR = (441 + 625 + 506.25) - 1564.08$$

$$SSR = 1572.25 - 1564.08$$

$$SSR = 8.17$$

**Step-6: Calculation of Sum of Squares Between Columns (i.e.. Between Technicians)**

$$SSC = \frac{C_1^2}{n} + \frac{C_2^2}{n} + \frac{C_3^2}{n} + \dots + \frac{C_k^2}{n} - C.F$$

$$SSC = \left( \frac{30^2}{3} + \frac{33^2}{3} + \frac{37^2}{3} + \frac{37^2}{3} \right) - 1564.08$$

$$SSC = (300 + 363 + 456.33 + 456.33) - 1564.08$$

$$SSC = 1575.66 - 1564.08$$

$$= 11.58.$$

**Step-7: Calculation of Sum of Squares of Residual or Error (SSE)**

$$SSE = TSS - SSR - SSC$$

$$= 36.92 - 8.17 - 11.58$$

$$= 17.17$$

**ANOVA Table**

Sources of Variation	Sum of Squares	Degree of Freedom	Mean Sum of Squares	F-ratio	F-ratio (Table)
Between Technicians	11.58	$(c - 1) = (4 - 1) = 3$	$\frac{11.58}{3} = 3.86$	$\frac{3.86}{2.86} = 1.35$	4.76
Between Samples	8.17	$(r - k) = (3 - 1) = 2$	$\frac{8.17}{2} = 4.09$	$\frac{4.09}{2.86} = 1.43$	5.14
Residual or Error	17.17	$(c - 1)(r - 1) = 3 \times 2 = 6$	$\frac{17.17}{6} = 2.86$		
<b>Total</b>	<b>36.92</b>	<b><math>(n - 1) = 12 - 1 = 11</math></b>			

**Step-8 : Table values of F-ratio at 5% level of significance**

$$F_{(6, 3)} = 4.76$$

$$F_{(6, 2)} = 5.14$$

**(i) Calculated  $F_{(6, 3)} < \text{Table } F(6, 3)$** 

Hence null hypothesis is accepted i.e., there is no significant difference between technicians opinion about mean moisture content in the sample.

**(ii) Calculated  $F_{(6, 2)} < \text{Table } F_{(6, 2)}$** 

Hence, null hypothesis is accepted i.e., there is no significant difference between samples. Mean moisture content is same in all the samples.

13. Four different drugs have been developed for a certain disease. These drugs are used under three different environments. It is assumed that the environment might affect efficacy of drugs. The number of cases of recovery from the disease per 100 people who have taken the drugs is tabulated as follows :

Environment	Drug A1	Drug A2	Drug A3	Drug A4
I	19	8	23	8
II	10	9	12	6
III	11	10	13	16

Test whether the drugs differ in their efficacy to treat the disease, also whether there is any effect of environment on the efficacy of disease.

*Sol :*

#### Null Hypothesis

$H_0$  = There is no significant difference in the efficacy of drugs to treat the disease.

$H_0$  = There is no significant effect of environment on the efficacy of disease.

Environment	Drug				Total
	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>	
I	19	8	23	8	58
II	10	9	12	6	37
III	11	10	13	16	50
<b>Total</b>	<b>40</b>	<b>27</b>	<b>48</b>	<b>30</b>	<b>GT=145</b>

$$\text{Correction Factor, (CF)} = \frac{(\text{Grand Total})^2}{N}$$

$$= \frac{(145)^2}{12} = 1752.08$$

#### Total Sum of Squares (TSS)

$$(\text{TSS}) = \sum_i \sum_j X_{ij}^2 - \text{C.F}$$

$$= [(19)^2 + (8)^2 + (23)^2 + (8)^2 + (10)^2 + (9)^2 + (12)^2 + (6)^2 + (11)^2 + (10)^2 + (13)^2 + (16)^2] - \text{C.F}$$

$$= 2025 - 1752.08$$

$$\therefore \text{TSS} = 272.92$$

**Sum of Squares Between Drugs (Column)**

$$\begin{aligned}
 SSC &= \sum_j \frac{T_j^2}{n_j} - \frac{(GT)^2}{N} \\
 &= \left[ \frac{(40)^2}{3} + \frac{(27)^2}{3} + \frac{(48)^2}{3} + \frac{(30)^2}{3} \right] - 1752.08 \\
 &= (533.33 + 243 + 768 + 300) - 1752.08 \\
 &= 1844.33 - 1752.08
 \end{aligned}$$

$$\therefore SSC = 92.25$$

$$\begin{aligned}
 \text{Degree of freedom (r)} &= (C - 1) = (4 - 1) \\
 &= 3
 \end{aligned}$$

**Sum of Squares Between Environment (Rows)**

$$\begin{aligned}
 SSR &= \sum_i \frac{T_i^2}{n_i} - \frac{(GT)^2}{N} \\
 &= \left[ \frac{(58)^2}{4} + \frac{(37)^2}{4} + \frac{(50)^2}{4} \right] - C.F \\
 &= (841 + 342.25 + 625) - 1752.08 \\
 &= 1808.25 - 1752.08
 \end{aligned}$$

$$\therefore SSR = 56.17$$

Degree of freedom,

$$V_m = (r - 1) - (3 - 1) - 2$$

$$\begin{aligned}
 \text{Residual} &= \text{Total sum of squares} - (\text{Sum of squares between columns} \\
 &\quad + \text{Sum of squares between rows}) \\
 &= TSS - (SSC + SSR) = 272.92 - (92.25 + 56.17) \\
 &= 272.92 - 148.42 = 124.5
 \end{aligned}$$

**ANOVA TABLE**

Sources of variation	Sum of squares	Degrees of Freedom	Means squares	Variance Ratio (F)
Between Drugs	92.25	$(C-1) = (4-1)=3$	$\frac{92.25}{3} = 30.75$	$F_{(3,6)} = \frac{30.75}{20.75} = 1.48$
Between Environment	56.17	$(r-1) = (3-1)=2$	$\frac{56.17}{2} = 28.09$	$F_{(2,6)} = \frac{28.09}{20.75} = 1.354$
Residual or Error (e)	124.5	$(C-1)(r-1)=3 \times 2=6$	$\frac{124.5}{6} = 20.75$	
<b>Total</b>	<b>272.92</b>	<b><math>(12 - 1) = 11</math></b>		

[Note: As level of significance is not given in the problem, assume 5% level of significance]

Critical value of $F_{0.05}$	Computed value of F
Drugs at $V_0(3,6) = 4.76$	1.48
Environment at $V_m(2,6) = 5.14$	1.354

Table values are calculated as per 5% level of significance.

### Decision

#### 1. Drugs

Since the calculated value of F(1.48) is less than the table value (4.76), null hypothesis is accepted. Hence, there is no significant difference in the efficacy drugs.

#### 2. Environment

Since the calculated value of F(1.354) is less than the table Value (5.14), null hypothesis is accepted. Hence, there is no affect of environment on the efficacy of disease.

14. Suppose that we are interested in establishing the yield producing ability of four types of soya beans A, B, C and D. We have three blocks of land X,Y and Z which may be different in fertility. Each block of land is divided into four plots and the different types of soya beans are assigned to the plots in each block by a random procedure. The following results are obtained:

Soya Bean

Block	Type A	Type B	Type C	Type D
X	5	9	11	10
Y	4	7	8	10
Z	3	5	8	9

Test whether A,B,C and D are significantly different.

Sol:

### Null Hypothesis

$H_0$  : There is no significant difference between A,B,C and D.

Soya bean

Block	Type A	Type B	Type C	Type D	Total
X	5	9	11	10	35
Y	4	7	8	10	29
Z	3	5	8	9	25
Total	12	21	27	29	GT = 89

$$\text{Correction Factor (CF)} = \frac{(\text{Grand Total})^2}{N} = \frac{(89)^2}{12} = 660.08$$



**Total Sum of Squares (TSS)**

$$\begin{aligned}
 &= \sum_i \sum_j x_{ij}^2 - \frac{(GT)^2}{N} \\
 &= [(5)^2 + (9)^2 + (11)^2 + (10)^2 + (4)^2 + (7)^2 + (8)^2 + (10)^2 + (3)^2 + (5)^2 + (8)^2 \\
 &\quad + (9)^2] - 660.08 \\
 &= [25 + 81 + 121 + 100 + 16 + 49 + 64 + 100 + 9 + 25 + 64 + 81] - 660.8 \\
 &= 735 - 660.08
 \end{aligned}$$

$$\therefore \text{TSS} = 74.92$$

**Sum of Squares Between Soya Bean (Columns)**

$$\begin{aligned}
 \text{SSB} &= \sum_j \frac{T_j^2}{n_j} - \frac{(GT)^2}{N} \\
 &= \frac{(12)^2}{3} + \frac{(21)^2}{3} + \frac{(27)^2}{3} + \frac{(29)^2}{3} - 660.08 \\
 &= [48 + 147 + 243 + 280.33] - 660.08 \\
 &= 718.33 - 660.08
 \end{aligned}$$

$$\therefore \text{SSB} = 58.25$$

$$\begin{aligned}
 \text{Degree of freedom (r)} &= (K - 1) \\
 &= (4 - 1) \\
 &= 3
 \end{aligned}$$

$$\text{Mean sum of squares between the soya beans} = \frac{58.25}{3} = 19.42$$

**Sum of Squares within Blocks (SSW)**

$$\begin{aligned}
 \text{SSW} &= \text{TSS} - \text{SSB} \\
 &= 74.92 - 58.25 \\
 &= 16.67
 \end{aligned}$$

Mean sum of squares within the blocks

$$= \frac{16.67}{12 - 4} = \frac{16.67}{8} = 2.08$$

**ANOVA TABLE**

Sources of variation	Sum of squares	Degrees of Freedom	Means squares
Between soya bean type	58.25	$(k - 1) = (4 - 1) = 3$	19.42
Within blocks	16.67	$(n - k) = (12 - 4) = 8$	2.08
Total		$(n - 1) = (12 - 1) = 11$	

$$\begin{aligned} \text{F-Ratio} &= \frac{\text{Mean square between soya bean type}}{\text{Mean square within blocks}} \\ &= \frac{19.42}{2.08} = 9.34 \end{aligned}$$

**[Note:** Assuming level of significance as 5%]

F-Ratio<sub>(3, 8)</sub>, calculated = 9.34

F-Ratio from table for  $V_1 = 3$  and  $V_2 = 8$  at 5% level of significance = 4.07

### Decision

The calculated value of F is more than the table value. Therefore we reject null hypothesis ( $H_0$ ) which means that there is a significant difference between types of soya beans.

## 4.3 CHI-SQUARE DISTRIBUTION

### Q10. Explain briefly about Chi-Square Test.

*Ans :*

The magnitude of discrepancy between the theory and observation is given by the quantity  $\chi^2$  (a Greek letter, pronounced as "chi-square"). If  $\chi = 0$ , the observed and expected frequencies completely coincide. As the value of  $\chi^2$  increases, the discrepancy between the observed and theoretical frequencies increases. Thus,  $\chi^2$  affords a measure of the correspondence between theory and observation.

### Definition

If a set of events  $A_1, A_2, \dots, A_n$  are observed to occur with frequencies  $O_1, O_2, \dots, O_n$  respectively and according to probability rules  $A_1, A_2, \dots, A_n$  are expected to occur with frequencies  $E_1, E_2, \dots, E_n$  respectively with  $O_1, O_2, \dots, O_n$  are called observed frequencies and  $E_1, E_2, \dots, E_n$  are called expected frequencies.

If  $O_i$  ( $i = 1, 2, \dots, n$ ) is a set of observed (experimental) frequencies and  $E_i$  ( $i = 1, 2, \dots, n$ ) is the corresponding set of expected (theoretical) frequencies, then  $\chi^2$  is defined as  $\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$  with  $(n - 1)$  degrees of freedom.

$\chi^2$  is used to test whether differences between observed and expected frequencies are significant.

### Note :

If the data is given in a series of 'n' numbers then degrees of freedom =  $n - 1$ .

In case of Binomial distribution, d.f. =  $n - 1$ .

In case of Poisson distribution, d.f. =  $n - 2$

In case of Normal distribution, d.f. =  $n - 3$

Chi-square Distribution is an important continuous probability distribution and it is used in both large and small tests. In chi-square tests,  $\chi^2$  - distribution is mainly used

- (i) To test the goodness of fit,
- (ii) To test the independence of attributes,
- (iii) To test if the population has a specified value of the variance  $\sigma^2$ .

### 4.3.1 Test for a Specified Population Variance

**Q11. Explain briefly about Test for a Specified Population Variance.**

*Ans :*

We want to test if the given normal population has specified variance  $\sigma^2 = \sigma_0^2$  (say).

We set up Null Hypothesis ( $H_0$ ):  $\sigma^2 = \sigma_0^2$ .

Alternative Hypothesis ( $H_1$ ):  $\sigma^2 \neq \sigma_0^2$ .

#### Test Statistic

$$\chi^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma_0^2} = \frac{ns^2}{\sigma_0^2} \sim \text{a Chi Square Distribution with } (n-1) \text{ D.F. The value of } s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

is the sample variance.

#### Conclusion

By comparing the calculated value of Chi-Square with the tabulated value of Chi-Square for (n-1) D.F. at the selected Level of Significance, we may accept or reject the  $H_0$ .

#### PROBLEMS

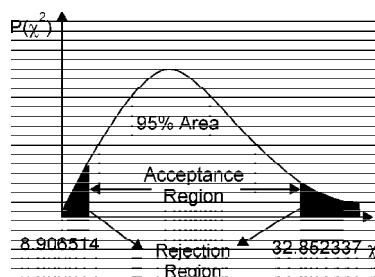
- 15. A manufacturer's representative claims that his company has perfected an adjustment to a machine that will reduce the variability in the diameter of the screws produced by the machine. The original variance was 0.010 inches. To determine whether or not his claim is reasonable, a random sample of 20 screws is selected from those produced after the adjustment has been made. The sample variance is 0.009. Does this evidence support the representative's claim ?**

*Sol. :*

1. Null Hypothesis ( $H_0$ ) :  $s^2 = 0.010$   
Alternative Hypothesis ( $H_1$ ) :  $s^2 \neq 0.010$
2. The data given is:  $n = 20$ ,  $s^2 = 0.009$   $\sigma_0^2 = 0.010$
3. Test Statistic

$$\chi^2 = \frac{ns^2}{\sigma_0^2} = \frac{20 \times 0.009}{0.010} = 18$$

Degrees of freedom =  $n - 1 = (20 - 1) = 19$



**Table Value**

At 5% LOS (selected), the table of  $c^2$  for 19 D.F. in two tailed test is between (8.906514, 32.852377) (found from the tables in Appendix).

**Conclusion**

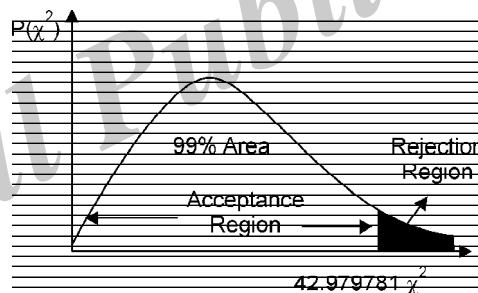
Since the computed value of  $c^2 = 18$  is falling within the acceptance Region (limits obtained from the tables) for 19 D.F. at 5% LOS, we may accept  $H_0$  and conclude that the sample evidence presents no reason to doubt the representative's claim.

**16. A random sample of size 25 from a normal population gives a sample mean of 42 and a sample standard deviation of 6. Test the hypothesis that the population S.D. is 9. Clearly state the alternative hypothesis you allow for and the level of significance adopted.**

*Sol.:*

1. Null Hypothesis ( $H_0$ ) :  $s^2 = 81$   
Alternative Hypothesis ( $H_1$ ) :  $s^2 > 81$
2. The data given is:  $n = 25$ ,  $s = 6$ ,  $\sigma_0^2 = (9)^2 = 81$
3. Test Statistic :  $\chi^2 = \frac{ns^2}{\sigma_0^2} = \frac{25 \times 36}{81} = 11.11$

Degrees of freedom =  $n - 1 = (25 - 1) = 24$

**Table Value**

At 1% LOS (selected), the table value of  $\chi^2$  for 24 D.F. (one tailed-right tail) is 42.979781

**Conclusion**

Since the computed value of  $c^2 = 11.11$  is less than the table value of  $\chi^2 = 42.979781$  for 24 D.F. at 1% LOS, we may accept  $H_0$  and conclude that the population S.D. is 9.

**4.3.2 Test for Goodness of Fit**

**Q12. Explain briefly about Chi-square test as test of goodness of fit.**

*Ans.:*

(Nov.-20, Imp.)

One of the very popular applications of  $\chi^2$  test is test of goodness of fit. It enables us to ascertain how the theoretical distributions such as Binomial, Poisson, Normal etc. can fit into empirical distributions obtained from sample data. When an ideal frequency curve whether normal or some other type is fitted to the data, we are interested out how well this curve fits with the observed facts.

A test of the concordance (goodness of fit) of the two can be made just by inspection, but such a test is obviously inadequate. Precision can be secured by applying the  $\chi^2$  test. The following are the steps in testing the goodness of fit :

1. Null and alternative hypotheses are established, and a significance level is selected for rejection of the null hypothesis.
2. A random sample of independent observations is drawn from a relevant statistical population.
3. A set of expected or theoretical frequencies is derived under the assumptions that the null hypothesis is true. This generally takes the form of assuming that a particular probability distribution is applicable to the statistical population under consideration.
4. The observed frequencies are compared with the expected, or theoretical frequencies.
5. If the calculated value of  $\chi^2$  is less than the table value at a certain level of significance (generally 5% level) and for certain degrees of freedom the fit is considered to be good, i.e., the divergence between the actual and expected frequencies is attributed to random fluctuations of sampling.

On the other hand, if the calculated value of  $\chi^2$  is greater than the table value, the fit is considered to be poor, i.e., it cannot be attributed to fluctuations of sampling rather it is due to the inadequacy of the theory to fit the observed facts.

### Goodness of Fit

$\chi^2$  test help us to find out how well the assumed theoretical distribution fit to the observed data. When some theoretical distribution is fitted to the given data, the statistician or managers will be interested in knowing as to how this distribution fits with the observed data.

This method of  $\chi^2$  test helps in answering this question.

If the calculated value of  $\chi^2$  is less than the table value at a certain level of significance, the fit is considered to be good one i.e., divergence between the observed and expected frequencies is attributable to fluctuations of sampling. But if the reverse occurs, the fit is not considered to be a good one. In short,

$$\chi_{\text{cal}}^2 < \chi_{\text{table}}^2 \Rightarrow \text{Good fit}$$

$$\chi_{\text{cal}}^2 > \chi_{\text{table}}^2 \Rightarrow \text{Not a good fit.}$$

If  $1, 2, \dots, n$  is a set of observed (experimental) frequencies and  $E_i$  ( $i = 1, \dots, n$ ) is the corresponding set of theoretical frequencies then  $\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$  with the condition that,  $\sum_{i=1}^n O_i = \sum_{i=1}^n E_i = N = \text{Total frequency}$  follows,  $\chi^2$  - Distribution with  $(n - 1)$  d.o.f.

### Steps for Test of Goodness of Fit

1. Null hypothesis : Good fit exists between the theoretical distribution and given data.
2. Alternative hypothesis : No good fit.
3. Level of significance is  $\alpha$ .
4. Critical region : Reject null hypothesis if  $\chi^2 > \chi_{\alpha}^2$  with  $v$  d.o.f. i.e., theoretical distribution is a poor fit.
5. Computations :  $\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$
6. **Decision** : Accept null hypothesis, if  $\chi^2 < \chi_{\alpha}^2$ , i.e., the theoretical distribution is a good fit to the data.

**PROBLEMS**

17. The number of automobile accidents per week in a certain community are as follows : 12, 8, 20, 2, 14, 10, 15, 6, 9, 4. Are these frequencies in agreement with the be that accident conditions were the same during this 10 week period.

*Sol :*

Expected frequency of accidents each week =  $\frac{100}{10} = 10$ .

**Null Hypothesis  $H_0$  :** The accident conditions were the same during the 10 week period.

**Alternative Hypothesis :** The accidents conditions are different during the 10 week period.

Observed Frequency ( $O_i$ )	Expected Frequency ( $E_i$ )	$(O_i - E_i)$	$\frac{(O_i - E_i)^2}{E_i}$
12	10	2	0.4
8	10	-2	0.4
20	10	10	10.0
2	10	-8	6.4
14	10	4	1.6
10	10	0	0.0
15	10	5	2.5
6	10	-4	1.6
9	10	-1	0.1
4	10	-6	3.6
100	100		26.6

Now  $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 26.6$  i.e., Calculated  $\chi^2 = 26.6$

Here  $n = 10$  observations are given

Degrees of freedom (d.f) =  $n - 1 = 10 - 1 = 9$

Tabulated  $\chi^2 = 16.9$

Since Calculated  $\chi^2 >$  Tabulated  $\chi^2$  therefore, the Null Hypothesis is rejected and conclude that the accident conditions were not the same during the 10 week period.

18. A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured a third class, 90 were placed in second class and 20 got a first class. Do these figures commensurate with the general examination result which is in the ratio of 4 : 3 : 2 : 1 for the various categories respectively.

*Sol :*

**Null Hypothesis  $H_0$  :** The observed results commensurate with the general examination results.

Expected frequencies are in the ratio of 4 : 3 : 2 : 1

Total frequency = 500

If we divide the total frequency 500 in the ratio 4 : 3 : 2 : 1, we get the expected frequencies as 200, 150, 100, 50.

#### Calculations for $\chi^2$

Class	Observed Frequency ( $O_i$ )	Expected Frequency ( $E_i$ )	$(O_i - E_i)$	$\frac{(O_i - E_i)^2}{E_i}$
Failed	220	200	20	2.00
Third	170	150	20	2.667
Second	90	100	-10	1.000
First	20	50	-30	18.00
	500	500		23.667

$$\text{Calculated } \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 23.667$$

Degrees of freedom = 4 - 1 = 3 i.e.,  $v = 3$

For  $v = 3$ ,  $\chi_{0.05}^2 = 7.81$

i.e., tabulated value of  $\chi^2$  at 5% level for 3 d.f = 7.81

Since calculated  $\chi_{0.05}^2 >$  tabulated  $\chi_{0.05}^2$ , we reject the null hypothesis.

i.e., the observed results are not commensurate with the general examination results.

19. The following figures show the distribution of digits in numbers chosen a random from a telephone directory.

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	1026	1107	997	966	1075	933	1107	972	964	853

Test whether the digits may be taken to occur equally frequently in the director.

*Sol :*

**Null Hypothesis  $H_0$  :** The digits occur equally frequently in the directory.

**Alternative Hypothesis  $H_1$  :** The digits do not occur equally frequently.

Under this null hypothesis the expected frequencies for each of the digits

$$0, 1, 2, \dots, 9 \text{ is } \frac{10,000}{10} = 1000.$$

**Calculations for  $\chi^2$**

Digits	Observed Frequency ( $O_i$ )	Expected Frequency ( $E_i$ )	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	1026	1000	676	0.676
1	1107	1000	11449	11.449
2	997	1000	9	0.009
3	966	1000	1156	1.156
4	1075	1000	5625	5.625
5	933	1000	4489	4.489
6	1107	1000	11449	11.449
7	972	1000	784	0.784
8	964	1000	1296	1.296
9	853	1000	21609	21.609
Total	10000	10,000		58.542

$$\text{Calculated } \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 58.542$$

$$\text{Degrees of freedom} = n - 1 = 10 - 1 = 9$$

The tabulated value of for 9 d.f. at 5% level = 16.919

Since calculated  $\chi^2 >$  tabulated  $\chi^2$ , we reject the null hypothesis  $H_0$  and conclude that the digits do not occur equally frequently in the directory.

20. A die is thrown 264 times with the following results. Show that the die is biased. [Given  $\chi^2_{0.05} = 11.07$  for 5 d.f.]

No. appeared on the die	1	2	3	4	5	6
Frequency	40	32	28	58	54	52

*Sol:*

**Null Hypothesis  $H_0$ :** The die is unbiased.

The expected frequency of each of the numbers 1, 2, 3, 4, 5, 6 is

$$\frac{264}{6} = 44$$



## Calculations for X

Observed frequency ( $O_i$ )	Expected frequency ( $E_i$ )	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
40	44	16	0.3636
32	44	144	3.2727
28	44	256	5.8181
58	44	196	4.4545
54	44	100	2.2727
52	44	64	1.4545
264	264		17.6362

$$\therefore \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 17.6362$$

The number of degrees of freedom =  $n - 1 = 5$

The tabulated value of  $\chi^2$  for 5 d.f at 5% level = 11.07

Since calculated  $\chi^2 >$  tabulated  $\chi^2$ , we reject the null hypothesis  $H_0$

i.e., we reject the hypothesis that the die is unbiased. Hence the die is biased.

21. The following results are obtained when a dice is thrown 132 times. Test the Hypothesis that the dice is unbiased

No. Turned up	1	2	3	4	5	6
Frequency	16	20	25	14	29	28

*Sol :*

## Chi-Square Test

## Step - 1

## Null Hypothesis :

$H_0$  : The dice is unbiased

## Alternative Hypothesis :

$H_1$  : The dice is not unbiased

$$\text{Expected (E)} = \frac{132}{6} = 22 \text{ times}$$

**Step - 2 : Computing Test Statistics  $\chi^2$** 

S.No.	O	E	O - E	(O - E) <sup>2</sup>	$\frac{(O - E)^2}{E}$
1	16	22	- 6	36	1.64
2	20	22	- 2	4	0.18
3	25	22	3	9	0.41
4	14	22	- 8	64	2.91
5	29	22	7	49	2.23
6	28	22	6	36	1.64
					$\Sigma \left[ \frac{(O - E)^2}{E} \right] = 9.01$

$$\begin{aligned}\chi^2 &= \Sigma \left[ \frac{(O - E)^2}{E} \right] \\ &= 9.01\end{aligned}$$

**Step - 3 : Level of Significance :**

Level of significance =  $\alpha$  = 0.05 (Assumed)

Degree of freedom =  $n - 1$

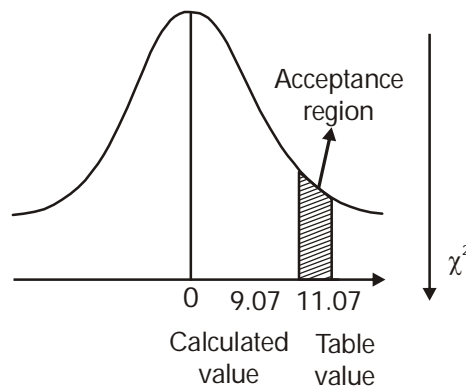
$$= 6 - 1 = 5$$

**Step - 4 :**

Table value of  $\chi^2$  for 5 d. at 5% level of significance is 11.07.

**Step - 5 :**

Table calculated value of  $\chi^2$  is 9.01 which is less than the table value i.e., 11.07. Hence, Null hypothesis is accepted, which states that the dice is unbiased.

**Fig. :**

### 4.3.3 Test for Independence of Attributes

**Q13. Explain the Chi-Square Test for independence of attributes.**

*Ans :*

**(Imp.)**

An attribute means a quality or characteristic. Example of attributes are drinking, smoking, blindness, honesty, beauty etc.

An attribute may be marked by its presence (position) or absence in a number of a given population. Let the observations be classified according to two attribute and the frequencies  $O_i$  in the different categories be shown in a two-way table called contingency table. We have to test on the basis of cell frequencies whether the two attributes are independent or not. We take the Null - Hypothesis  $H_0$  that there is no association between the attributes i.e., we assume that the two attributes are independent. The expected

frequencies ( $E_i$ ) of any cell =  $\frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$

The test statistic  $\chi^2 = \sum_i \left[ \frac{(O_i - E_i)^2}{E_i} \right]$  approximately follows Chi-square distribution with d.f. = (No. of rows - 1)  $\times$  (No. of columns - 1)

If the calculated value of  $\chi^2$  is less than the table value at a specified level (generally 5%) of significance, the hypothesis holds good i.e., the attributes are independent and do not bear any association. On the other hand, if the calculated value of  $\chi^2$  is greater than the table value at a specified level of significance, we say that the results of the experiment do not support the hypothesis, in other words, the attributes are associated.

Let us consider two attributes A and B. A is divided into two classes and B is divided into two classes. The various cell frequencies can be expressed in the following table known as  $2 \times 2$  contingency table.

	A	a	b
	B	c	d
a	b		a + b
c	d		c + d
a + c	b + d		N = a + b + c + d

The expected frequencies are given by

$E(a) = \frac{(a+c)(a+b)}{N}$	$E(b) = \frac{(b+d)(a+b)}{N}$	a + b
$E(c) = \frac{(a+c)(c+d)}{N}$	$E(d) = \frac{(b+d)(c+d)}{N}$	c + d
a + c	b + d	N = a + b + c + d

The value of  $\chi^2$  is given by  $\chi^2 = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$  where

$N = a + b + c + d$  with d.f. =  $(2 - 1)(2 - 1) = 1$ . We use this formula when the expected frequencies are in fractions (or decimals).

**PROBLEMS**

22. On basis of information given below about the treatment of 200 patients suffering from a disease, state whether the new treatment is comparatively superior to the conventional treatment.

	Favourable	Not favourable	Total
New	60	30	90
Conventional	40	70	110

*Sol :*

**Null Hypothesis  $H_0$  :** No difference between new and conventional treatment (or) New and conventional treatment are independent.

The number of degrees of freedom is  $(2 - 1)(2 - 1) = 1$

Expected frequencies are given in the table :

$$\text{Expected frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

$\frac{90 \times 100}{200} = 45$	$\frac{90 \times 100}{200} = 45$	90
$\frac{100 \times 110}{200} = 55$	$\frac{100 \times 110}{200} = 55$	110
100	100	200

Calculation of  $\chi^2$  :

Observed Frequency ( $O_i$ )	Expected Frequency ( $E_i$ )	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
60	45	225	5
30	45	225	5
40	55	225	4.09
70	55	225	4.09
200	200		18.18

$$\therefore \chi^2 = \sum \frac{(O - E)^2}{E} = 18.18$$

Tabulated  $\chi^2$  for 1 d.f. at 5% level of significance is 3.841.

Since calculated  $\chi^2 >$  tabulated  $\chi^2$  we reject the null hypothesis  $H_0$  i.e., new and conventional treatment are not independent. The new treatment is comparatively superior to conventional treatment.

23. The following table gives the classification of 100 workers according to sex and nature of work. Test whether the nature of work is independent of the sex of the worker.

	Stable	Unstable	Total
Males	40	20	60
Females	10	30	40
Total	50	50	100

*Sol :*

**Null Hypothesis  $H_0$  :** The nature of work is independent of the sex of the workers. Expected frequencies are given in the table :

$\frac{50 \times 60}{100} = 30$	$\frac{50 \times 40}{100} = 20$	60
$\frac{50 \times 40}{100} = 20$	$\frac{50 \times 60}{100} = 30$	40
50	50	100

Calculation of  $\chi^2$  :

$O_i$	$E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
40	30	100	3.333
20	30	100	3.333
10	20	100	5.000
30	20	100	5.000
100	100	$\sum \frac{(O_i - E_i)^2}{E_i}$	16.66

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 16.66$$

$$\therefore \text{Calculated } \chi^2 = 16.66$$

Tabulated value of  $\chi^2$  for  $(2 - 1) (2 - 1) = 1$  d.f. at 5% level of significance is 3.84.

Since calculated  $\chi^2 >$  tabulated  $\chi^2$ , we reject the null hypothesis  $H_0$ , i.e., the nature of work is not independent of the sex of the workers.

i.e., there is difference in the nature of work on the basis of sex.

24. In an anti malarial campaign in a certain area quinine was administered to 812 persons. Out of a total population of 3248 persons the number of fever cases is shown below :

	Fever	No Fever	Total
Quinine	20	792	812
No Quinine	220	2216	2436
Total	240	3008	3248

Discuss the usefulness of quinine in checking malaria.

*Sol :*

It is a  $\chi^2$  test

#### Null Hypothesis

$H_0$  = Quinine is not effective in checking malaria

$H_a$  = Quinine is effective in checking malaria

#### Computing Test Statistic

$$\chi^2 = \sum \left( \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right)$$

$$E_{ij} = \frac{RT \times CT}{GT}$$

#### Calculation of Expected Frequencies

Treatment	Fever	No Fever	Total
Quinine	$\frac{812 \times 240}{3,248} = 60$	$\frac{812 \times 3,008}{3,248} = 752$	812
	$E_{11}$	$E_{12}$	
No Quinine	$\frac{2,436 \times 240}{3,248} = 180$	$\frac{2,436 \times 3,008}{3,248} = 2,256$	2,436
	$E_{21}$	$E_{22}$	
Total	240		3,248

Calculation of  $\chi^2$ 

Group	$O_{ij}$	$E_{ij}$	$O_{ij} - E_{ij}$	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
	(1)	(2)	(3) = (1) - (2)	(4)	(5) = $\frac{(4)}{(2)}$
11	20	60	- 40	1,600	26.667
12	792	752	40	1,600	2.128
21	220	180	40	1,600	8.889
22	2,216	2,256	- 40	1,600	0.709
				Total	38.393

$$\therefore \chi^2 = \sum \left( \frac{O_{ij} - E_{ij}}{E_{ij}} \right) = 38.393$$

Level of significance,  $\alpha = 0.05$  (Assumed)

$$\begin{aligned} \text{Degree of freedom} &= (c - 1)(r - 1) = (2 - 1)(2 - 1) \\ &= 1 \times 1 = 1 \end{aligned}$$

Table of  $\chi^2$  at 1 d.f and 0.05 is 3.84

Since calculated  $\chi^2 \geq \chi_{tab}^2$ , we reject null hypothesis. Hence, quinine is effective in checking malaria.

25. Three sales persons account for the following amount of sales (in thousand of rupees) over a period of 4 months. Test for the significant difference in their performance.

Sales	Monthly Sales in Thousand of Rupees			
X	48	49	50	49
Y	47	49	48	48
Z	49	51	50	50

Sol.:

It is a  $\chi^2$  test

#### Null Hypothesis

$H_0$  = There is no significant difference between the sales/

#### Alternate Hypothesis

$H_a$  = There is a significant difference between the sales.

## Computing Test Statistic

Sales	Monthly Sales in Thousand of Rupees				Total
X	48	49	50	49	196
Y	47	49	48	48	192
Z	49	51	50	50	200
Total	144	149	148	147	588

$$\chi^2 = \sum \left[ \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right]$$

$O_{ij}$  = Observed frequency at the cell in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column

$E_{ij}$  = Expected frequency at the cell in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column

$$\text{Expected frequency} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

$$\text{i.e., } E_{ij} = \frac{RT - CT}{GT} \quad (\text{Calculated for each box})$$

## Calculation of Expected Frequencies

Sales	Monthly Sales in Thousand of Rupees				Total
X	$\frac{196 \times 144}{588} = 48$ $E_{11}$	$\frac{196 \times 149}{588} = 49.67$ $E_{12}$	$\frac{196 \times 148}{588} = 49.33$ $E_{13}$	$\frac{196 \times 147}{588} = 49$ $E_{14}$	196
Y	$\frac{192 \times 144}{588} = 47.02$ $E_{21}$	$\frac{192 \times 149}{588} = 48.65$ $E_{22}$	$\frac{192 \times 148}{588} = 48.33$ $E_{23}$	$\frac{192 \times 147}{588} = 48$ $E_{24}$	192
Z	$\frac{200 \times 144}{588} = 48.98$ $E_{31}$	$\frac{200 \times 149}{588} = 50.68$ $E_{32}$	$\frac{200 \times 148}{588} = 50.34$ $E_{33}$	$\frac{200 \times 147}{588} = 50$ $E_{34}$	200
Total	144	149	148	147	588



Group	O <sub>ij</sub>	E <sub>ij</sub>	O <sub>ij</sub> - E <sub>ij</sub>	(O <sub>ij</sub> - E <sub>ij</sub> ) <sup>2</sup>	$\frac{(O_{ij} - E_{ij})}{E_{ij}}$
	(1)	(2)	(3) = (1) - (2)	(4)	(5) = $\frac{(4)}{(5)}$
11	48	48	0	0	0
12	49	49.67	- 0.67	0.4489	0.0090
13	50	49.33	0.67	0.4489	0.0091
14	49	49	0	0	0
21	47	47.02	- 0.02	0.0004	0.000008
22	49	48.65	0.35	0.1225	0.0025
23	48	48.33	- 0.33	0.1089	0.0022
24	48	48	0	0	0
31	49	48.98	0.02	0.0004	0.000008
32	51	50.68	0.32	0.1024	0.0020
33	50	50.34	- 0.34	0.1156	0.0023
34	50	50	0	0	
					0.027116

$$\therefore \chi^2 = \sum \left[ \frac{O_{ij} - E_{ij}}{E_{ij}} \right] = 0.027116$$

Level of significance,  $\alpha = 0.05$  (Assumed)

$$\begin{aligned} \text{Degrees of freedom} &= (c - 1) (r - 1) = (4 - 1) (3 - 1) \\ &= 3 \times 2 = 6 \end{aligned}$$

Table of  $\chi^2$  at 6 d.f. and 0.05 is 12.592

Since calculated  $\chi^2 < \chi^2_{\text{tab}}$ ,  $H_0$  is accepted. Hence, there is no significant difference between the sales.

## Short Question and Answers

### 1. Test for Goodness of Fit

*Ans :*

One of the very popular applications of  $\chi^2$  test is test of goodness of fit. It enables us to ascertain how the theoretical distributions such as Binomial, Poisson, Normal etc. can fit into empirical distributions obtained from sample data. When an ideal frequency curve whether normal or some other type is fitted to the data, we are interested out how well this curve fits with the observed facts.

A test of the concordance (goodness of fit) of the two can be made just by inspection, but such a test is obviously inadequate. Precision can be secured by applying the  $\chi^2$  test. The following are the steps in testing the goodness of fit :

- (i) Null and alternative hypotheses are established, and a significance level is selected for rejection of the null hypothesis.
- (ii) A random sample of independent observations is drawn from a relevant statistical population.
- (iii) A set of expected or theoretical frequencies is derived under the assumptions that the null hypothesis is true. This generally takes the form of assuming that a particular probability distribution is applicable to the statistical population under consideration.
- (iv) The observed frequencies are compared with the expected, or theoretical frequencies.
- (v) If the calculated value of  $\chi^2$  is less than the table value at a certain level of significance (generally 5% level) and for certain degrees of freedom the fit is considered to be good, i.e., the divergence between the actual and expected frequencies is attributed to random fluctuations of sampling.

On the other hand, if the calculated value of  $\chi^2$  is greater than the table value, the fit is considered to be poor, i.e., it cannot be attributed to fluctuations of sampling rather it is due to the inadequacy of the theory to fit the observed facts.

### Goodness of Fit

$\chi^2$  test help us to find out how well the assumed theoretical distribution fit to the observed data. When some theoretical distribution is fitted to the given data, the statistician or managers will be interested in knowing as to how this distribution fits with the observed data.

This method of  $\chi^2$  test helps in answering this question.

If the calculated value of  $\chi^2$  is less than the table value at a certain level of significance, the fit is considered to be good one i.e., divergence between the observed and expected frequencies is attributable to fluctuations of sampling. But if the reverse occurs, the fit is not considered to be a good one. In short,

$$\chi_{cal}^2 < \chi_{table}^2 \Rightarrow \text{Good fit}$$

$$\chi_{cal}^2 > \chi_{table}^2 \Rightarrow \text{Not a good fit.}$$

If  $O_i = 1, 2, \dots, n$  is a set of observed (experimental) frequencies and  $E_i$  ( $i = 1, \dots, n$ ) is the corresponding set of theoretical frequencies then

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad \text{with the condition that,}$$

$\sum_{i=1}^n O_i = \sum_{i=1}^n E_i = N = \text{Total frequency}$  follows,  $\chi^2$  - Distribution with  $(n - 1)$  d.o.f.

### 2. Paired t-test.

*Ans :*

Paired observations arise in many practical situations where each homogeneous experimental unit receives both population conditions. As a result, each experimental unit has a pair of observations, one for each population.

For instance, to test the effectiveness of "drug" some 11 persons blood pressure is measured "before" and "after" the intake of certain drug. Here the individual person is the experimental unit and the two populations are blood pressure "before"

and "after" the drug is given. Thus for each observation in one sample, there is a corresponding observation in the other sample pertaining to the same character. Hence the two samples are not independent.

Consider another example. Suppose a business concern is interested to know whether a particular media of promoting sales of a product is really effective or not. In this case we have to test whether the average sales before and after the sales promotion are equal.

If  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , be the pairs of sales data before and after the sales production in a business concern, we apply paired t - test to examine the significance of the difference of the two situations.

$$\text{Let } d_i = x_i - y_i \text{ (or) } y_i - x_i \\ \text{for } i = 1, 2, 3, \dots, n$$

Let the Null Hypothesis be  $H_0 : \mu_1 = \mu_2$  (i.e.,  $\mu = 0$ ), there is no significant difference between the means in two situations.

Then the Alternative Hypothesis is

$$H_1 : \mu_1 \neq \mu_2$$

Assuming that  $H_0$  is true, the test statistic for  $n$  paired observations (which are dependent) by taking the differences  $d_1, d_2, \dots, d_n$  of the paired data.

$$t = \frac{\bar{d} - \mu}{S/\sqrt{n}} = \frac{\bar{d}}{S/\sqrt{n}} \quad (\because \mu = 0)$$

where

$$t = \bar{d} = \frac{1}{n} \sum d_i \text{ and } S^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$$

$$\text{or } S^2 = \frac{\sum d^2 - n(\bar{d})^2}{n-1}$$

$$\text{or } \frac{1}{n-1} \left[ \sum d^2 - \frac{(\sum d)^2}{n} \right]$$

are the mean and variance of the differences  $d_1, d_2, \dots, d_n$  respectively and  $p$  is the are of the population of differences.

The above statistic follows student's t-distribution with  $(n - 1)$  degrees of freedom.

### 3. What is Small sample test?

*Ans :*

Small sample size referred to size of sample which is less than 30. In case of small sample size the z-test is not appropriate test statistic as the assumptions on which it is based do not hold good in case of small sample. The theoretical work on t-distribution was done by W.S. Gosset (1876-1937) under the pen name "student" as he was the employee of the company Gunnies & Sons, a Dublin bravery, Ireland, which did not allowed it employees to publish research findings under their own names. The t-distribution is used when sample size is less than 30 and the population standard deviation is not known.

### 4. Assumptions of t-test.

*Ans :*

The following are the pre-requisites for the application of t-test :

- (i) The population from which a sample is drawn is normal.
- (ii) The samples have been drawn at random.
- (iii) The population standard deviation is not known.
- (iv) Sample size should small i.e., less than 30.

### 5. State the applications of the t-distribution.

*Ans :*

- (i) To test the significance of the sample mean, when population variance is not given
- (ii) To test the significance of the mean of the sample i.e., to test if the sample mean differs significantly from the population mean.
- (iii) To test the significance of the difference between two sample means or to compare two samples.
- (iv) To test the significance of an observed sample correlation coefficient and sample regression coefficient.

**6. What is ANOVA?***Ans :*

The variance test is also known as ANOVA. ANOVA is the acronym for Analysis of Variance. Analysis of variance is a statistical technique specially designed to test whether the means of more than two quantitative populations are equal i.e., to make inferences about whether those samples are drawn from the populations having the same mean.

The test is called 'F' test as it was developed by R.A Fisher in 1920's. The test is conducted in situations where we have three or more to consider, at a time an alternative procedure (to t-test) needed for testing the hypothesis that all samples could likely be drawn from the same population.

**Example**

Five fertilizers are applied to four plots, each of wheat and yield of wheat on these plots is given. We are interested in finding out whether the effects of these fertilizers on the yields are significantly different or, in other words, whether the samples have come from the same population. ANOVA answers this question.

**7. Applications of ANOVA.***Ans :*

The applications of ANOVA are as follows,

- (i) Anova is used in education, industry, business, psychology fields mainly in their experiment design.
- (ii) Anova helps to save time and money as several population means can be compared simultaneously.
- (iii) Anova is used to test the linearity of the fitted regression line and correlation ratio, significance test statistic of anova

$$= F(r - 1, n - r).$$

**8. Chi-Square Test.***Ans :*

The magnitude of discrepancy between the theory and observation is given by the quantity  $\chi^2$  (a Greek letter, pronounced as "chi-square"). If  $\chi = 0$ , the observed and expected frequencies

completely coincide. As the value of  $\chi^2$  increases, the discrepancy between the observed and theoretical frequencies increases. Thus,  $\chi^2$  affords a measure of the correspondence between theory and observation.

**Definition**

If a set of events  $A_1, A_2, \dots, A_n$  are observed to occur with frequencies  $O_1, O_2, \dots, O_n$  respectively and according to probability rules  $A_1, A_2, \dots, A_n$  are expected to occur with frequencies  $E_1, E_2, \dots, E_n$  respectively with  $O_1, O_2, \dots, O_n$  are called observed frequencies and  $E_1, E_2, \dots, E_n$  are called expected frequencies.

If  $O_i$  ( $i = 1, 2, \dots, n$ ) is a set of observed (experimental) frequencies and  $E_i$  ( $i = 1, 2, \dots, n$ ) is the corresponding set of expected (theoretical)

frequencies, then  $\chi^2$  is defined as  $\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$  with  $(n - 1)$  degrees of freedom.

$\chi^2$  is used to test whether differences between observed and expected frequencies are significant.

**9. Chi-Square Test for independence of attributes.***Ans :*

An attribute means a quality or characteristic. Example of attributes are drinking, smoking, blindness, honesty, beauty etc.

An attribute may be marked by its presence (position) or absence in a number of a given population. Let the observations be classified according to two attribute and the frequencies  $O_i$  in the different categories be shown in a two-way table called contingency table. We have to test on the basis of cell frequencies whether the two attributes are independent or not. We take the Null - Hypothesis  $H_0$  that there is no association between the attributes i.e., we assume that the two attributes are independent. The expected frequencies ( $E_i$ ) of

$$\text{any cell} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

The test statistic  $\chi^2 = \sum_i \left[ \frac{(O_i - E_i)^2}{E_i} \right]$  approximately follows Chi-square distribution with d.f. = (No. of rows – 1)  $\times$  (No. of columns – 1)

If the calculated value of  $\chi^2$  is less than the table value at a specified level (generally 5%) of significance, the hypothesis holds good i.e., the attributes are independent and do not bear any association. On the other hand, if the calculated value of  $\chi^2$  is greater than the table value at a specified level of significance, we say that the results of the experiment do not support the hypothesis, in other words, the attributes are associated.

#### 10. What are the properties of t-distribution?

*Ans :*

- (i) The shape of t-distribution is bell-shaped, which is similar to that of a normal distribution and is symmetrical about the mean.
- (ii) The t-distribution curve is also asymptotic to the t-axis, i.e., the two tails of the curve on both sides of  $t = 0$  extends to infinity.

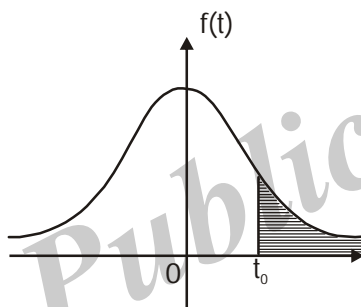


Fig.: t-distribution

- (iii) It is symmetrical about the line  $t = 0$ .
- (iv) The form of the probability curve varies with degrees of freedom i.e., with sample size.
- (v) It is unimodal with Mean = Median = Mode.
- (vi) The mean of standard normal distribution and as well as t-distribution zero but the variance of t-distribution depends upon the parameter  $v$  which is called the degrees of freedom.

## Exercise Problems

1. Two salesmen there A and B are working in a certain district. From a sample survey conducted by the Head office, the following results were obtained. State whether there is any significant difference in the average sales between two salesmen.

<b>No. of Sales</b>	20	18
Average sales (in Rs.)	170	205
Standard deviation (in Rs.)	20	25

**[Ans :  $H_0 : \mu_1 = \mu_2$ ,  $t = -4.65$ , significant]**

2. Memory capacity of 10 students were tested before and after training. State whether the training was effective (or) not from the following scores.

Before training	12	14	11	8	7	10	3	0	5	6
After training	15	16	10	7	5	12	10	2	3	8

**[Ans :  $H_0 : \mu_1 = \mu_2$ ,  $H_1 : \mu_1 \neq \mu_2$ ,  $t = -1.3646$ , not significant]**

3. In one sample of 10 observations, the sum of the squares of the deviations of the sample values from mean was 120 and in the other sample of 12 observations, it was 314. Test whether the difference in variances is significant at 5% variances level.

**[Ans : Calculated F = 2.14, not significant]**

4. The random samples were drawn from two normal populations and the following results were obtained.

<b>Sample I</b>	16	17	18	19	20	21	22	24	26	27	-	-
<b>Sample II</b>	19	22	23	25	26	28	29	30	31	32	35	36

**[Ans :  $\sigma_1^2 = \sigma_2^2$ , F = 1.935, not significant]**

## Choose the Correct Answers

1. When the sample size is 30 or less than 30 it is considered as [ a ]  
 (a) Small samples (b) Large samples  
 (c) Sample mean (d) S.D sample.
2. The \_\_\_\_\_ one of the important applications of t-distribution [ b ]  
 (a) Only test of hypothesis  
 (b) Test of hypothesis about the population mean  
 (c) Test of hypothesis for large samples  
 (d) Test of sample size
3. \_\_\_\_\_ is the acronym for analysis of variances [ c ]  
 (a) ANVI (b) AIVS  
 (c) ANOVA (d) Analysis
4. There are \_\_\_\_\_ types of ANOVA [ d ]  
 (a) One (b) Three  
 (c) Five (d) Two
5. Correction factor is given by \_\_\_\_\_ [ a ]  
 (a)  $\frac{T^2}{N}$  (b)  $\frac{T}{N}$   
 (c)  $\frac{N}{T}$  (d)  $\sqrt{TN}$
6. F-ratio is calculated is [ b ]  
 (a) Variance (b)  $\frac{\text{Greater variance}}{\text{Smaller variance}}$   
 (c) Greater variance  $\times$  smaller variance (d) Greater variance – smaller variance
7. One of the application of  $\chi^2$  - distribution is [ c ]  
 (a) To know the sample size (b) To know the degree of freedom  
 (c) To conduct  $\chi^2$  test of goodness of fit. (d)  $\chi^2$  test for dependence of attributes.
8.  $\chi^2 = \Sigma$  [ d ]  
 (a)  $\left[ \frac{O_{ij} \times E_{ij}}{E_{ij}} \right]$  (b)  $\frac{(E_{ij} - O_{ij})}{(E_{ij})^2}$   
 (c)  $[O_{ij} - E_{ij} + E_{ij}]$  (d)  $\left[ \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right]$

9. The variance of t-distribution is always greater than one and is defined only when [ a ]
- (a)  $n \geq 3$  (b)  $n = 3$   
(c)  $n \leq 3$  (d)  $n < 3$
10. The t-distribution has \_\_\_\_\_ dispersion than the normal distribution. [ b ]
- (a) Lesser (b) Greater  
(c) Equal (d) No.

Rahul Publications



## Fill in the blanks

1. \_\_\_\_\_ test help us to find out how well the assumed theoretical distribution fits to the observed data.
2. If  $\chi^2_{\text{cal}} < \chi^2_{\text{table}}$  denotes that it is \_\_\_\_\_
3. \_\_\_\_\_ is essentially a procedure for testing the difference among different groups of data for homogeneity.
4. \_\_\_\_\_ is calculated as a ratio of squares of the mean between samples and squares of the mean within samples.
5. The correlation suggested by yates is popularly known as \_\_\_\_\_ .
6. \_\_\_\_\_ for a sample of size 'n' is calculated as  $(n - 1)$ .
7. Expected frequency = \_\_\_\_\_ .
8. The sampling distribution of small samples follow \_\_\_\_\_ .
9. Sum of squares of residual or error is \_\_\_\_\_ .
10. \_\_\_\_\_ test is used when large samples are tested and/or population variance is known.

### ANSWERS

1.  $\chi^2$
2. Good fit
3. ANOVA
4. F-ratio
5. Yates correction
6. Degree of freedom
7. 
$$\frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$
8. Students t-distribution
9.  $SS_{\text{res}} = SST - (SSC + SSR)$
10. Z - test

# UNIT V

- (i) **Correlation Analysis:** Scatter diagram, Positive and negative correlation, limits for coefficient of correlation, Karl Pearson's coefficient of correlation, Spearman's Rank correlation, concept of multiple and partial Correlation.
- (ii) **Regression Analysis:** Concept, least square fit of a linear regression, two lines of regression, properties of regression coefficients.
- (iii) **Time Series Analysis:** Components, Models of Time Series-Additive, Multiplicative and Mixed models; Trend analysis-Free hand curve, Semi averages, moving averages, Least Square methods

## 5.1 CORRELATION ANALYSIS

**Q1. Define the term correlation.**

**(OR)**

**What do you mean by correlation?**

*Ans :* (Aug.-17)

**Meaning**

Correlation is the study of the linear relationship between two variables. When there is a relationship of 'quantitative measure between two set of variables, the appropriate statistical tool for measuring the relationship and expressing each in a precise way is known as correlation.

**For example,** there is a relationship between the heights and weights of persons, demand and prices of commodities etc.

Correlation analysis is the statistical tool we can use to describe the degree to which one variable is linearly related to another.

**Definitions**

- (i) **According to L.R. Connor** "If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in other(s) then they are said to be correlated."
- (ii) **According to A.M. Tuttle** "Correlation is an analysis of covariation between two or more variables".
- (iii) **According to Croxton and Cowden** "When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship

and expressing it in a brief formula is known as correlation."

- (iv) **According to Ya Lun Chou** "Correlation analysis attempts to determine the 'degree of relationship' between variable".

**Q2. Explain the Significance of Measuring Correlation.**

**(OR)**

**How do you say that the correlation between the two variables is significant (or) not.**

*Ans :* (June-19, Imp.)

1. Correlation is very useful to economists to study the relationship between variables, like price and quantity demanded. To businessmen, it helps to estimate costs, sales, price and other related variables.
2. In economic theory we come across several types of variables which show some kind of relationship. For example, there exists a relationship between price, supply and quantity demanded; convenience, amenities, and service standards are related to customer retention; yield a crop related to quantity of fertilizer applied, type of soil, quality of seeds, rainfall and so on. Correlation analysis helps in measuring the degree of association and direction of such relationship.
3. The relation between variables can be verified and tested for significance, with the help of the correlation analysis. The effect of correlation is to reduce the range of uncertainty of our prediction.

4. The coefficient of correlation is a relative measure and we can compare the relationship between variables, which are expressed in different units.
5. Correlations are useful in the areas of health care such as determining the validity and reliability of clinical measures or in expressing how health problems are related to certain biological or environmental factors. For example, correlation coefficient can be used to determine the degree of inter-observer reliability for two doctors who are assessing a patient's disease.
6. Sampling error can also be calculated.
7. Correlation is the basis for the concept of regression and ratio of variation.
8. The decision making is heavily facilitated by reducing the range of uncertainty and hence empowering the predictions.

**Q3. Explain the scope of correlation analysis.**

*Ans :*

**(i) One of the variable may be affecting the other**

A correlation coefficient calculated from the data on quantity demanded and corresponding price of tea would only reveal that the degree of association between them is very high. It will not give us any idea about whether price is affecting demand of tea or vice-versa. In order to know this, we need to have some additional information apart from the study of correlation.

For example if, on the basis of some additional information, we say that the price of tea affects its demand, then price will be the cause and quantity will be the effect. The causal variable is also termed as independent variable while the other variable is termed as dependent variable.

**(ii) The two variables may act upon each other**

Cause and effect relation exists in this case also but it may be very difficult to find out which of the two variables is independent.

For example, if we have data on price of wheat and its cost of production, the correlation between them may be very high because higher price of wheat may attract farmers to produce more wheat and more production of wheat may mean higher cost of production, assuming that it is an increasing cost industry. Further, the higher cost of production may in turn raise the price of wheat. For the purpose of determining a relationship between the two variables in such situations, we can take any one of them as independent variable.

**(iii) The two variables may be acted upon by the outside influences**

In this case we might get a high value of correlation between the two variables, however, apparently no cause and effect type relation seems to exist between them.

For example, the demands of the two commodities, say X and Y, may be positively correlated because the incomes of the consumers are rising. Coefficient of correlation obtained in such a situation is called a spurious or nonsense correlation.

**(iv) A high value of the correlation coefficient may be obtained due to sheer coincidence (or pure chance)**

This is another situation of spurious correlation. Given the data on any two variables, one may obtain a high value of correlation coefficient when in fact they do not have any relationship. For example, a high value of correlation coefficient may be obtained between the size of shoe and the income of persons of a locality.

**Q4. State the applications of correlation.**

*Ans :*

- (a) Relationship between advertising expenditure and sales revenue of a company for a period of time.
- (b) Relationship between time or the day and the temperatures recorded for a particular city.
- (c) Relationship between ages of the husbands and wives for a group of couples under study.

- (d) Relationship between heights and weights of a group of students.
- (e) Relationship between age and blood pressure of men.
- (f) Relationship between imports of raw material and exports of finished goods of a firm in a particular year.
- (g) Relationship between amount invested (capital) and profits earned by different firms.
- (h) Relationship between price and demand for a particular commodity.

### 5.1.1 Types of Correlation

#### 5.1.1.1 Positive and negative correlation

**Q5. Explain various Types of Correlation.**

(or)

**What are the different Types of correlations.**

*Ans.:*

(Imp.)

Broadly speaking, there are four types of correlation, namely,

- A) Positive correlation,
- B) Negative correlation,
- C) Linear correlation and
- D) Non-Linear Correlation.

#### A) Positive correlation

If the values of two variables deviate in the same direction i.e., if increase in the values of one variable results, on an average, in a corresponding increase in the values of the other variable or if a decrease in the values of one variable results, on an average, in a corresponding decrease in the values of the other variable, the corresponding correlation is said to be positive or direct.

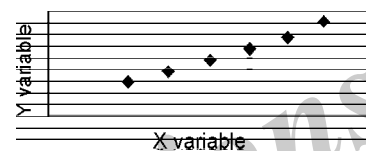
#### Examples

- i) Sales revenue of a product and expenditure on Advertising.
- ii) Amount of rain fall and yield of a crop (up to a point)

- iii) Price of a commodity and quantity of supply of a commodity
- iv) Height of the Parent and the height of the Child.
- v) Number of patients admitted into a Hospital and Revenue of the Hospital.
- vi) Number of workers and output of a factory.

#### i) Perfect Positive Correlation

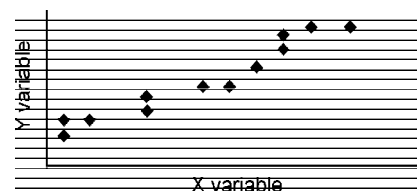
If the variables X and Y are perfectly positively related to each other then, we get a graph as shown in fig. below.



**Fig.: Perfect Positive Correlation ( $r = +1$ )**

#### ii) Very High Positive Correlation

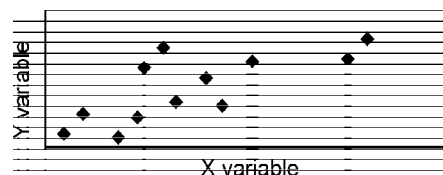
If the variables X and Y are related to each other with a very high degree of positive relationship then we can notice a graph as in figure below.



**Fig.: Very High Positive Correlation  
( $r = \text{nearly } +1$ )**

#### iii) Very Low Positive Correlation

If the variables X and Y are related to each other with a very low degree of positive relationship then we can notice a graph as in fig. below.



**Fig.: Very Low Positive Correlation  
( $r = \text{near to } +0$ )**

**B) Negative Correlation**

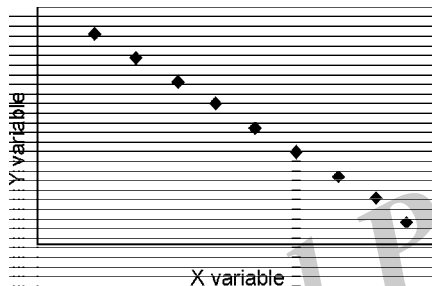
Correlation is said to be negative or inverse if the variables deviate in the opposite direction i.e., if the increase (decrease) in the values of one variable results, on the average, in a corresponding decrease (increase) in the values of the other variable.

**Examples**

1. Price and demand of a commodity.
2. Sales of Woolen garments and the day temperature.

**i) Perfect Negative Correlation**

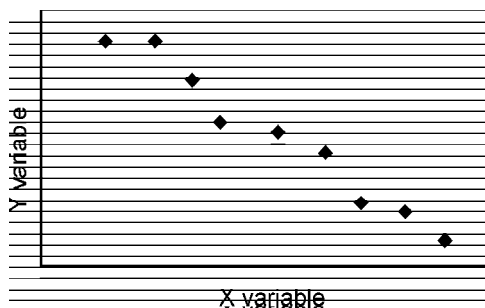
If the variables X and Y are perfectly negatively related to each other then, we get a graph as shown in fig. below.



**Fig.: Perfect Negative Correlation ( $r = -1$ )**

**ii) Very High Negative Correlation**

If the variables X and Y are related to each other with a very high degree of negative relationship then we can notice a graph as in fig. below.

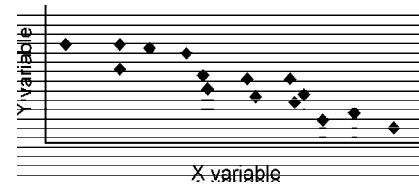


**Fig.: Very High Negative Correlation**

( $r = \text{near to } -1$ )

**iii) Very low Negative Correlation**

If the variables X and Y are related to each other with a very low degree of negative relationship then we can notice a graph as in fig. below.

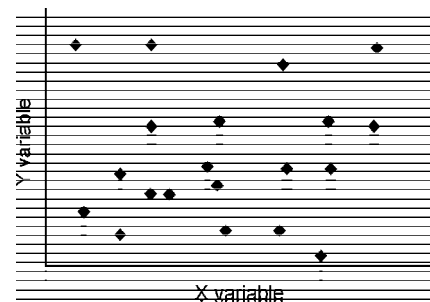


**Fig.: Very Low Negative Correlation**

( $r = \text{near to } 0 \text{ but negative}$ )

**iv) No Correlation**

If the scatter diagram show the points which are highly spread over and show no trend or patterns we can say that there is no correlation between the variables.



**Fig.: No Correlation ( $r = 0$ )**

**C) Linear Correlation**

Two variables are said to be linearly related if corresponding to a unit change in one variable there is a constant change in the other variable over the entire range of the values.

If two variables are related linearly, then we can express the relationship as

$$Y = a + bX$$

where ' $a$ ' is called as the "intercept" (If  $X = 0$ , then  $Y = a$ ) and ' $b$ ' is called as the "rate of change" or slope.

If we plot the values of X and the corresponding values of Y on a graph, then the graph would be a straight line as shown in fig. below.

### Example

X	1	2	3	4	5
Y	8	11	14	17	20

For a unit change in the value of x, a constant 3 units changes in the value of y can be noticed. The above can be expressed as :  $Y = 5 + 3x$ .

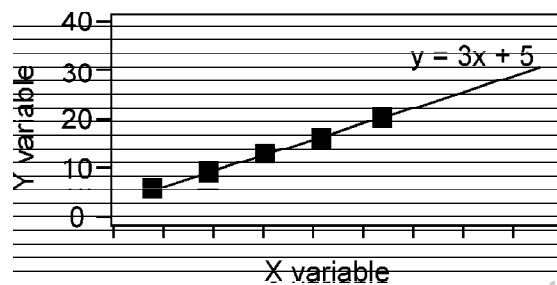


Fig.: Linear Correlation

### D) Non Linear (Curvilinear) Correlation

If corresponding to a unit change in one variable, the other variable does not change in a constant rate, but change at varying rates, then the relationship between two variables is said to be nonlinear or curvilinear as shown in fig. below. In this case, if the data are plotted on the graph, we do not get a straight line curve.

Mathematically, the correlation is non-linear if the slope of the plotted curve is not constant. Data relating to Economics, Social Science and Business Management do exhibit often non-linear relationship. We confine ourselves to linear correlation only.

### Example

X	-3	-2	-1	0	1	2	3
Y	9	4	1	0	1	4	9

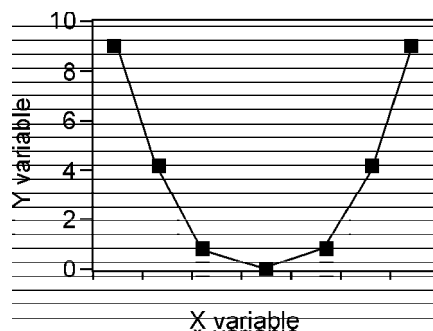


Fig.: Non Linear Correlation

### 5.1.2 limits for coefficient of correlation

**Q6. State the Limits for Coefficient of Correlation.**

*Ans :*

#### Limits for Co-efficient of Correlation for (x, y)

The value of the coefficient of correlation should lie between +1 and -1. If  $r = +1$ , the correlation is perfect and positive and if  $r = -1$  the correlation is perfect and negative. When  $r = 0$ , it means that there is no relationship between the two variables.

$$\text{Hence, } -1 < r(x, y) \leq 1$$

#### Note

The correlation coefficient describes not only the magnitude of correlation but also its direction. Thus, +1 would mean that correlation is positive and the magnitude of correlation is 1.

Similarly -1 means correlation is negative and the magnitude of correlation is again 1.

Degree of Correlation	The Value of $r(x,y)$ (Positive and Negative)
1. Perfect correlation	Exactly 1
2. Very high degree of correlation	0.90 and above but less than 1
3. Fairly high degree of correlation	0.75 and above but less than 0.90
4. Moderate degree of correlation	0.50 and above but less than 0.75
5. Low degree of correlation	0.25 and above but less than 0.50
6. Very high degree of correlation	Below 0.25
7. Absence of correlation	Equal to 0

### 5.2 METHODS OF CORRELATION

**Q7. What are the various methods of Correlation.**

*Ans :*

The various method of studying linear correlation can be shown diagrammatically as follows:

- (i) Scatter Diagram
- (ii) Two way frequency table
- (iii) Spearman's Rank Correlation Method
- (iv) Method of Concurrent Deviations
- (v) Method of Least Squares.

### 5.2.1 Scatter diagram

**Q8. What is Scatter Diagram? Explain the procedure of Scatter Diagram.**

*Ans :*

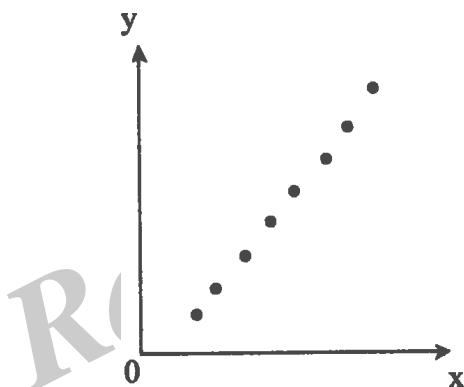
Scatter diagram method is the simplest way of diagrammatic representation of a bivariate distribution and helps in ascertaining the correlation between the two variables under study i.e., it portrays the relationship between these two variables graphically.

#### Procedure of Scatter Diagram

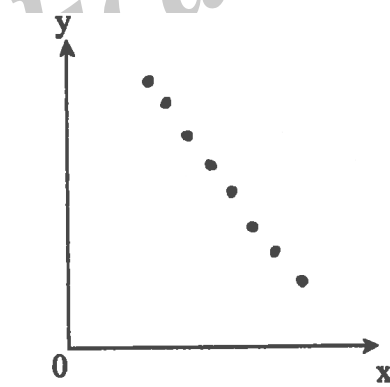
Given pair of values  $(x_1, y_1)$   $(x_2, y_2)$ ...  $(x_n, y_n)$  of two variables X and Y. Take the independent variables on the X axis and the dependent variable on the Y-axis. The V points denoted by the pair of values are plotted on the graph. The diagram of dots thus obtained is the scatter diagram. Regarding the correlation between the two variables, the scatter diagram can be interpreted as follows,

- (i) If the points reveal any upward or downward trend, the variables are said to be correlated, otherwise uncorrelated.
- (ii) If the points are very close to each other, a good amount of correlation exists, else poor correlation exists.
- (iii) Upward trend indicates positive correlation and downward trend indicates negative correlation.

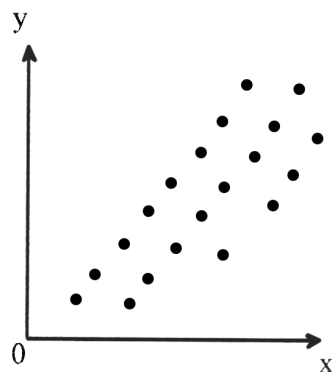
#### Different Forms of Scatter Diagram



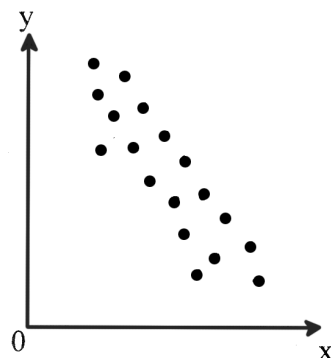
Perfect Positive Correlation



Perfect Negative Correlation

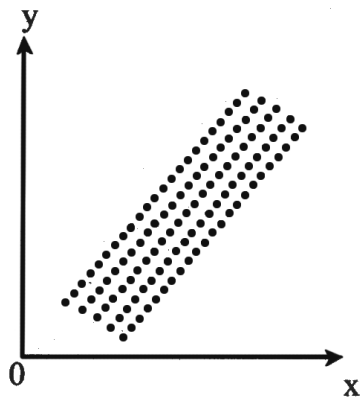
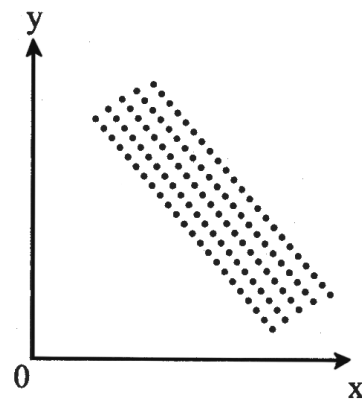
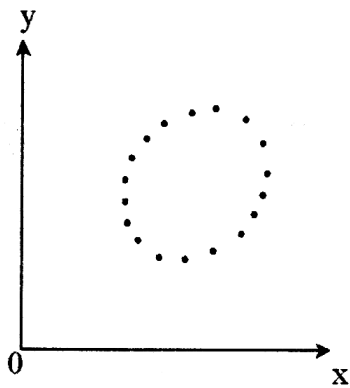
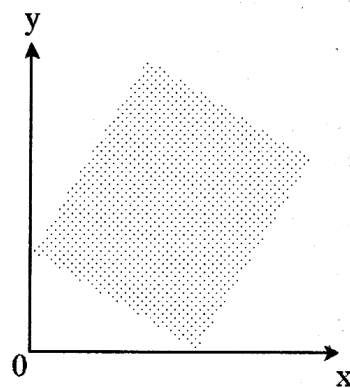
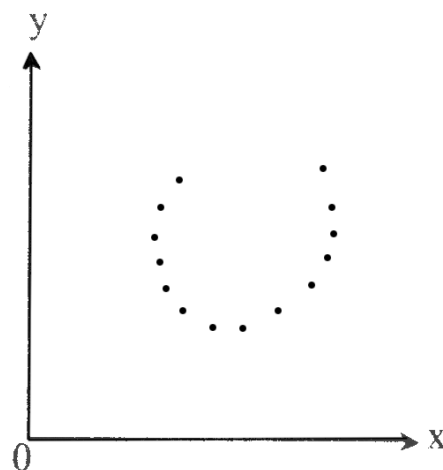


Low Degree Positive Correlation



Low Degree Negative Correlation



**High Degree Positive Correlation****High Degree negative Correlation****No Correlation****No Correlation****No Correlation**

**Q9. State the Merits and Demerits of Scatter Diagram.***Ans :***Merits**

- Scatter diagram is a simple and attractive method of finding out the nature of correlation between two variables.
- It is a non-mathematical method of studying correlation. It is easy to understand.
- We can get a rough idea at a glance whether it is a positive or negative correlation.
- It is not influenced by extreme items.
- It is a first step in finding out the relationship between two variables.

**Demerits**

- The major limitation of the method is that it only gives a visual picture of the relationship of two variables. It only tells us whether there is correlation between the variables, and if so, then in which direction, positive (or) negative.
- It does not give an idea about the precise degree of relationship as it is not amenable to mathematical treatment.

**5.2.2 Karl Pearson's coefficient of correlation****Q10. What is Karl Pearson's Coefficient of Correlation? Explain properties of Coefficient of Correlation.***Ans :***(Imp.)**

Karl Pearson's Coefficient of Correlation is arrived at with the help of a statistical formula that takes into account the mean and standard deviation of the two variables, the number of such observations and the covariance between them. Since Karl Pearson's coefficient of correlation is a number, it can describe the strength of the correlation in greater detail and more objectively. A value of  $-1$  signifies "absolute" negative correlation, a value between  $-1$  and  $-0.5$  signifies strong negative correlation, a value between  $-0.5$  and  $-0.25$  signifies moderate negative correlation and a value between  $-0.25$  and  $0$  signifies weak negative

correlation. Similarly, a value of  $+1$  signifies "absolute" positive correlation, a value between  $+1$  and  $+0.5$  signifies strong positive correlation, a value between  $+0.5$  and  $+0.25$  signifies moderate positive correlation and a value between  $+0.25$  and  $0$  signifies weak positive correlation.

**Properties of Karl Pearson's Coefficient of Correlation**

1. It is based on Arithmetic Mean and Standard Deviation.
2. It lies between  $-1$
3. It measures both direction as well as degree of change. If  $r$  is less than  $0$ , there is negative correlation, which means the direction of change of the two variables will be opposite. If  $r$  is more than  $0$ , there is positive correlation, which means the direction of change of the two variables will be same. Higher the value of  $r$ , greater is the degree of correlation. Hence, Karl Pearson's coefficient of correlation is said to be the ideal measure of correlation.
4. It is independent of change in scale. In other words, if a constant amount is added/subtracted from all values of a variable, the value of  $r$  does not change.
5. It is independent of change in origin. Thus, if a constant amount is multiplied with or divides all values of a variable,  $r$  does not change.
6. It is independent of direction. In other words, Correlation of  $X$  and  $Y$  is same as Correlation of  $Y$  and  $X$ .
7. It is a pure number without any units. In other words, it is independent of the unit of measurement of the 2 variables.
8. It takes into account all items of the variable(s).
9. It does not prove causation but is simply a measure of covariation.
10. Correlation coefficient of two variables  $X$  and  $Y$  is the Geometric Mean of two regression coefficients, regression coefficient of  $X$  on  $Y$  and regression coefficient of  $Y$  on  $X$ . Symbolically,

$$r = \text{Square root of } (b_{xy} * b_{yx})$$

11. Correlation coefficient can be calculated between two unrelated variables and such a number can be misleading. Such correlation is called accidental correlation, spurious correlation or non sense correlation.

**Q11. Explain Merits and Demerits of Coefficient of Correlation.**

*Ans :*

**Merits**

1. It takes into account all items of the variable(s).
2. It is a numerical measure and hence more objective.
3. It measures both direction as well as degree of change.
4. It facilitates comparisons between two series.
5. It is capable of further Algebraic treatment
6. It is more practical and hence popular and is more commonly used.

**Demerits**

1. It is not easy to calculate as complex formulae are involved.
2. It is more time consuming compared to methods such as rank correlation
3. It assumes a linear relationship between the two variables which may not be correct
4. It is impacted by extreme values as it is based on mean and standard deviation.
5. It is not easy to interpret.

**Q12. Explain the various methods of Coefficient of Correlation.**

*Ans :*

**i) Direct Method when deviations are taken from actual mean**

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

Where

$$x = X - \bar{X} \text{ and } y = Y - \bar{Y}$$

**Steps :**

1. Find the means of the two series ( $\bar{X}$  ,  $\bar{Y}$  )
2. Take the deviations of X series from the mean of X and denote these deviations as x.
3. Square these deviations and obtain the total. Denote it as  $\sum x^2$ .
4. Take the deviations of Y series from the Mean of Y and denote these deviations as y.
5. Square these deviations, obtain the total and denote it as  $\sum y^2$ .

6. Multiply the deviations of X and Y series, obtain the total and denote it  $\Sigma xy$ .
7. Substitute the above values in the formula.

## ii) Short-Cut Method

**When deviations are taken from assumed mean.**

$$r = \frac{N\Sigma xy - \Sigma x\Sigma y}{\sqrt{N\Sigma x^2 - (\Sigma x)^2} \sqrt{N\Sigma y^2 - (\Sigma y)^2}}$$

Where,  $x = X - A_1$

$y = Y - A_2$

$A_1$  = Assumed mean for X series

$A_2$  = Assumed mean for Y series

The values of coefficient of correlation as obtained by above formulae will always lie between  $\pm 1$ . When there is perfect positive correlation its value is +1 and when there is perfect negative correlation, its value is -1. When  $r = 0$  means that there is no relationship between the two variables. We normally get values which lie between +1 and -1.

### Q13. Explain briefly about probable error.

*Ans :*

The probable error of the coefficient of correlation helps in interpretation. The probable error of the coefficient of correlation is obtained as follows:

$$\text{P.E. of } r = 0.6745 \frac{1-r^2}{\sqrt{N}}$$

Where

$r$  = Coefficient of correlation

$N$  = Number of pairs of observations.

If the probable error is added to and subtracted from the coefficient of correlation, it would give two such limits within which we can reasonably expect the value of coefficient of correlation to vary.

Symbolically  $P(\rho) = r \pm \text{P.E.}$

Where 'P' denotes the correlation in the population. Suppose, the Coefficient of correlation for a pair of 10 observations is 0.8 and its P.E. is 0.05. the limits of the correlation in the population would be  $r \pm \text{P.E.}$  i.e.  $0.8 \pm 0.05$  or  $0.75 - 0.85$ .

If the value of  $r$  is less than the probable error then  $r$  is not at all significant, i.e. there is no evidence of correlation. If the value of  $r$  is more than six times the probable error, it is significant. Hence it can be said that  $r$  is significant, when

$$r > 6 \text{ P.E. or } \frac{r}{\text{P.E.}} > 6.$$

**Q14. How do you interpret the coefficient of correlation?***Ans :*

The investigator must take utmost care while interpreting the value of 'r' and interpretation depends very much on experience.

**Rules of Interpretation**

1. When  $r = 0$ , it implies that the variables are uncorrelated.
2. When  $r = +1$ , it implies that there is a perfect positive relationship between the variables.
3. When  $r = -1$ , it implies that there is a perfect negative relationship between the variables.
4. The closer the value of 'r' to  $+1$  or  $-1$ , the closer the relationship between the variables.
5. The closeness of the relationship is not proportional to r.

**PROBLEMS ON KARL PEARSON'S COEFFICIENT OF CORRELATION AND****PROBABLE ERROR**

1. Find the coefficient of correlation from the following data :

X	46	54	56	56	58	60	62
Y	36	40	44	54	42	58	54

*Sol :***Calculation of Correlation**

X	(X - $\bar{X}$ ) (x)	Y	(Y - $\bar{Y}$ ) (y)	$x^2$	$y^2$	xy
46	-10	36	-11	100	121	110
54	-2	40	-7	4	49	14
56	0	44	-3	0	9	0
56	0	54	7	0	49	0
58	+2	42	-5	4	25	-10
60	+4	58	11	16	121	44
62	+6	54	7	36	49	42
				160	423	200

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

$$x = (x - \bar{x})$$

$$\bar{X} = \frac{392}{7} = 56,$$

Where,

$$x^2 = 160, y^2 = 1423, xy = 200$$

$$= \frac{200}{\sqrt{160 \times 423}}$$

$$y = (y - \bar{y})$$

$$\bar{Y} = \frac{328}{7} = 46.85 = 47$$

$$= \frac{200}{260.15} = 0.768$$

2. Find Karl Pearson's Coefficient of Correlation for the following data:

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

Sol :

Calculations of Karl pearson's coefficient of correlation

X	$X - \bar{X} = x$	$x^2$	Y	$Y - \bar{Y} = y$	$y^2$	xy
65	-3	9	67	-2	4	6
66	-2	4	68	-1	1	2
67	-1	1	65	-4	16	4
67	-1	1	68	-1	1	1
68	0	0	72	3	9	0
69	1	1	72	3	9	3
70	2	4	69	0	0	0
72	4	16	71	2	4	8
		$\Sigma x^2 = 36$			$\Sigma y^2 = 44$	$\Sigma xy = 24$

$$\bar{x} = \frac{\Sigma x}{N} = \frac{544}{8} = 68$$

$$\bar{y} = \frac{\Sigma y}{N} = \frac{552}{8} = 69$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{24}{\sqrt{36 \times 44}}$$

$$= \frac{24}{\sqrt{1584}} = \frac{24}{39.799}$$

$$r = 0.6031.$$

3. Calculate Karl Pearson's Coefficient of Correlation for the following data.

X	7	6	5	4	3	2	1
Y	18	16	14	12	10	6	8

Sol :

Calculations of Karl pearson's coefficient of correlation

X	Y	X - $\bar{X}$ (x)	Y - $\bar{Y}$ (y)	xy	x <sup>2</sup>	y <sup>2</sup>
7	18	3	6	18	9	36
6	16	2	4	8	4	16
5	14	1	2	2	1	4
4	12	0	0	0	0	0
3	10	-1	-2	+ 2	1	4
2	6	-2	-6	12	4	36
1	8	-3	-4	12	9	16
				<b>54</b>	<b>28</b>	<b>112</b>

$$\bar{X} = \frac{28}{7} = 4$$

$$\bar{Y} = \frac{84}{7} = 12$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$r = \frac{54}{\sqrt{28 \times 112}}$$

$$= \frac{54}{\sqrt{3136}} = \frac{54}{56} = 0.96$$

4. If the Coefficient of Correlation between two variables X and Y is 0.86, the Covariance is 36. And the Standard Deviation of X is 4, find the Standard Deviation of Y.

*Sol:*

Given

$$\sigma_x = 4$$

$$\begin{aligned}\text{Covariance} &= \frac{\sum xy}{N} \\ &= 36\end{aligned}$$

$$\begin{aligned}\text{Here } r &= \frac{\sum xy}{N\sigma_x\sigma_y} \\ &= \frac{\sum xy}{N} \times \frac{1}{\sigma_x \cdot \sigma_y}\end{aligned}$$

$$r = 36 \times \frac{1}{4\sigma_y}$$

$$\text{Here } r = 0.86$$

$$0.86 = \frac{36}{4 \cdot \sigma_y}$$

$$\sigma_y = \frac{36 \cdot 4}{4 \times 0.86}$$

$$\sigma_y = \frac{9}{0.86}$$

$$= 10.46.$$

5. Calculate coefficient of correlation from the following data:

X	100	200	300	400	500	600	700
Y	0.3	0.5	0.6	0.8	1.0	1.1	1.3

*Sol:*

(Dec.-20)

$$\text{Average of X} = \bar{X} = \frac{2800}{7} = 400$$

$$\text{Average of Y} = \bar{Y} = \frac{5.6}{7} = 0.8$$



Karl Pearson's correlation coefficient / coefficient of correlation

$$r = \frac{N\sum UV - \sum U \sum V}{\sqrt{N\sum U^2 - (\sum U)^2} \sqrt{N\sum V^2 - (\sum V)^2}}$$

X	Y	U = X - 400	V = Y - 0.8	U <sup>2</sup>	V <sup>2</sup>	UV
100	0.3	- 300	- 0.5	90,000	0.25	150
200	0.5	- 200	- 0.3	40,000	0.09	60
300	0.6	- 100	- 0.2	10,000	0.04	20
400	0.8	0	0	0	0	0
500	1.0	100	0.2	10,000	0.04	20
600	1.1	200	0.3	40,000	0.09	60
700	1.3	300	0.5	90,000	0.25	150
$\Sigma X = 2800$	$\Sigma Y = 5.6$	$\Sigma U = 0$	$\Sigma V = 0$	$\Sigma U^2 = 2,80,000$	$\Sigma V^2 = 0.76$	$\Sigma UV = 460$

$$r = \frac{N\sum UV - \sum U \sum V}{\sqrt{N\sum U^2 - (\sum U)^2} \sqrt{N\sum V^2 - (\sum V)^2}}$$

$$r = \frac{(7 \times 460) - (0 \times 0)}{\sqrt{(7 \times 2,80,000) - (0)^2} \sqrt{(7 \times 0.76) - (0)^2}}$$

$$= \frac{3220}{\sqrt{1960000} \sqrt{5.32}}$$

$$= \frac{3220}{3229} = 0.997$$

$\therefore$  Karl Pearson's coefficient of correlation = 0.997

### Conclusion

The is a positive correlation between the variables X and Y.

### 6. Find Karl Pearson's coefficient correlation to the following:

X	40	45	53	55	38	42	45	62
Y	71	78	87	73	74	71	76	75

*Sol:*

(July-18, Imp.)

$$\text{Average of } \bar{X} = \frac{380}{8} = 48$$

$$\text{Average of } \bar{Y} = \frac{605}{8} = 76$$

Calculation of Karl Pearson's coefficient correlation

x	y	x - $\bar{x}$ (x)	y - $\bar{y}$ (y)	x <sup>2</sup>	y <sup>2</sup>	xy
40	71	-8	-5	64	25	40
45	78	-3	2	9	4	-6
53	87	5	11	25	121	55
55	73	7	-3	49	9	-21
38	74	-10	-2	100	4	20
42	71	-6	-5	36	25	30
45	76	-3	0	9	0	0
62	75	-14	1	196	1	-14
				488	189	104

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{104}{\sqrt{488 \times 189}} = \frac{104}{\sqrt{92,232}}$$

$$\frac{104}{303.69} = 0.34$$

7. Find Karl Pearson's coefficient correlation to the following :

X	48	39	65	80	73	60	52	120
Y	10	50	120	225	90	60	55	25

Sol :

(Aug.-17)

x	y	x - $\bar{x}$ (x)	y - $\bar{y}$ (y)	x <sup>2</sup>	y <sup>2</sup>	xy
48	10	-19	-69	361	4761	1311
39	50	-28	-29	784	841	812
65	120	-2	41	4	1681	-82
80	225	13	146	169	21,316	1898
73	90	6	11	36	121	66
60	60	-7	-19	49	361	133
52	55	-15	-24	225	576	360
120	25	53	-54	2809	2916	-2862
537	635			4437	32,573	1636

$$\bar{x} = \frac{537}{8} = 67.125 = 67$$

$$\bar{y} = \frac{635}{8} = 79.375 = 79$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{1636}{\sqrt{4437 \times 32,573}}$$

$$\frac{1636}{144526401} = 0.136$$

8. Following figures gives the rain fall in inches and production in '00 mds for Rabi crops for number of years. Find the coefficient of correlation between rainfall and production.

Rainfall	20	22	24	26	28	30	32
Production	15	18	20	32	40	39	40

Sol.:

(Nov.-20)

Let the rainfall be X and the production be Y

**Averages of X and Y**

$$\bar{X} = \frac{\Sigma X}{N} = \frac{182}{7} = 26 \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{204}{7} = 29.14$$

As  $\bar{Y}$  is in fraction, it is advisable to take deviation from arbitrary point (i.e., assumed mean). Let the assumed mean of Y be 29.

Karl Pearson's correlation coefficient

$$r = \frac{N\Sigma UV - \Sigma U\Sigma V}{\sqrt{N\Sigma U^2 - (\Sigma U)^2} \sqrt{N\Sigma V^2 - (\Sigma V)^2}}$$

X	Y	U = X - 26	V = Y - 29	U <sup>2</sup>	V <sup>2</sup>	UV
20	15	-6	-14	36	196	84
22	18	-4	-11	16	121	44
24	20	-2	-9	4	81	18
26	32	0	3	0	9	0
28	40	2	11	4	121	22
30	39	4	10	16	100	40
32	40	6	11	36	121	66
<b><math>\Sigma X = 182</math></b>	<b><math>\Sigma Y = 204</math></b>	<b><math>\Sigma U = 0</math></b>	<b><math>\Sigma V = 1</math></b>	<b><math>\Sigma U^2 = 112</math></b>	<b><math>\Sigma V^2 = 749</math></b>	<b><math>\Sigma UV = 274</math></b>

$$\begin{aligned}
 r &= \frac{7 \times 274 - (0) \times (1)}{\sqrt{(7 \times 112) - (0)^2} \sqrt{(7 \times 749) - (1)^2}} \\
 &= \frac{1918 - 0}{\sqrt{784 - 0} \sqrt{5243 - 1}} \\
 &= \frac{1918}{28 \times 72.40} \\
 &= \frac{1918}{2027.2} = 0.946
 \end{aligned}$$

Karl Pearson's correlation coefficient = 0.95

It indicates that there is very high degree of positive correlation between rainfall and production.

### 5.3 SPEARMAN'S RANKCORRELATION

**Q15. What is Rank correlation ? State the properties rank correlation.**

*Ans :*

(Nov.-20)

The Karl Pearson's method is based on the assumption that the population being studied is normally distributed. When it is known that the population is not normal or when the shape of the distribution is not known, there is need for a measure of correlation that involves no assumption about the parameter of the population.

It is possible to avoid making any assumptions about the populations being studied by ranking the observations according to size and basing the calculations on the ranks rather than upon the original observations. It does not matter which way the items are ranked, item number one may be the largest or it may be the smallest. Using ranks rather than actual observations gives the coefficient of rank correlation.

This method of finding out covariability or the lack of it between two variables was developed by the British Psychologist Charles Edward Spearman in 1904.

#### Properties

1. It is based on subjective ranking of variables.
2. It lies between -1

3. It measures both direction as well as degree of change. If  $r$  is less than 0, there is negative correlation, which means the direction of change of the two variables will be opposite. If  $r$  is more than 0, there is positive correlation, which means the direction of change of the two variables will be same. Higher the value of  $r$ , greater is the degree of correlation.
4. It is independent of direction. In other words, Correlation of X and Y is same as Correlation of Y and X.
5. It is a pure number without any units. In other words, it is independent of the unit of measurement of the 2 variables.
6. It is not impacted by extreme values as only ranking matters.

**Q16. State the Merits and Demerits of rank correlation.**

*Ans :*

#### Merits

1. It is easy to understand and calculate
2. It is not impacted by extreme values.
3. It is a numerical measure and provides objectivity to subjective ranking.
4. It is the only method of finding correlation with respect to qualitative factors such as honesty, beauty, etc.
5. It measures both direction as well as degree of change.
6. It facilitates comparisons between two series.
7. It can be applied to irregular data also.
8. It is ideal when the number of observations is very small.

#### Demerits

1. It cannot be applied to grouped data
2. It lacks the precision of Karl Pearson's Coefficient of Correlation.
3. All the information concerning the variable is not used.
4. The Computation becomes complicated as the number of observations increase.

**Q17. Explain how correlation is calculated through spearman's rank correlation method.**

*Ans :*

### Computation

Spearman's correlation is computed by using the following formula:

$$r_s = 1 - \frac{6\sum D^2}{N^3 - N}$$

Where

$r_s$  = rank coefficient of correlation.

D = Difference of rank between paired items in two series.

N = No. of Pairs of observation.

Spearman's rank correlation coefficient lies between +1 and -1.

Under this method two types of problems can be given. (i) Where ranks are given (ii) Where ranks are not given.

#### i) When ranks are given

##### Steps:

- Take the difference of the two ranks and denote the difference as D.
- Square these differences and obtain the total. Denote it as  $D^2$  and apply the formula.

#### ii) When ranks are not given

When the actual data is given and ranks are not given, we have to assign the ranks. Ranks can be assigned by taking either the higher value as 1 or the lowest value as 1. The same should be followed for both the variables. Once the ranks are assigned, the rest of the working remains the same.

##### ➤ Equal Ranks

Where there is more than one item with the same value a common rank is given to such items. If

two individual are ranked equal at fourth place, they are each given the rank  $\frac{4+5}{2} = 4.5$ . In the same

manner if three items are ranked at 6th place, they are given  $\frac{6+7+8}{3} = 7$  rank.

When equal ranks are given to some items, an adjustment is made in the formula by adding  $\frac{1}{12}$  ( $m^3-m$ ) to the value of  $\sum D^2$  where m stands for the number of items whose ranks are equal. If there are more than one such group of items with common ranks, this value is added as many times the number of such groups. It may be written as.

$$r_s = 1 - \frac{6 \left[ \sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) \dots \right]}{N^3 - N}$$

**PROBLEMS ON RANK CORRELATION**

9. Calculate the coefficient of correlation from the following data by the spearman's Rank Differences method :

Prices of Tea (₹)	75	88	95	70	60	80	81	50
Prices of Coffee(₹)	120	134	150	115	110	140	142	100

Sol :

Price of Tea (X)	R <sub>1</sub>	Price of Coffee (Y)	R <sub>2</sub>	R <sub>1</sub> -R <sub>2</sub> =d	d <sup>2</sup>
75	4	120	4	0	0
88	7	134	5	2	4
95	8	150	8	0	0
70	3	115	3	0	0
60	2	110	2	0	0
80	5	140	6	-1	1
81	6	142	7	-1	1
50	1	100	1	0	0
					6

$$r_s = 1 - \frac{6 \sum (d)^2}{(N)^3 - N} = 1 - \frac{6 \times 6}{(8)^3 - 8} = 1 - \frac{36}{504}$$

$$r_s = 1 - 0.071 = 0.238$$

10. Compute Rank coefficient of correlation for the following data :

X	15	20	28	12	40	60	20	80
Y	40	30	50	30	20	10	30	60

Sol :

X	R <sub>1</sub>	Y	R <sub>2</sub>	R <sub>1</sub> - R <sub>2</sub> = d	d <sup>2</sup>
15	2	40	6	-4	16
20	3.5	30	4	-0.5	0.25
28	5	50	7	-2	4
12	1	30	4	-3	9
40	6	20	2	4	16
60	7	10	1	6	36
20	3.5	30	4	-0.5	0.25
80	8	60	8	0	0
					81.5

$$r_s = 1 - \frac{6 \left[ \sum d^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) \right]}{(N)^3 - N}$$

$$r_s = 1 - \frac{6 \left[ 81.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3) \right]}{(8)^3 - 8}$$

$$r_s = 1 - \frac{6[81.5 + 0.5 + 2]}{512 - 8}$$

$$r_s = 1 - \frac{6(84)}{504}$$

$$r_s = 1 - \frac{504}{504}$$

$$r_s = 1 - 1$$

$$= 0$$

11. Calculate Rank correlation co-efficient from the following data :

	A	B	C	D	E
x :	1	2	3	5	4
f :	2	1	4	3	5

Sol :

Rank correlation co-efficient

X	R <sub>1</sub>	Y	R <sub>2</sub>	(R <sub>1</sub> - R <sub>2</sub> ) 'D'	D <sup>2</sup>
1	5	2	4	1	1
2	4	1	5	-1	1
3	3	4	2	1	1
5	1	3	3	-2	4
4	2	5	1	1	1
					8

$$1 - \frac{6 \sum D^2}{N^3 - N}$$

$$1 - \frac{6 \times 8}{5^3 - 5}$$

$$1 - \frac{48}{125 - 5}$$

$$1 - \frac{48}{120}$$

$$1 - 0.4 = 0.6$$

12. From the following data calculate rank correlation

x	20	25	60	45	80	25	55	65	25	75
f	45	50	55	50	60	70	72	78	80	63

Sol.:

X	R <sub>1</sub>	Y	R <sub>2</sub>	R <sub>1</sub> - R <sub>2</sub> (D)	D <sup>2</sup>
20	10	45	10	0	0
25	8	50	8.5	- 0.5	0.25
60	4	55	7	- 3	9
45	6	50	8.5	- 2.5	6.25
80	1	60	6	- 5	25
25	8	70	4	4	16
55	5	72	3	2	4
65	3	78	2	1	1
25	8	80	1	7	49
75	2	63	5	- 3	9
					ΣD <sup>2</sup> = 119.5

$$1 - \frac{6 \left( \Sigma D^2 + \frac{1}{12} (m^3 - m) + \frac{1}{12} (n^3 - n) \right)}{N^3 - N}$$

$$1 - \frac{6 \left\{ 119.5 + \frac{1}{12} (3^3 - 3) + \frac{1}{12} (2^3 - 2) \right\}}{10^3 - 10}$$

$$\frac{1 - 6 \{ 119.5 + 2 + 0.5 \}}{990}$$

$$1 - \frac{6(122)}{990}$$

$$1 - \frac{732}{990}$$

$$1 - 0.73 = 0.27$$



13. From the ranks of 10 student's in Accountancy and Statistics given below,  
Calculate Rank Correlation Coefficient

Ranks in Accountancy	1	2	3	4	5	6	7	8	9	10
Rank in Statistics	1	3	5	6	7	4	8	10	9	2

Sol.:

$R_1$	$R_2$	$R_1 - R_2 (D)$	$D^2$
1	1	0	0
2	3	-1	1
3	5	-2	4
4	6	-2	4
5	7	-2	4
6	4	2	4
7	8	-1	1
8	10	-2	4
9	9	0	0
10	2	8	16
			38

$$r_k = 1 - \frac{6\sum d^2}{N^3 - N} = 1 - \frac{6(38)}{10^3 - 10} = 1 - \frac{228}{1000 - 10} = 1 - \frac{228}{990}$$

$$= 1 - 0.230 = 0.7696$$

#### 5.4 CONCEPT OF MULTIPLE AND PARTIAL CORRELATION

Q18. Define the following terms :

- (i) Multiple Correlation
- (ii) Partial Correlation

Ans.:

- 1) **Multiple Correlations** : In multiple correlations we study together the relationship between three or more factors like production, rainfall and use of fertilizers.
- 2) **Partial Correlation** : In partial correlation though more than two factors are involved but correlation is studied only between two factors and the other factors are assumed to be constant.

**For example**, If out of the three related variables, say, marks in statistics, marks in accountancy, and marks in English, we study the correlation between the two variables, viz., marks in statistics, and marks in Accountancy ignoring the effect of the other variable, i.e., marks in English, it will be a case of partial correlation. On the other hand, when the relationship between any two or more variables is studied at a time, it is a case of multiple correlation. If the relationship between the volume of profits, volume of sales, and the volume of cost of sales at a time are studied it will be a case of multiple correlation. In actual practice, however, the study of multiple correlation is not popular.

### 5.5 REGRESSION ANALYSIS

#### 5.5.1 Concept

**Q19. What do you understand by Regression?**

(OR)

**What is meant by Regression?**

(OR)

**Define Regression?**

*Ans :*

Regression analysis which confines itself to a study of only two variables is called simple regression. The regression analysis which studies more than two variables at a time is called multiple regression. In the simple regression analysis there are two variables-one of which is known as 'independent variable' or 'regressor' or 'predictor'. On the basis of the values of this variable the values of the other variable are predicted. The other variable whose values are predicted is called the 'dependent' or 'regressed' variable.

#### Definitions

1. "Regression is the measure of the average relationship between two or more variables in terms of the original units of the data."
2. **According to Morris Hamburg** The term 'regression analysis' refers to the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process."
3. **According to Taro Yamane** "One of the most frequently used techniques in economics and business research, to find a relation between two or more variables that are related causally, is regression analysis."
4. **According to YaLum Chou** "Regression analysis attempts to establish the 'nature of the relationship between variables that is, to study the functional relationship between the variables and thereby provide a mechanism for prediction, or forecasting."

**Q20. What is the importance of regression analysis ?**

*Ans :*

(June-19)

1. Regression analysis helps in establishing a functional relationship between two or more variables. Once this is established it can be used for various advanced analytical purposes.
2. Since most of the problems of economic analysis are based on cause and effect relationship, the regression analysis is a highly valuable tool in economics and business research.
3. This can be used for prediction or estimation of future production, prices, sales, investments, income, profits and population which are indispensable for efficient planning of an economy and are of paramount importance to a businessman or an economist.
4. Regression analysis is widely used in statistical estimation of demand curves, supply curves, production functions, cost functions, consumption functions, etc. Economists have discovered many types of production functions by fitting regression lines to input and output data.

**Q21. What are the objectives of Regression Analysis.**

*Ans :*

1. The first objective of regression analysis is to provide estimates of values of the dependent variable from values of independent variable. This is done with the help of the regression line. The regression line describes the average relationship existing between X and Y variables, more precisely, it is a line which displays mean values of Y for given values of X.
2. The second objective of regression analysis is to obtain a measure of the error involved in using the regression line as a basis for estimation. For this purpose standard error of estimate is obtained. This helps in understanding the correlation existing between X and Y.

3. In general, we can model the expected value of  $y$  as an  $n^{\text{th}}$  order polynomial, yielding the general polynomial regression model

$$Y = a_0 + a_1x + a_2x^2 + a_3x^3 + \dots + a_nx^n + \varepsilon$$

Conveniently, these models are all linear from the point of view of estimation, since the regression function is linear in terms of the unknown parameters  $a_0, a_1, \dots$ . Therefore, for least squares analysis, the computational and inferential problems of polynomial regression can be completely addressed using the techniques of multiple regressions. This is done by treating  $x, x^2, \dots$  as being distinct independent variables in a multiple regression model.

## Q22. Explain different types of Regression.

*Ans :*

The various types of Regression are as follows:

### 1. Simple Regression

In statistics, simple regression is the least squares estimator of a linear regression model with a single predictor variable. In other words, simple linear regression fits a straight line through the set of  $n$  points in such a way that makes the sum of squared residuals of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible.

### 2. Multiple Regression

Multiple regression analysis represents a logical extension of two-variable regression analysis. Instead of a single independent variable, two or more independent variables are used to estimate the values of a dependent variable. However, the fundamental concepts in the analysis remain the same.

### 3. Linear Regression

Linear regression is a form of regression which is used for modeling the relationship between scalar variables like  $X$  and  $F$  under linear regression, linear functions are used to model the data and the unknown parameters, of models are estimated from the data. Hence, these models are known as linear models.

## 4. Non-linear regression

In the non-linear regression the explained variable (dependent variable) changes at varying rate with a given change in the explaining variable (independent variable). It is also known as Curvilinear regression. Under non linear regression, the observational data are modeled by a function i.e., a non linear blend of model parameters and depends on one or more independent variable.

## Q23. What are the limitations of regression analysis ?

*Ans :*

1. It assumes a linear relationship between two variables which need not be the case always.
2. It assumes a static relationship between the two variables over a period of time. However, relationships between variables can change with a change in other factors. For example, the change in demand for a given change in price can be estimated using regression. However, the impact of price on demand will be different when a family or a nation is poor and when such a family or nation has abundance of wealth or resources.
3. Regression analysis provides meaningful insights only up to a certain limit. For example, increasing production results in a decrease in marginal cost. However, beyond a certain point, increase in production can result in the costs going up.

## Q24. What do you mean by Linear and Non Linear Regression?

*Ans :*

(Dec.-20)

### (i) Linear Regression

Linear regression is a form of regression which is used for modeling the relationship between scalar variables like  $X$  and  $F$  under linear regression, linear functions are used to model the data and the unknown parameters, of models are estimated from the data. Hence, these models are known as linear models.

Linear models more commonly refers to those models, where the conditional mean of variable 'F' for a given value of variable X will be an affine function of X. A linear regression may also refer to a model, where median or other quantile of the conditional distribution of 'F' for a given value of 'X' is termed as linear function of X. Similar, to all types of regression analysis, linear regression also aims on the conditional probability distribution of 'F' for a given 'X', instead of joint probability distribution of 'F' and X.

## (ii) Non-Linear Regression

In the non-linear regression the explained variable (dependent variable) changes at varying rate with a given change in the explaining variable (independent variable). It is also known as Curvilinear regression. Under non linear regression, the observational data are modeled by a function i.e., a non linear blend of model parameters and depends on one or more independent variable.

Method of successive approximations are used for fitting the data. The data in non linear regression contains of error free independent variable 'X' and its relatively observed dependent variable 'Y'.

**Example :** The output of rice increases rapidly with the application of the initial dose of fertilizer; there after it increases at a falling rate. The relationship in such case, when shown on graph will yield a 'curve'.

**Q25. Differentiate between linear and non-linear regression.**

(OR)

**Compare and contrast linear and non-linear regression.**

*Ans :*

The differences between linear and non-linear regression are as follows,

S.No.	Basis	Linear Regression	Non-Linear Regression
1.	Meaning	Linear regression is a form of regression which is used for modelling the relation ship between a scalar variable 'X' and 'Y'.	Non linear regression is a type of regression, where the observational data are modeled by a function i.e., a non linear blend of model parameters.
2.	Curve	If the regression curve is a straight line, then the regression is termed as linear regression.	If the curve of the regression is not a straight line. then the regression is termed as curved or non-linear regression.
3.	Model form	Under this, the parameters are considered as linear combinations.	Under this, the parameter are considered as functions
4.	Solution	Under linear regression, the solution for parameters is represented as closed form	Under non linear regression it is necessary for parameters to be solved repeatedly by using optimization algorithms.
5.	Uniqueness	The solution under linear regression is unique.	The Sum of the Squared Errors (SSE) may not be appear as unique.
6.	Parameters estimation	In case of uncorrelated errors, estimation Parameters are unbiased.	Incase of uncorrelated errors, estimation of Parameters are usually biased.
7.	Equation	The equation of regression curve is the equation of a straight line i.e., first degree equation in variables X and Y.	The regression equation will be functional relation between variables X and Y involving terms in x and Y involving terms in x and y of degree more than one.

**Q26. Differentiate between Correlation and Regression.****(OR)****What are the differences between Correlation and Regression.***Ans :***(July-18, Imp.)**

S.No.	Basis for Comparison	Correlation	Regression
1.	<b>Meaning</b>	Correlation is a statistical measure which determines co-relationship or association of two variables.	Regression describe how an independent variable is numerically related to the dependent variable.
2.	<b>Usage</b>	To represent linear relationship between two variables.	To fit a best line and estimate one variable on the basis of another variable.
3.	<b>Dependent and Independent variables</b>	No difference	Both variables are different
4.	<b>Indicates</b>	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (Y).
5.	<b>Objective</b>	To find a numerical value expressing the relationship between variables.	To estimate values of random variable on the basis of the values of fixed variable.

**5.5.2 Least square fit of a linear regression****Q27. Discuss about Least square fit of a linear regression***Ans :*

The regression equation of Y on X is expressed as follows:

$$Y = a + bX$$

It may be noted that in this equation 'Y' is a dependent variable, i.e., its value depends on X. 'X' is independent variable, i.e., we can take a given value of X and compute the value of Y.

'a' is "Y-intercept" because its value is the point at which the-regression line crosses the Y-axis, that is, the vertical axis, 'b' is the "slope" of line. It represents change in Y variable for a unit change in X variable.

'a' and 'b' in the equation are called numerical constants because for any given straight line, their value does not change.

If the values of the constants 'a' and 'b' are obtained, the line is completely determined. But the question is how to obtain these values. The answer is provided by the method of Least Squares which states that the line should be drawn through the plotted points in such a manner that the sum of the squares of the deviations of the actual Y values from the computed Y values is the least, or in other words, in order to obtain a line which fits the points best  $\sum(Y-Y_c)^2$ , should be minimum. Such a line is known as the line of 'best fit'.

A straight line fitted by least squares has the following characteristics:

- 1) It gives the best fit to the data in the sense that it makes the sum of the squared deviations from the line,  $\sum(Y - Y_c)^2$  smaller than they would be from any other straight line. This property accounts for the name 'Least Squares'.
- 2) The deviations above the line equal those below the line, on the average. This means that the total of the positive and negative deviations is zero, or  $\sum(Y - Y_c) = 0$ .
- 3) The straight line goes through the overall mean of the data  $(\bar{X}, \bar{Y})$ .
- 4) When the data represent a sample from a large population the least squares line is a 'best' estimate of the population regression line.

### 5.5.3 Two lines of regression

**Q28. What do you mean by line of regression? Derive the equations of lines of regression.**

*Ans :*

(Imp.)

In a bivariate distribution, if the variables are related then the points when plotted in the scatter diagram will lie near a straight line which is called the line of regression and the regression is said to be linear regression. If points lie on some non-linear curve then the regression is said to be curvilinear regression.

#### (I) Regression of Y on X.

The regression equation of Y on X is expressed as follows :

$$Y = a + bX$$

It may be noted that in this equation 'Y' is a dependent variable, i.e., its value depends on X. 'X' is independent variable, i.e., we can take a given value of X and compute the value of Y.

'a' is "Y-intercept" because its value is the point at which the regression line crosses the Y-axis, that is, the vertical axis, 'b' is the "slope" of line. It represents change in Y variable for a unit change in X variable.

'a' and 'b' in the equation are called numerical constants because for any given straight line, their value does not change.

If the values of the constants 'a' and 'b' are obtained, the line is completely determined. But the question is how to obtain these values. The answer is provided by the method of Least Squares which states that the line should be drawn through the plotted points in such a manner that the sum of the squares of the deviations of the actual Y values from the computed Y values is the least, or in other words, in order to obtain a line which fits the points best  $\sum(Y - Y_c)^2$ , should be minimum. Such a line is known as the line of 'best fit'.

A straight line fitted by least squares has the following characteristics :

- (i) It gives the best fit to the data in the sense that it makes the sum of the squared deviations from the line,  $\sum(Y - Y_c)^2$ , smaller than they would be from any other straight line. This property accounts for the name 'Least Squares'.
- (ii) The deviations above the line equal those below the line, on the average. This means that the total of the positive and negative deviations is zero, or  $\sum(Y - Y_c) = 0$ .
- (iii) The straight line goes through the overall mean of the data  $(\bar{X}, \bar{Y})$ .
- (iv) When the data represent a sample from a large population the least squares line is a 'best' estimate of the population regression line.

With a little algebra and differential calculus it can be shown that the following two equations, if solved simultaneously, will yield values of the parameters *a* and *b* such that the least squares requirement is fulfilled :

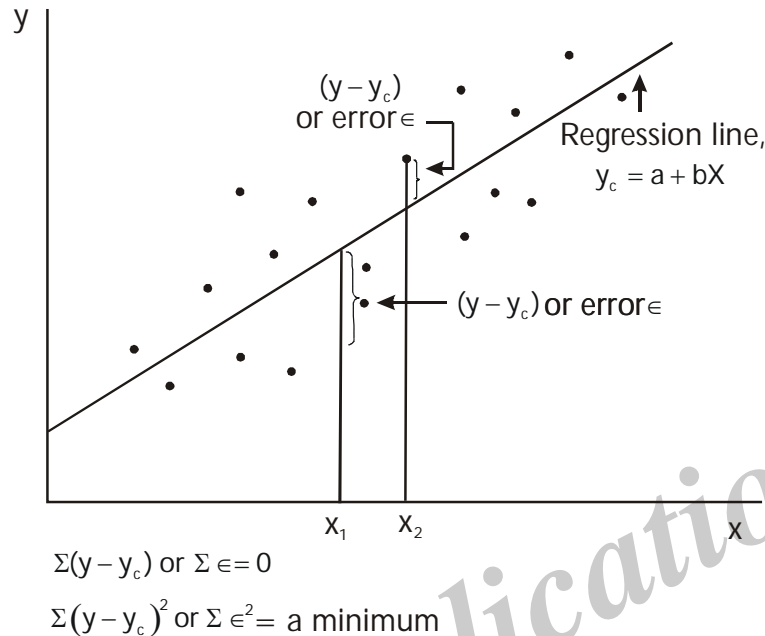
$$\sum Y = Na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

These equations are usually called the *normal equations*. In the equations  $\sum X$ ,  $\sum XY$ ,  $\sum X^2$  indicate totals which are computed from the

observed pairs of values of two variables  $X$  and  $Y$  to which the least squares estimating line is to be fitted and  $N$  is the number of observed pairs of values.

This will be shown in figure below.



**Figure : Regression of Y on X: Least Squares**

If the parameter estimates be  $a$  for  $\alpha$  and  $b$  for  $\beta$ , then the line would be

$$Y_c = a + bX$$

Since we seek to minimize  $\Sigma(Y - Y_c)^2$ , which works out to be  $\Sigma(Y - a - bX)^2$ , we can find the values of  $a$  and  $b$  by applying calculus. This results in a pair of what are called normal equations. The normal equations are :

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

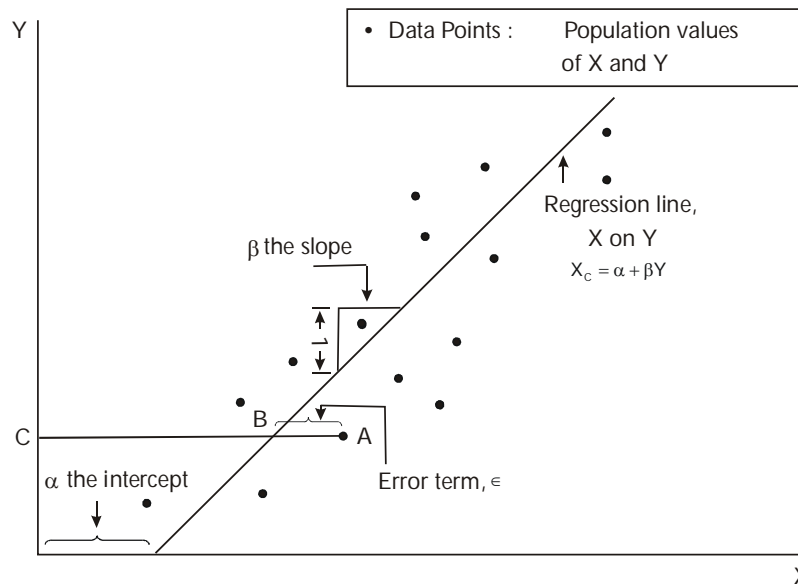
## (II) Regression of X on Y

In general, usually  $X$ -variable is taken to be independent and the  $Y$ -variable as dependent one. However, if the  $X$ -variable is treated as the dependent variable and  $Y$  as an independent variable, we can also have regression of  $X$  on  $Y$ . In the regression of  $X$  on  $Y$ , the population regression model is :

$$X = \alpha + \beta Y + \epsilon \text{ or } X_c = a + bY$$

in which  $X$  is the dependent variable (the variable to be predicted);  $Y$  is the independent variable (the predictor variable);  $\alpha$  is the population  $X$ -intercept;  $\beta$  is the population slope (measured as change in the  $X$  variable corresponding to a unit change in  $Y$ ); and  $\epsilon$  is the error term. Here the  $Y$ -variable is fixed and the randomness in the  $X$  variable comes from the error term,  $\epsilon$ .

The population regression model is shown in Figure below.



For a given pair of  $X$  and  $Y$  values, represented by a point, say  $A$ , the actual value of  $X$  equal to  $AC$  is composed of the non-random part  $BC$  (given by the regression line) and the random component  $AB$ .

### Assumptions

- (i) There exists a linear relationship between  $X$  and  $Y$  variables.
- (ii) The values of independent variable  $Y$  are fixed while those of dependent variable  $X$  are random - with randomness arising from the error term.
- (iii) The errors,  $e$ , are normally distributed with mean equal to zero, and constant variance  $\sigma^2$ . Further, they are independent in different observations.

Observe here that if  $X$  and  $Y$  are plotted on the graph on  $X$ -axis and  $Y$ -axis respectively, then we consider horizontal deviations in the case of regression of  $X$  on  $Y$  and vertical deviations in the case of regression of  $Y$  on  $X$ . The estimates of the parameters are given by  $a$  and  $b$ , which are obtained from a pair of normal equations given below :

$$\Sigma X = Na + b\Sigma Y$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

We calculate the input values and substitute them into the above equations, solve them simultaneously to get  $a$  and  $b$ . This yields the regression equation  $X = a + bY$ . This equation is then used to get the expected values of  $X$  for given values of  $Y$ .

### Some important points may be noted as follows:

- (i) For a given set of paired data, there are two regression lines - one showing regression of  $Y$  on  $X$  and the second one showing regression of  $X$  on  $Y$ . One of these is obtained by minimizing the squared vertical deviations and the other by horizontal deviations. As such, they are different lines with separate parameter values.
- (ii) The slope parameters of the regression lines are of particular significance. To distinguish, they are designated as  $b_{yx}$  and  $b_{xy}$ , called regression co-efficient of  $Y$  on  $X$  and, regression co-efficient of  $X$  on  $Y$ , respectively.



- (iii) For a given set of data, both the regression lines would be either positively sloped or negatively. Thus, both the regression co-efficients would be positive or both would be negative. Positive co-efficients indicate positive correlation and negative co-efficients mean negative correlation. The sign of the other parameter,  $a$ , is not important and for the two equations, it may bear same or opposite signs.
- (iv) Each of the regression lines passes through the mean values of the variables. When both the regression lines for a given set of paired data are plotted on a graph, their point of intersection yields the mean values of the variables  $X$  and  $Y$ . Thus, when two regression equations are solved simultaneously, the  $X$  and  $Y$  values obtained are  $\bar{X}$  and  $\bar{Y}$ , respectively.
- (v) The closer are the regression lines to each other, the higher is the degree of correlation between the variables and more away are they from each other, the weaker is the correlation. When the variables are perfectly correlated, the two regression lines coincide. Thus, while usually there are two regression lines for a set of data, when the correlation is perfectly positive or perfectly negative, there would be only one regression line.

The two regression lines in respect of a given set of data are both sloped positively in the case of positive correlation and negatively in the case of negative correlation between the variables. They intersect at  $\bar{X}$  and  $\bar{Y}$ , and their closeness to each other is indicative of the degree of correlation between the variables.

#### 5.5.4 Properties of regression coefficients

**Q29. What are the properties of regression coefficient?**

*Ans :*

- (i)  $r^2 = b_{xy} * b_{yx}$  In other words,  $r$  is the Geometric mean between the two regression coefficients  $b_{xy}$  and  $b_{yx}$
- (ii) Both the regression coefficients will have the same sign, i.e. either they will be positive or negative. Also, the coefficient of correlation will have the same sign as that of regression coefficients
- (iii) The arithmetic mean of the two regression coefficients is greater than the correlation coefficient. In other words,  $(b_{xy} + b_{yx})/2 > r$ .
- (iv) If one regression coefficient is greater than 1, the other has to be less than one. This is an extension of the first property as the product of the two coefficients is equal to square of the correlation coefficient  $r$ . Since  $r$  lies between -1 and +1,  $r^2$  cannot be greater than 1. Thus,  $b_{xy} * b_{yx}$  cannot be greater than 1. Thus, if one regression coefficient is greater than 1, the other has to be less than one.

**14. Given that the means of two variables  $X$  and  $Y$  are 68 and 150, their standard deviations are 2.5 and 20 respectively and the coefficient of correlation between them is +0.6, write down the equation of  $X$  and  $Y$ .**

*Sol :*

Regression Equation of  $X$  on  $Y$ .

$$= X - \bar{X} = r = \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$

$$= \bar{X} = 68, \bar{Y} = 150,$$

$$\sigma_x = 2.5, \sigma_y = 20, r = 0.6$$

$$= X - 68 = 0.6 \frac{2.5}{20}$$

$$= (Y - 150) \Rightarrow X - 68$$

$$= 0.075 Y - 11.25$$

$$\Rightarrow X = 0.075 Y - 11.25 + 68$$

$$\Rightarrow X = 0.075 Y + 56.75$$

15. For some bivariate data, the following results were obtained. The mean value of  $X = 53.2$  The mean value of  $Y = 27.9$ , the regression coefficient of  $Y$  on  $X = -0.15$ , and the regression coefficient of  $X$  on  $Y = -0.2$  find the most probable value of  $Y$  when  $X = 60$ .

*Sol.:*

Regression Equation of  $Y$  on  $X$

$$Y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - \bar{X}) = \bar{X} = 53.2,$$

$$\bar{Y} = 27.9 r \frac{\sigma_y}{\sigma_x} = 1.5$$

$$= Y - 27.9 = -1.5 (X - 53.2)$$

$$\Rightarrow Y - 27.9 = -1.5 (X - 53.2)$$

$$\Rightarrow Y - 27.9 = -1.5 X + 79.8 + 27.9$$

$$\Rightarrow Y = 107.7 - 1.5 X$$

When  $X = 60$ ,  $Y$  will be  $107.7 - 1.5(60)$

$$= 107.7 - 90 = 17.7$$

16. If  $\gamma=0.6$ ,  $\sigma_x=1.5$  and  $\sigma_y=2$ , Find the  $b_{xy}$  and  $b_{yx}$ .

*Sol.:*

Regression equation  $x$  on  $y$  co-efficient is  $b_{xy}$

$$b_{xy} = \gamma \frac{\sigma_x}{\sigma_y}$$

there  $\gamma=0.6$ ,  $\sigma_x=1.5$  and  $\sigma_y=2$

$\gamma$  = Coarctation co-efficient

$\sigma_x$  = Standard Deviation of 'x' series

$\sigma_y$  = Standard Deviation of 'y' series

$$b_{xy} = 0.6 * \frac{1.5}{2}$$

$$0.6 * 0.75$$

$$\boxed{b_{xy} = 0.45}$$

➤ Regression equation  $y$  on  $x$  co-efficient is  $b_{yx}$

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} = 0.6 * \frac{2}{1.5}$$

$$0.6 * 1.333 = \boxed{b_{yx} = 0.8}$$

17. From the following data obtain the two regression equations and calculate the correlation co-efficient.

x	2	4	6	8	10	12	14	16	18
y	18	16	20	24	22	26	28	32	30

Calculate the value of y when x = 6.2

*Sol:*

- (i) X on Y

$$X - \bar{X} = b_{xy} (Y - \bar{Y})$$

- (ii) Y on X

$$Y - \bar{Y} = b_{yx} (X - \bar{X})$$

$$\bar{X} = \frac{\sum x}{n} = \frac{90}{9} = 10$$

$$\bar{Y} = \frac{\sum y}{n} = \frac{216}{9} = 24$$

#### Calculation of Regression Equation and Correlation Coefficient

X	Y	x - $\bar{x}$	y - $\bar{y}$	$x^2$	$y^2$	xy
		x	y			
2	18	-8	-6	64	36	48
4	16	-6	-8	36	64	48
6	20	-4	-4	16	16	16
8	24	-2	0	4	0	0
10	22	0	-2	0	4	0
12	26	2	2	4	4	4
14	28	4	4	16	16	16
16	32	6	8	36	64	48
18	30	8	6	64	36	48
90	216	0	0	240	240	228

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{228}{240} = 0.95$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{228}{240} = 0.95$$

**Equation X on Y**

$$X - \bar{X} = b_{xy} (Y - \bar{Y})$$

$$X - 10 = 0.95 (Y - 24)$$

$$X = 0.95 Y - 22.8 + 10$$

$$X = 0.95 Y - 12.8$$

**Equation Y on X**

$$Y - \bar{Y} = b_{yx} (X - \bar{X})$$

$$Y - 24 = 0.95 (X - 10)$$

$$Y - 24 = 0.95 X - 9.5 + 24$$

$$Y = 0.95 X - 14.5$$

Regression equation y on x = 6.2

$$y = 0.95x - 14.5$$

$$y = 0.95(6.2) - 14.5$$

$$y = -8.61$$

**18. Given :**

$$\Sigma x = 56, \Sigma y = 40, \Sigma x^2 = 524, \Sigma y^2 = 256, \Sigma xy = 364, N = 8$$

(i) Find the two Regression equations and

(ii) The Correlation Coefficient.

*Sol.:*

We have

$$\bar{X} = \frac{\Sigma x}{N} = \frac{56}{8} = 7; \quad \bar{y} = \frac{\Sigma y}{N} = \frac{40}{8} = 5$$

$$b_{yx} = \text{co-efficient of regression of y on x} = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{N(\Sigma y^2) - (\Sigma x)^2}$$

$$\frac{8(364) - (56)(40)}{8(256) - (56)^2} = \frac{2912 - 2240}{4096 - 3136} = \frac{672}{1056}$$

$$b_{yx} = 0.6363$$

$$b_{xy} = \text{co-efficient of regression of } x \text{ on } y = \frac{N\sum xy - (\sum x)(\sum y)}{N(\sum y^2) - (\sum y)^2}$$

$$\frac{8(364) - (56)(40)}{8(256) - (40)^2} = \frac{2912 - 2240}{2048 - 1600} = \frac{672}{448} = \boxed{b_{xy} = 1.504}$$

(i) **Two Regression equations**

Regression equation x on y

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

$$(x - 7) = 1.504(y - 5)$$

$$(x - 7) = 1.504(y) - 1.504(5)$$

$$x = 1.504(y) - 7.522 + 7$$

$$x = 1.504(y) - 0.522 \dots\dots (1)$$

Regression equation y on x

$$(y - \bar{y}) = b_{yx}(x - \bar{x})$$

$$(y - 5) = 0.6363(x - 7)$$

$$y - 5 = 0.6363(x) - 0.6363(7)$$

$$y = 0.6363(x) - 4.4541 + 5$$

$$y = 0.6363(x) + 0.5459 \dots\dots (2)$$

(ii) **The correlation co-efficient  $\gamma_{xy}$  between x and y is given by**

$$\gamma_{xy}^2 = b_{yx} \cdot b_{xy} = (0.6363)(1.504)$$

$$r_{xy}^2 = 0.9569$$

$$\boxed{\gamma_{xy} = 0.9782}$$

**19. Following are the marks in Statistics and English in an Annual Examination.**

Particular	Statistics (X)	English (Y)
Mean	40	50
Standard Derivation	10	16
Co-efficient Correlation		0.5

(i) Estimate the score of English, when the score in Statistics is 50.

(ii) Estimate the score of statistics, when the score in English is 30.

*Sol:*

Given mean of X denoted as  $\bar{X} = 40$ .

Given mean of Y denoted as  $\bar{Y} = 50$ .

SD of X denoted as  $\sigma_x = 10$ .

SD of Y denoted as  $\sigma_y = 16$ .

Coefficient of correlation denoted as  $r = 0.5$

#### Regression Equation X on Y

$$[X - \bar{X}] = [r] \left[ \frac{\sigma_x}{\sigma_y} \right] [Y - \bar{Y}]$$

$$X - 40 = [0.5] \left[ \frac{10}{16} \right] [Y - 50]$$

$$X - 40 = [0.5] [0.625] [Y - 50]$$

$$X - 40 = [0.3125] [Y - 50]$$

$$X - 40 = 0.3125y - 15.625$$

$$X = 0.3125y - 15.625 + 40$$

$$X = 0.3125y + 24.375$$

#### Regression Equation Y on X

$$[Y - \bar{Y}] = [r] \left[ \frac{\sigma_y}{\sigma_x} \right] [X - \bar{X}]$$

$$[Y - 50] = [0.5] \left[ \frac{16}{10} \right] [X - 40]$$

$$Y - 50 = [0.5] [1.6] [X - 40]$$

$$Y - 50 = (0.8) (X - 40)$$

$$Y - 50 = 0.8X - 32$$

$$Y = 0.8X - 32 + 50$$

$$Y = 0.8X + 18$$

Estimation of English (Y) when Statistics (X) is 50

$$Y = 0.78X + 18$$

$$= 0.8(50) + 18$$

$$= 40 + 18$$

$\therefore Y = 18$  marks.

Estimation of statistics (X) when English (Y) is 30

$$\begin{aligned}
 X &= 0.3125 Y + 24.375 \\
 &= 0.3125(30) + 24.375 \\
 &= 9.375 + 24.375 \\
 X &= 33.75 \text{ marks.}
 \end{aligned}$$

20. You are given the following information about advertisement expenditure and sales :

Particulars	Adv. Exp (X) ( ` Crores)	Sales (Y) ( ` Crores)
Mean	20	120
S.D	5	25
Correlation coefficient		0.8

- (i) Calculate two regression equations  
(ii) Find likely sales when Adv. Expenses is ` 25 Crores  
(iii) What should be the Adv. Budget if the company wants to attain sales target of ` 150 crores.

*Sol :*

Let, the variable  $x$  represent advertisement expenses and  $y$  represent sales (in ` crores). Then, we have,

$$\bar{x} = 20, \bar{y} = 120, \sigma_x = 5, \sigma_y = 25, r = 0.8$$

(i) Two Regression Equations

(a) Regression Equation of X on Y

$$\begin{aligned}
 x - \bar{x} &= r \left[ \frac{\sigma_x}{\sigma_y} \right] (y - \bar{y}) \\
 \Rightarrow x - 20 &= 0.8 \left[ \frac{5}{25} \right] (y - 120) \\
 \Rightarrow x - 20 &= 0.16 (y - 120) \\
 \Rightarrow x - 20 &= 0.16 y - 19.2 \\
 \Rightarrow x &= 0.16 y - 19.2 + 20 \\
 \Rightarrow x &= 0.16 y + 0.8
 \end{aligned}$$

(b) Regression Equation of Y on X

$$\begin{aligned}
 y - \bar{y} &= r \left[ \frac{\sigma_y}{\sigma_x} \right] (x - \bar{x}) \\
 \Rightarrow y - 120 &= 0.8 \left[ \frac{25}{5} \right] (x - 20) \\
 \Rightarrow y - 120 &= 4(x - 20) \\
 y - 120 &= 4x - 80
 \end{aligned}$$

$$\Rightarrow y = 4x - 80 + 120$$

$$\Rightarrow y = 4x + 40$$

∴ The two equations are,

$$x = 0.16y + 0.8$$

$$y = 4x + 40$$

- (ii) For proposed advertisement expenditure of ₹ 25 crores, sales,

$$y = 4x + 40$$

$$= 4(25) + 40$$

$$= ₹ 140 \text{ crore}$$

- (iii) To achieve the sales target of ₹ 150 crores the company should have the advertising budget,

$$x = 0.16y + 0.8$$

$$= 0.16(150) + 0.8$$

$$= ₹ 24.8 \text{ crore.}$$

21. From the following data, calculate the regression equations taking deviation of items from the mean of X and Y series.

X	6	2	10	4	8
Y	9	11	5	8	7

*Sol:*

Regression Equation of X on Y

$$X - \bar{X} = b_{xy} (Y - \bar{Y})$$

$$\bar{X} = \frac{\sum X}{N} = \frac{30}{5} = 6$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{40}{5} = 8$$

$$\therefore \bar{X} = 6, \bar{Y} = 8$$

**Calculation of Regression equation and correlation coefficient**

X	Y	X - $\bar{X}$ (x)	Y - $\bar{Y}$ (y)	x <sup>2</sup>	y <sup>2</sup>	xy
6	9	0	1	0	1	0
2	11	-4	3	16	9	-12
10	5	4	-3	16	9	-12
4	8	-2	0	4	0	0
8	7	2	-1	4	1	-2
$\sum X = 30$	$\sum Y = 40$	$\sum x = 0$	$\sum y = 0$	$\sum x^2 = 40$	$\sum y^2 = 20$	$\sum xy = 26$



Number of Pairs  $N = 5$

### Regression Co-efficient

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{-26}{20} = -1.3$$

$$\therefore X - \bar{X} = b_{xy} (Y - \bar{Y})$$

$$X - 6 = 1.3 (Y - 8)$$

$$X - 6 = 1.3 Y + 10.4$$

$$X = -1.3Y + 10.4 + 6$$

$$X = 1.3Y + 16.4$$

(or)

$$X = 16.4 - 1.3 Y$$

### Regression Equation of Y on X

$$(Y - \bar{Y}) = b_{yx} (X - \bar{X})$$

### Regression Coefficient

$$b_{yx} = \frac{\Sigma x^2}{\Sigma y^2} = \frac{-26}{40} = -0.65$$

$$\therefore Y - \bar{Y} = b_{yx} (X - \bar{X})$$

$$Y - 8 = -0.65 (X - 6)$$

$$Y - 8 = -0.65x + 3.9$$

$$Y = -0.65x + 3.9 + 8$$

$$Y = -0.65x + 11.9$$

or

$$Y = 11.9 - 0.65x$$

22. The following table gives the age of cars of a certain make and annual main tenance costs. Obtain the regression equation for costs related to age.

Age of cars in year	2	4	6	8
Maintenance Cost in ` ('00) Hundered of `	10	20	25	30

*Sol :*

(Dec.-20)

Let the age of cars be 'X' and the maintenance costs be 'Y'.

Regrassion equation of Y on X is

$$Y - \bar{Y} = b_{yx} (X - \bar{X})$$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{20}{4} = 5$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{85}{4} = 21.25$$

X	Y	$d_x = X - 5$	$d_y = Y - 21$	$d_x^2$	$d_x d_y$
2	10	-3	-11	9	33
4	20	-1	-1	1	1
6	25	1	4	1	4
8	30	3	9	9	27
$\Sigma X = 20$	$\Sigma Y = 85$	$\Sigma d_x = 0$	$\Sigma d_y = 1$	$\Sigma d_x^2 = 20$	$\Sigma d_x d_y = 65$

$$b_{xy} = \frac{n \Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n \Sigma d_x^2 - (\Sigma d_x)^2}$$

$$= \frac{4(65 - (0)(1))}{4(20) - (0)^2}$$

$$= \frac{260 - 0}{80 - 0} = \frac{260}{80} = 3.25$$

Regression equation Y on X is,

$$Y - \bar{Y} = b_{xy} (X - \bar{X})$$

$$Y - 21.25 = 3.25(X - 5)$$

$$Y - 21.25 = 3.25 X - 16.25$$

$$Y = 3.25 X + 21.25 - 16.25$$

$$Y = 3.25 X + 5$$

**23. Find both regression lines to the following:**

**Mean (X) = 15**

**Mean (Y) = 110**

**Variance (X) = 25**

**Variance (Y) = 62.5 and r = 0.81.**

*Sol.:*

**(July-18, Imp.)**

Given that,

$$\text{Mean } (\bar{X}) = 15$$

$$\text{Mean } (\bar{Y}) = 110$$

$$\text{Variance (X)} = 25 \quad \sigma_x = \sqrt{25} = 5$$

$$\text{Variance (Y)} = 625 \quad \sigma_y = \sqrt{625} = 25$$

$$\text{Regression coefficient of Y on X} = b_{yx} = r \frac{\sigma_y}{\sigma_x} = 0.81 \times \frac{25}{5} = 4.05$$

$$\text{Regression coefficient of X on Y} = b_{xy} = r \frac{\sigma_x}{\sigma_y} = 0.81 \times \frac{5}{25} = 0.162$$

Thus, the line of regression of Y on X is :

$$(Y - \bar{Y}) = b_{yx} (X - \bar{X})$$

$$(Y - 110) = 4.05 (X - 15)$$

$$Y = 4.05 X - 60.75 + 110$$

$$Y = 4.05X + 49.25$$

and the line of regression of X on Y is :

$$(X - \bar{X}) = b_{xy} (Y - \bar{Y})$$

$$(X - 15) = 0.162 (Y - 110)$$

$$X = 0.162 Y - 17.82 + 15$$

$$X = 0.162 Y - 2.82$$

24. Find both regression lines to the following. Estimate the coefficient correlation and comment.

X	13	48	88	42	22	10	6
Y	8	52	82	84	22	10	6

*Sol :*

(Aug.-17)

$$\bar{x} = \frac{\Sigma x}{n} = \frac{229}{7} = 32.71$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{264}{7} = 37.71$$

As the values of  $\bar{x}$  and  $\bar{y}$  are in fraction, let us assume the values of means to simplify the calculations.

Assuming,  $\bar{x} = 33$  and  $\bar{y} = 38$ .

x	y	u = x - 33	v = y - 38	u <sup>2</sup>	v <sup>2</sup>	uv
13	8	- 20	- 30	400	900	600
48	52	15	14	225	196	210
88	82	55	44	3025	1936	2420
42	84	9	46	81	2116	414
22	22	-11	-16	121	256	176
10	10	- 23	- 28	529	784	644
6	6	-27	-32	729	1024	864
$\Sigma x = 229$	$\Sigma y = 264$	$\Sigma u = - 2$	$\Sigma v = - 2$	$\Sigma u^2 = 5110$	$\Sigma v^2 = 7212$	$\Sigma uv = 5328$

When deviations are taken from assumed mean, the formula used for computing  $b_{xy}$  and  $b_{yx}$  are as follows,

$$b_{xy} = b_{uv} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2}$$

$$= \frac{7(5328) - (-2)(-2)}{7(7212) - (-2)^2} = \frac{37296 - 4}{50484 - 4} = \frac{37292}{50480} = 0.74$$

$$b_{yx} = b_{vu} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2}$$

$$= \frac{7(5328) - (-2)(-2)}{7(5110) - (-2)^2} = \frac{37296 - 4}{35770 - 4} = \frac{37292}{35766} = 1.04$$

#### Regression Line x on y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\text{Actual } \bar{x} = 32.71, \bar{y} = 37.71$$

$$x - 32.71 = 0.74(y - 37.71)$$

$$x - 32.71 = 0.74y - 27.91$$

$$x = 0.74y + 32.71 - 27.91$$

$$x = 0.74y + 4.8$$

#### Regression Line y on x

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 37.71 = 1.04(x - 32.71)$$

$$y - 37.71 = 1.04x - 34.02$$

$$y = 1.04x + 37.71 - 34.02$$

$$y = 1.04x + 3.69$$

**Correlation Coefficient**

$$r = \pm \sqrt{b_{xy} \times b_{yx}}$$

$$= \pm \sqrt{0.74 \times 1.04} = \pm \sqrt{0.7696} = \pm 0.88$$

25. Find the equation of the line of regression of x on y for the following data:

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y	5.3	5.7	6.3	7.2	8.2	8.7	8.4

Find the (i) x on y

(ii) y on x

Sol.:

(Nov.-20)

$$\bar{X} = \frac{17.5}{7} = 2.5 \approx 3$$

$$\bar{Y} = \frac{49.8}{7} = 7.11 \approx 7$$

X	Y	$d_x = X - 3$	$d_y = Y - 7$	$d_x^2$	$d_y^2$	$d_x d_y$
1.0	5.3	-2	-1.7	4	2.89	3.4
1.5	5.7	-1.5	-1.3	2.25	1.69	1.95
2.0	6.3	-1	-0.7	1	0.49	0.7
2.5	7.2	-0.5	0.2	0.25	0.04	-0.1
3.0	8.2	0	1.2	0	1.44	0
3.5	8.7	0.5	1.7	0.25	2.89	0.85
4.0	8.4	1	1.4	1	1.96	1.4
$\Sigma X = 17.5$	$\Sigma Y = 49.8$	$\Sigma d_x = -3.5$	$\Sigma d_y = 0.8$	$\Sigma d_x^2 = 8.75$	$\Sigma d_y^2 = 11.4$	$\Sigma d_x d_y = 8.2$

**Regression Equation of X on Y**

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$b_{xy} = \frac{n \Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n \Sigma d_y^2 - (\Sigma d_y)^2}$$

$$= \frac{7(8.2) - (-3.5)(0.8)}{7(11.4) - (0.8)^2}$$

$$= \frac{57.4 + 2.8}{79.8 - 0.64}$$

$$= \frac{60.2}{79.16} = 0.76$$

∴ The regression equation X on Y as follows,

$$X - 2.5 = 0.76(Y - 7.11)$$

$$X - 2.5 = 0.76 Y - 5.40$$

$$X = 0.76 Y + 2.5 - 5.40$$

$$X = 0.76 Y - 2.90$$

### Regression Equation of Y on X

$$Y - \bar{Y} = b_{yx} (X - \bar{X})$$

$$b_{xy} = \frac{n\sum d_x d_y - (\sum d_x)(\sum d_y)}{n\sum d_x^2 - (\sum d_x)^2}$$

$$= \frac{7(8.2) - (-3.5)(0.8)}{7(8.75) - (-3.5)^2}$$

$$= \frac{57.4 + 2.8}{61.25 - 12.25}$$

$$b_{xy} = \frac{60.2}{49} = 1.22$$

∴ The regression equation Y on X as follows,

$$Y - 7.11 = 1.22 (X - 2.5)$$

$$Y - 7.11 = 1.22 X - 3.05$$

$$Y = 1.22 X + 7.11 - 3.05$$

$$Y = 1.22 X - 4.06$$

## 5.6 TIME SERIES ANALYSIS

**Q30. Define time series.**

**(OR)**

**What is time series.**

*Ans :*

**(Imp.)**

### Meaning

A time series is a statistical data that are collected, observed or recorded at regular intervals of time. The term time series applies, for example,

to the data recorded periodically showing the total annual sales of retail stores, the total quarterly value of construction contracts awarded, the total amount of unfilled orders in durable goods industries at the end of each month, weekly earnings of workers in an industrial town, hourly temperature in a particular city.

### Definitions

Some of the important definitions of time series, given by different experts are as under:

- i) **According to Morris Hamburg**, "A time series is a set of statistical observations arranged in chronological order."
- ii) **According to Patterson**, "A time series consists of statistical data which are collected, recorded observed over successive increments."
- iii) **According to Ya-Lun-Chou**, "A time series may be defined as a collection of magnitudes belonging to different time periods, of some variable or composite of variables, such as production of steel, per capita income, gross national product, price of tobacco, or index of industrial production."
- iv) **According to Wessel and Wellet**, "When quantitative data are arranged in the order of their occurrence, the resulting statistical series is called a time series."
- v) **According to Spiegel**, "A time series is a set of observations taken at specified times, usually at 'equal intervals'. Mathematically, a time series is defined by the values  $Y_1 Y_2 \dots$  of a variable Y (temperature, closing price of a share, etc.) at times  $t_1, t_2 \dots$ . Thus Y is a function of  $t_1$  symbolized by  $Y = F(t)$ ."
- vi) **According to Cecil H. Mayers**, "A time series may be defined as a sequence of repeated measurement of a variables made periodically through time".

It is clear from the above definitions that time series consist of data arranged chronologically. Thus if we record the data relating to population, per capita income, prices, production, etc., for the last 5, 10, 15, 20 years or some other time period, the series so emerging would be called time series.

It should be noted that the term 'time series' is usually used with reference to economic data and the economists are largely responsible for the development of the techniques of time series analysis. However, the term 'time series' can apply to all other phenomena that are related to time such as the number of accidents occurring in a day, the variation in the temperature of a patient during a certain period, number of marriages taking place during a certain period, etc.

**Q31. What are the characteristics of time series ?**

*Ans :* (Imp.)

The essentials of Time Series are

- i) It must consist of a set of values that are homogeneous. For example, production data for a year and sales data for the next year will not be a time series.
- ii) The values must be with reference to time. In other words, in a time series, we have at least 2 variables, with one variable necessarily being time.
- iii) The data must be available for a reasonably long period of time.
- iv) The gaps between various time values should as far as possible be equal.
- v) The values of the second variable should be related to time. For example, the number of people being hired by the BPO industry can be tracked in relation to time. However, if we are talking about average height of students in a class, this data may not have a significant relationship with time and may not constitute a time series.

**Q32. What are the objectives of time series ?**

*Ans :*

While it is true that past performance does not necessarily guarantee future results, the quality of forecasts that management can make is strongly related to the information that can be extracted and used from past data. Thus, the objective of time series analysis is to interpret the changes in a given variable with reference to the given situation and attempt to anticipate the future course of events.

Analysis of Time Series is done with the following objectives:

- i) To evaluate past performance in respect of a particular variable.
- ii) To make future forecasts in respect of the particular variable.
- iii) To chart short term and long term strategies of the business in respect of the particular variable.

**5.6.1 Components**

**Q33. What are the components of time series analysis.**

*Ans :* (Imp.)

A time series is a historical series of statistical data. Since these statistical data are historical, they are subject to the influences of all the changes that may occur over any period of time. The usual classification of the influences affecting statistical data recording over time is one based upon the nature of the influence. Classified in this manner, these influences are as follows:

**1) Secular Variations or Trend (T)**

The general tendency of the time series data to increase or decrease or stagnate during a long-period of time is called the secular trend or simply trend.

**According to Simpson and Kafka** "Trend, also called secular or long-term trend, is the basic tendency of a series...to grow or decline over a period of time. The concept of trend does not include short-range oscillations, but rather the steady movements over a long time".

The trend thus refers to the general direction and the movement of time series considering a fairly long period of time. Most business and economic series are influenced by some secular forces of change 'in which the underlying tendency is one of growth or decline, for example, in series concerning population, national income, agricultural production, currency in circulation, bank deposits, etc., an upward tendency is usually observed while a downward tendency is

noticed in time series relating 10 birth and death rates specially in deaths by epidemics, tuberculosis, etc.

## 2) Seasonal Variations (S)

Seasonal variations refers to rhythmic forces of change inherent in most time series showing a regular or a periodic pattern of movement over a span of less than a year and has the same or almost the same pattern year after year.

**According to Hirsch** " a recurrent pattern of change within the period that result from operation of forces connected with climate or custom at different times of the period".

### Reasons for Seasonal Variations

#### i) Climate and Natural Forces

The result of natural forces like climate is causing seasonal variation. Umbrellas are sold more in rainy season. In winter season, sale of the woolen clothes will increase. In hot season, the sales of ice, ice-cream, fruit salad, etc., will increase, thus climate and weather play an important role in seasonal movement. Agricultural production depends upon the monsoon.

#### ii) Customs and Habits

Man-made conventions are the customs, habits, fashion, etc. There is the custom of wearing new clothes, preparing sweets and buying crackers for Deepavali, Onam, Christmas, etc. At that time, there is more demand for cloth, sweets and crackers. It will happen every year. In marriage season, the price of gold will increase.

### Purposes of Studying Seasonal Variations

- i) To analyze seasonal pattern in a short-period time series.
- ii) Once the seasonal factor is known it can be used for separating cyclical and irregular forces from the residual forces of changes.
- iii) The seasonal factor can be used for adjustment in the value projected on the basis of trend. Short-term forecasts are always affected by seasonal factors.
- iv) The observed value can be appraised in terms of seasonal factor and adjusted so as to get correct idea of the trend, if any, in an economic phenomenon. The study of seasonal variations is, therefore, necessary for proper appraisal of business facts influenced by a seasonal variation.
- v) The seasonal factor can also be used for pricing of articles and services so as to level up the seasonal variations in demand.
- vi) A study of the seasonal patterns is extremely useful in planning future operations and in formulation of policy decisions regarding purchase, production, inventory control, personnel requirement, selling and advertising programme, etc.

## 3) Cyclical Variations (C)

As the economy expands during a period of boom, we would expect to find that such data as sales, output or consumer expenditure also show a rising trend; and during a period of slump, we would expect them to show a downward trend. Thus a wavelike motion may be observed in the pattern of our data. They are prosperity (boom) recession, depression and recovery; and are shown below:



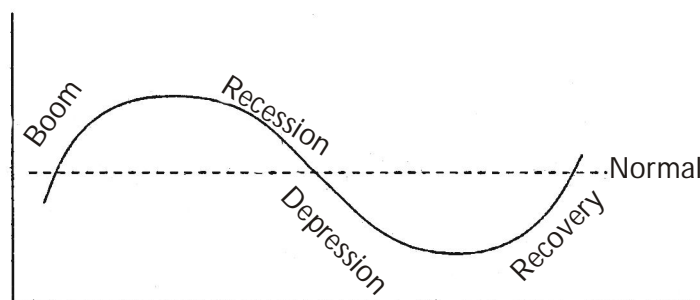


Fig.: Business Cycle

The study of cyclical variation is helpful to businessmen and economist for framing suitable policies and stabilizing the level of business activities.

#### 4) Irregular Variations (I)

Irregular variations are the effect of random factors. These are generally mixed up with seasonal and cyclical variations and are caused by irregular and accidental factors like floods, famines, wars, strikes, lockouts, etc. There is no regular period or time of their occurrence and that is why they are called random or chance fluctuations. Sometimes these factors are very effective and may even give rise to cyclical fluctuations. However since these are not regular factors, no advance preparation can be done to meet their consequences. Their effects are also unpredictable and irregular.

### 5.6.2 Models of Time Series

#### 5.6.2.1 Additive, Multiplicative and Mixed Models

**Q34. Describe the various models of time series.**

*Ans :*

There are two mathematical models which are commonly used for the decomposition of a time series into its components viz.

- (a) Decomposition by Additive model
- (b) Decomposition by Multiplicative model.
- (c) Decomposition by Mixed Model

**(a) Decomposition by Additive Model :** According to the additive mode, a time series can be expressed as

$$Y_t = T_t + S_t + C_t + R_t$$

where  $Y_t$  = is the time series value at time  $t$ ,

$T_t$  = represent the trend value.

$S_t$  = represents the seasonal variations.

$C_t$  = represents the cyclic movements and

$R_t$  = represents the random fluctuations at time  $t$ .

Obviously, the term  $S_t$  will not appear in a series of annual data. The additive model implicitly implies that seasonal forces (in different years), cyclical forces (in different cycles) and irregular

forces (in different long time periods) operate with equal absolute effect irrespective of the long trend value. As such  $C_t$  (and  $S_t$ ) will have positive or negative values according as whether we are in an above normal or below normal phase of the cycle (and year) and the total of positive and negative values for any cycle (and any year) will be zero.  $R_t$  will also have positive or negative values and in the long run  $R_t$  will be zero. Occasionally, there may be a few isolated occurrences of extreme  $R_t$  of episodic nature.

**(b) Decomposition by Multiplicative Model**

On the other hand if we have reasons to assume that the various components in a time series operate proportionately to the general level of the series, the traditional or classical multiplicative model is appropriate. According to the multiplicative model

$$Y_t = T_t \times S_t \times C_t \times R_t$$

where  $S_t$ ,  $C_t$  and  $R_t$  instead of assuming positive and negative values are indices fluctuating above or below unity and the geometric means of  $S_t$  in a year,  $C_t$  in a cycle and  $R_t$  in a long-term period are unity. In a time series with both positive and negative values the multiplicative model can not be applied unless the time series is translated by adding a suitable positive value. It may be pointed out that the multiplicative decomposition of a time series is same as the additive decomposition of the logarithmic values of the original time series.

In practice, most of the series relating to economic data confirm to the multiplicative model.

**(c) Decomposition by Mixed Models**

In addition to the additive and multiplicative models discussed above, the components in a time series may be combined in large number of ways. The different models, defined under different assumptions, will yield different results. Some of the defined models can be:

$$Y_t = (T_t \times S_t \times C_t) + R_t$$

$$Y_t = (T_t \times C_t) + (S_t \times R_t)$$

$$Y_t = T_t + (S_t \times C_t \times R_t)$$

$$Y_t = T_t + S_t + (C_t \times R_t)$$

$$Y_t = T_t \times (S_t + C_t) \times R_t$$

### 5.7 TREND ANALYSIS

**Q35. What is Trend Analysis? Explain the purpose of measuring trend.**

*Ans :*

The trend analysis refers to the concept of collecting information and attempting to spot a pattern, or trend, in the information. In some fields of study, the term "trend analysis" has more formally-defined meanings.

Although trend analysis is often used to predict future events, it could be used to estimate uncertain events in the past, such as how many ancient kings probably ruled between two dates, based on data such as the average years which other known kings reigned.

Trend analysis uses a technique called least squares to fit a trend line to a set of time series data and then project the line into the future for a forecast.

It is a special case of regression analysis, where the dependent variable is the variable to be forecasted and the independent variable is time.

#### Purposes of Measuring Trend

There are three basic purposes of measuring secular trend:

- 1) The first purpose is to study the past growth or decline of a series. The secular trend describes the basic growth tendency ignoring short-term fluctuations.
- 2) The second and most important purpose of measuring secular trend is to project the curve into the future as a long-term forecast. If the past growth has been steady and if the conditions that determine this growth may reasonably be expected to persist in the future, a trend curve may be projected over five to ten years into the future as a preliminary forecast.

- 3) The third purpose of measuring secular trend is to eliminate it, in order to clarify the cycles and other short-term movements in the data. A steep trend may observe minor cycles. Dividing the data, by the trend values yield ratios which make the curve fluctuate around a horizontal line, thus bringing the cycles into clear relief.

### 5.7.1 Free Hand Curve

**Q36. What is Free Hand Curve? Explain with illustration.**

*Ans :*

A trend is determined by just inspecting the plotted points on a graph sheet. Observe the up and down movements of the-points. Smooth out the irregularities by drawing a freehand curve or line through the scatter points. The curve so drawn would give a general notion of the direction of the change. Such a freehand smoothed curve eliminates the short- time swings and shows the long period general tendency of the changes in the data.

Drawing a smooth freehand curve requires a personal skill and judgement. The drawn curve should pass through the plotted points in such a manner that the variations in one direction are approximately equal to the variation in other direction. Different persons, however, drawn different curves at different directions, with different slopes and in different styles. This may lead to different conclusions. To overcome these limitations, we can use the semi-average method of measuring the trend.

#### Merits

- (i) It is very simple.
- (ii) It does not involve any calculations.
- (iii) It is very flexible and can be used irrespective of whether the trend is linear or curvilinear.
- (iv) If used by experienced statisticians, it is a better tool to study trend movement compared to other methods using rigid mathematical formulae.

#### Limitations

- (i) It is very subjective. Different persons may draw different lines and reach different conclusions from the same data. Hence, it is not a good forecasting tool.
- (ii) If properly attempted, it is very time-consuming effort.
- (iii) It requires high levels of experience and expertise to effectively use this method.

#### Example

Fit a trend line to the following data by the freehand method,

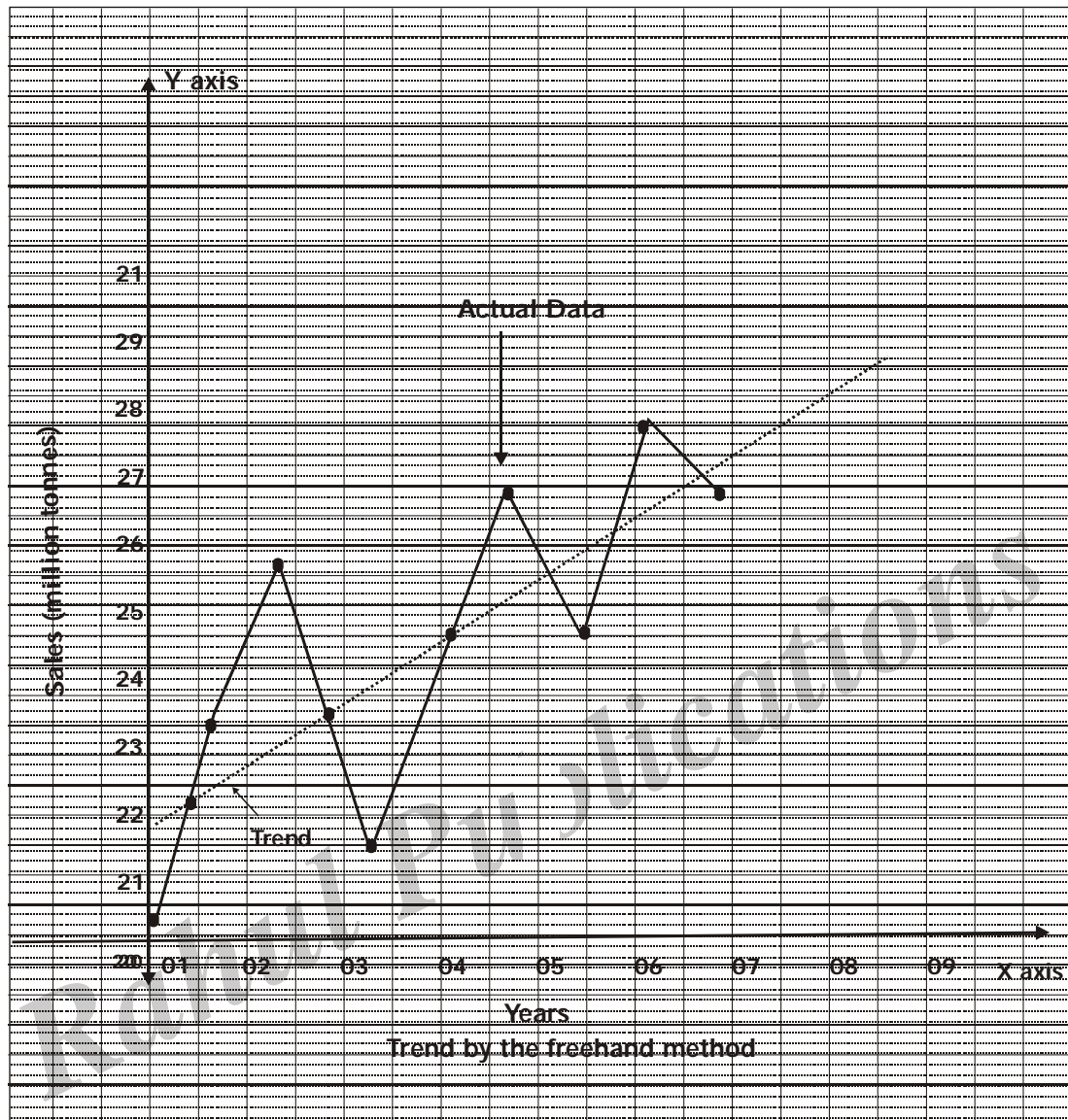
Year	2001	2002	2003	2004	2005	2006	2007	2008	2009
Sales (milllion tonnes)	19	22	24	20	23	25	23	26	25

*Sol :*

#### Steps

1. Time series data is plotted on the graph
2. The direction of the trend is examined on the basis of the plotted data (dots)
3. A straight line is drawn which shows the direction of the trend.

The actual data and the trend line are shown in the following graph.



### 5.7.2 Semi Averages

**Q37. Explain how trend analysis is done by semi averages method. State its merits and demerits.**

*Ans :*

- The entire time series is classified into two equal parts with respect to time. For even period, equal split. For odd period, equal parts obtained by omitting middle period.
- Compute the arithmetic mean of time series values for each half separately. These means are called semi-averages.
- Semi averages are plotted as points against the middle point of the respective time period covered by each part.
- The line joining these points gives the straight line trend fitting the given data.

**Merits of Semi-Average Method**

The following are some of the merits of semi-average method,

- Objectivity
- Ease of apply and understandability
- Extend both ways the line i.e., we can get past and future estimates.

**Demerits of Semi-Average Method**

Some of the demerits of semi-average method are,

- Linear trend assumption may not exist.
- A men may be questioned.
- Thus, values of trend are not precise and reliable.

**Example**

Using the following data, fit a trend line by using the method of semi-averages,

Year	1996	1997	1998	1999	2000	2001	2002
Output	700	900	1100	900	1300	1000	1600

*Sol :*

**Step 1**

The data provided in the problem is of seven years i.e., (an odd number). Thus, the middle year [ 1999] shall be ignored and the remaining years are divided into two equal time periods and their arithmetic averages is computed as follows,

$$\text{Average of the first three years} = \frac{700 + 900 + 1100}{3} = \frac{2700}{3} = 900$$

$$\text{Average of the last three years} = \frac{1300 + 1000 + 1600}{3} = \frac{3900}{3} = 1300$$

Therefore, the semi-averages are 900 and 1300

**Step 2**

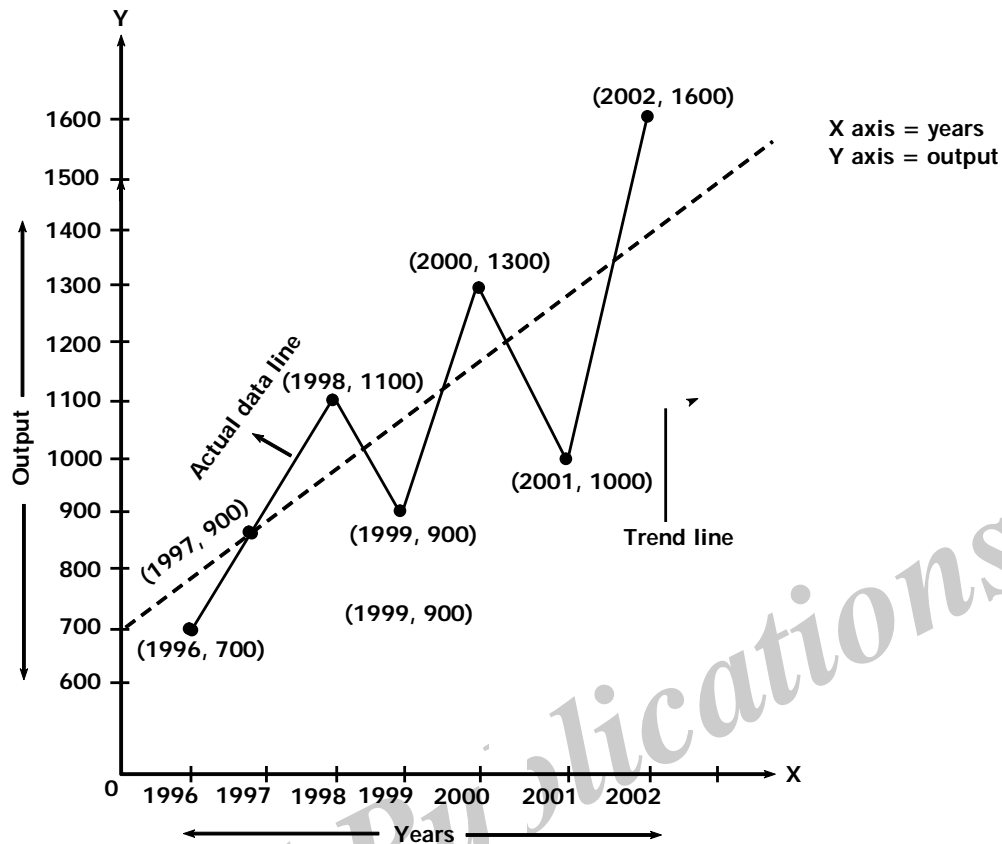
The next step is to plot the semi-averages against the mid-point (middle year) of each time period. Thus, it would , be year 1997 and 2001 respectively.

**Step 3**

The plotted points are joined in order to derive the trend line using the semi average method.

**Step 4**

The original data and the trend line is plotted on a graph as follows,



### 5.7.3 Moving Averages

**Q38. Discuss the method of moving averages in measuring trend. What are its merits and limitations of moving average method?**

*Ans :*

In moving average method, the average value for a number of years (month or weeks) is secured and this average is taken as the normal or trend value for the unit of time falling at the middle of the period covered in the calculation of the average.

The effect of averaging is to give a smoother curve, lessening the influence of the fluctuations that pull the annual figures away from the general trend.

The period of moving average is decided in the light of the length of the cycle. More applicable to data with cyclical movements.

Formula for 3 yearly moving average will be,

$$\frac{a+b+c}{3}, \frac{b+c+d}{3}, \frac{c+d+e}{3} \dots$$

Formula for 5 yearly moving average will be,

$$\frac{a+b+c+d+e}{5}, \frac{b+c+d+e+f}{5}, \frac{c+d+e+f+g}{5} \dots$$

**Methods**

The following two methods are followed in moving averages,

**a) Odd Yearly Method**

- i) Calculate 3/5...yearly totals
- ii) Now compute 3/5 yearly average by dividing the totals calculated in step (i) by the respective number of years, i.e. 3/5/...
- iii) Short term oscillations are calculated using the formula,  $Y - Y_C$

Where, Y - Actual value and  $Y_C$  - Estimated value.

**b) Even Yearly Method**

**Example :** 4 years

- i) Calculate 4 yearly moving totals and place at the centre of middle two years of the four years considered.
- ii) Divide 4 yearly moving totals by 4 to get 4 yearly average.
- iii) Take a 2 period moving average of the moving average which gives the 4 yearly moving average centered.

**Merits**

The merits of moving average are as follows,

- a) Of all the mathematical methods of fitting a trend, this method is the simplest.
- b) The method is flexible so that even if a few more observations are to be added, the entire calculations are not changed.
- c) If the period of the moving average happens to coincide with the period of the cycle, the cyclical fluctuations are automatically eliminated.
- d) The shape of the curve in case of moving average method is determined by the data rather than the statisticians choice of mathematical function.

**Limitations**

The following are the limitations of moving averages,

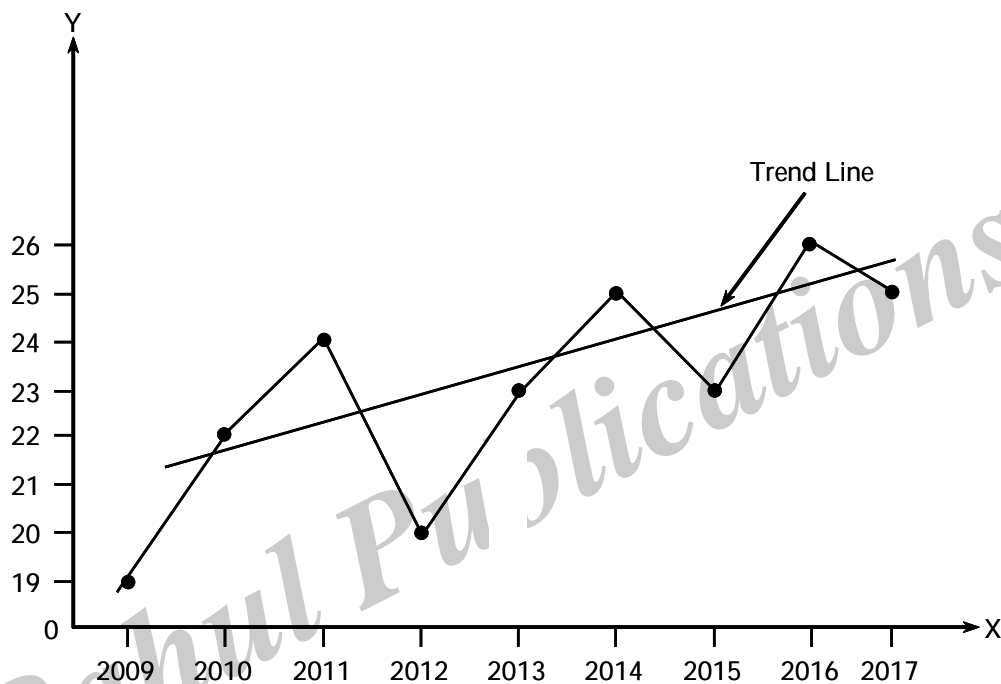
- a) Trend values cannot be computed for all the years. For example, in a 5 yearly moving we cannot compute trend values for the first two and the last two years.
- b) It is difficult to decide the period of moving average since there is no hard and fast rule for the purpose.
- c) Moving average cannot be used in forecasting as it is not represented by any mathematical function.
- d) When the trend is not linear, the moving average lies either above or below the true sweep of the data.

**PROBLEMS**

26. Fit a trend line to the following data by the freehand method.

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017
Sales (Rs.)	19	22	24	20	23	25	23	26	25

*Sol :*



27. From the following data fit a trend line by the method of Semi-Average.

Year :	2012	2013	2014	2015	2016	2017
Output :	20	16	24	30	28	32

*Sol :*

The trend value of first 3 years calculated as follows

$$\frac{20 + 16 + 24}{3} = \frac{60}{3} = 20.$$

The trend value of last 3 years are calculated as follows

$$\frac{30 + 28 + 32}{3} = \frac{90}{3} = 30$$

Therefore, the Semi Average are 20 and 30.



28. The following table gives the annual sales (in Rs.'000) of a commodity :

Year	Sales
1990	710
1991	705
1992	680
1993	687
1994	757
1995	629
1996	644
1997	783
1998	781
1999	805
2000	805

Determine the trend by calculating the 5-yearly moving average.

*Sol :*

#### Calculation of Trend by 5-Yearly Moving Average

Year	Sales	5-Yearly Moving Total	5-Yearly Moving Average
1990	710		
1991	705		
1992	680	710 + 705 + 680 + 687 + 757 = 3539	$\frac{3539}{5} = 707.8$
1993	687	705 + 680 + 687 + 757 + 629 = 3458	$\frac{3458}{5} = 691.6$
1994	757	680 + 687 + 757 + 629 + 644 = 3397	$\frac{3397}{5} = 679.4$

1995	629	687 + 757 + 629 + 644 + 783 = 3500	$\frac{3500}{5} = 700$
1996	644	757 + 629 + 644 + 783 + 781 = 3594	$\frac{3594}{5} = 718.8$
1997	783	629 + 644 + 783 + 781 + 805 = 3642	$\frac{3642}{5} = 728.4$
1998	781	644 + 783 + 781 + 805 + 872 = 3885	$\frac{3885}{5} = 777$
1999	805		
2000	872		

29. Calculate three year moving average for the following data:

Year :	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960
Value :	242	250	252	249	253	255	251	257	260	265	262

Sol.:

Calculation of Trend by 3 years moving average

Year	Value	3 year moving total	3 year moving average
1950	242	–	–
1951	250	242 + 250 + 252 = 744	744/3 = 248.00
1952	252	250 + 252 + 249 = 751	751/3 = 250.33
1953	249	252 + 249 + 253 = 754	754/3 = 251.33
1954	253	249 + 253 + 255 = 757	757/3 = 252.33
1955	255	253 + 255 + 251 = 759	759/3 = 253.00
1956	251	255 + 251 + 257 = 763	763/3 = 254.33
1957	257	251 + 257 + 260 = 768	768/3 = 256.00
1958	260	257 + 260 + 265 = 782	782/3 = 260.67
1959	265	260 + 265 + 262 = 787	787/3 = 262.33
1960	262	–	–

30. Calculate the trend values by the method of moving average, assuming a four yearly cycle, from the following data relating to sugar production in India.

Year	Sugar Production (Lakh Tonnes)	Year	Sugar Production (Lakh Tonnes)
1971	37.4	1977	48.4
1972	31.1	1978	64.6
1973	38.7	1979	58.4
1974	39.5	1980	38.6
1975	47.9	1981	51.4
1976	42.6	1982	84.4

Sol.:

Calculation of 4 years moving average

Year	Sugar Prod (lakh Tonnes)	4 yearly Moving Total	4 yearly moving Average	2 period Moving Total Average	Centered Moving
1971	37.4	-	-	-	-
1972	31.1	-	-	-	-
1973	38.7	$37.4 + 31.1 + 38.7 + 39.5 = 146.7$	36.675	$36.675 + 39.300 = 75.975$	37.99
1974	39.5	$31.1 + 38.7 + 39.5 + 47.9 = 157.2$	39.300	$39.300 + 42.175 = 81.475$	40.75
1976	47.9	$38.7 + 39.5 + 47.9 + 42.6 = 168.7$	42.175	$42.175 + 50.875 = 93.05$	46.525
1977	48.4	$47.9 + 42.6 + 48.4 + 64.6 = 203.5$	50.875	$50.875 + 53.500 = 104.375$	52.19
1978	64.6	$42.6 + 48.4 + 64.6 + 58.4 = 214.0$	53.500	$53.500 + 52.500 = 106.000$	53.00
1979	58.4	$48.4 + 64.6 + 58.4 + 38.6 = 210.00$	52.500	$52.500 + 53.250 = 105.750$	52.88
1980	38.6	$64.6 + 58.4 + 38.6 = 161.6$	52.500	$53.500 + 58.200 = 111.700$	55.73
1981	51.4	-	-	-	-
1982	84.4	-	-	-	-

**5.7.4 Least Square Methods****Q39. Define least square method. Explain merits and demerits of least square method.****(OR)****What is least square method and explain its advantages and disadvantages?***Ans :***(Imp.)**

Least square method is the most widely used method and provides us with a mathematical device to obtain an objective fit to the trend of a given time series. This method is so called because a trend line computed by this method is such that the sum of the squares of the deviation between the original data and the corresponding computed trend values is minimum. This method can be used to fit either a straight line trend or a parabolic trend.

The straight line trend equation is in the form of  $Y = a + bX$

Where, Y denotes the trend value of the dependent variable

X denotes the independent variable.

a and b are constants.

The values of a and b are obtained by solving the following normal equations.

$$\Sigma Y = Na + b \Sigma X$$

$$\Sigma XY = a \Sigma X + b \Sigma X^2$$

Where, N represents the number of years in the series.

When  $\Sigma X = 0$  the above normal equations are simplified to

$$a = \frac{\Sigma Y}{N}$$

$$b = \frac{\Sigma XY}{\Sigma X^2}$$

By substituting a and b values in straight line trend equation  $Y = a + bX$ , we get the straight line equation which can be used for estimation of future values.

**Merits of Least Squares Method**

The following are the merits of least squares method,

1. The method of least squares is a mathematical method of measuring trend and is free from subjectiveness.
2. This method provides the line of best fit since it is this line from where the sum of positive and negative deviations is zero and the sum of square of deviations is the least.
3. This method enables us to compute the trend values for all the given time periods in the series.
4. The trend equation can be used to estimate the values of the variable for any given time period 'f' in future and the forecasted values are quite reliable.
5. This method is the only technique which enables us to obtain the rate of growth per annum for yearly data in case of linear trend.

**Demerits of Least Squares Method**

Some of the demerits of least squares are as follows,

1. Fresh calculations become necessary even if a single new observation is added.
2. Calculations required in this method are quite tedious and time consuming as compared with other methods.
3. Future predictions based on this method completely ignore the cyclical, seasonal and erratic fluctuations.
4. This method cannot be used to fit growth curves, gomper  $t_z$  curve, logistic curve etc. to which most of the business and economic time series conform.

**PROBLEMS**

31. Obtain the straight line trend equation for the following data by the method of the least square. Tabulate the trend values.

Year :	2010	2011	2012	2013	2014	2015	2016
Sale (in '000 units)	140	144	160	152	168	176	180

*Sol :*

$$Y_c = a + bx$$

$$a = \frac{\sum y}{n} \quad b = \frac{\sum xy}{\sum x^2}$$

**Fitting of straight line trend**

Year	Sales	x	xy	x <sup>2</sup>	y <sub>c</sub> = a + bx
2010	140	-3	-420	9	139.42
2011	144	-2	-288	4	146.28
2012	160	-1	-160	1	153.14
2013	152	0	0	0	160
2014	168	1	168	1	166.86
2015	176	2	352	4	173.72
2016	180	3	540	9	180.58
	1120		192	28	

$$a = \frac{1120}{7} = 160$$

$$b = \frac{192}{28} = 6.86$$

$$y_c = 160 + 6.86(x)$$

$$\text{For } 2010 = 160 + 6.86(-3)$$

$$160 - 20.58 = 139.42$$

$$\text{For } 2011 = 160 + 6.86(-2)$$

$$160 - 13.72 = 146.28$$

$$\text{For } 2012 = 160 + 6.86(-1)$$

$$160 - 6.86 = 153.14$$

$$\text{For } 2013 = 160 + 6.86(0)$$

$$160 - 0 = 160$$

$$\text{For } 2014 = 160 + 6.86(1)$$

$$160 + 6.86 = 166.86$$

$$\text{For } 2015 = 160 + 6.86(2)$$

$$160 + 13.72 = 173.72$$

$$\text{For } 2016 = 160 + 6.86(3)$$

$$= 160 + 20.58$$

$$= 180.58$$

32. Fit a straight line by the Least Square Method and tabulate the trend values for the above data.

Year	2011	2012	2013	2014	2015	2016	2017
Production (in tons)	77	88	94	85	91	98	90

Sol.:

#### Fitting of Straight line trend

Years	Y	Years - Middle Year = X Years - 2014 = X	X <sup>2</sup>	XY	Y <sub>c</sub>
2011	77	-3	9	-231	83
2012	88	-2	4	-176	85
2013	94	-1	1	-94	87
2014	85	0	0	0	89
2015	91	+1	1	91	91
2016	98	+2	4	196	93
2017	90	+3	9	270	95
	623	0	28	56	623

$$y_c = a + bx$$

$$a = \frac{\Sigma y}{N} = \frac{623}{7} = 89$$

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{56}{28} = 2$$

### Calculation of Trend Values ( $Y_c$ )

$$2011 Y_c = a + bx = 89 + [2] [-3] = 89 - 6 = 83$$

$$2012 Y_c = a + bx = 89 + [2] [-2] = 89 - 4 = 85$$

$$2013 Y_c = a + bx = 89 + [2] [-1] = 89 - 2 = 87$$

$$2014 Y_c = a + bx = 89 + [2] [0] = 89 + 0 = 89$$

$$2015 Y_c = a + bx = 89 + [2] [1] = 89 + 2 = 91$$

$$2016 Y_c = a + bx = 89 + [2] [2] = 89 + 4 = 93$$

$$2017 Y_c = a + bx = 89 + [2] [3] = 89 + 6 = 95$$

### 33. Production figure of a Textile Industry are as follows.

Year	2011	2012	2013	2014	2015	2016	2017
Production (in '000 units)	12	10	14	11	13	15	16

For the above data

- Determine the straight line equation under the Least Square Method.
- Find the Trend Values and show the trend line on a graph paper.

*Sol:*

#### Fitting of straight line equation

Year	Production (y)	x	$x^2$	xy	Trend Values $y_c = a + bx$
2011	12	-3	9	-36	10.75
2012	10	-2	4	-20	11.5
2013	14	-1	1	-14	12.25
2014	11	0	0	0	13
2015	13	1	1	13	13.75
2016	15	2	4	30	14.5
2017	16	3	9	48	15.25
N = 7	91	0	28	21	

We know, Trend value,

$$y_c = a + bx$$

$$a = \frac{\Sigma y}{n} = \frac{91}{7} = 13$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{21}{28} = 0.75$$

For 2011

$$\begin{aligned} & 13 + 0.75 (-3) \\ &= 13 - 2.25 \\ &= 10.75. \end{aligned}$$

For 2012

$$\begin{aligned} & 13 + 0.75 (-2) \\ &= 13 - 1.5 \\ &= 11.5 \end{aligned}$$

For 2013

$$\begin{aligned} & 13 + 0.75 (-1) \\ &= 13 - 0.75 \\ &= 12.25 \end{aligned}$$

For 2014

$$\begin{aligned} & 13 + 0.75 (0) \\ &= 13 + 0 \\ &= 13 \end{aligned}$$

For 2015

$$\begin{aligned} & 13 + 0.75 (1) \\ &= 13 + 0.75 \\ &= 13.75 \end{aligned}$$

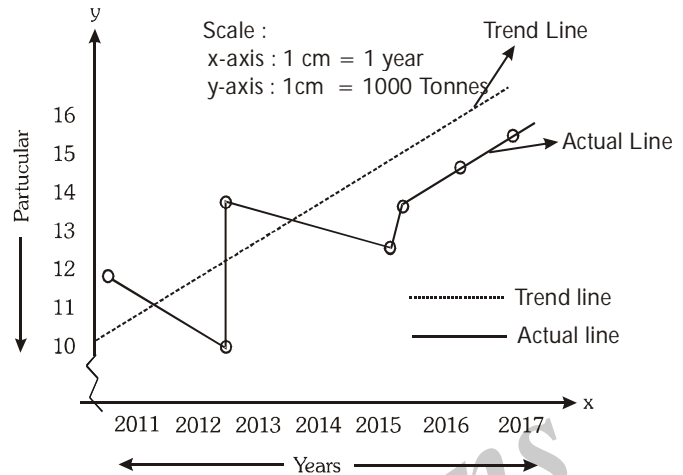
For 2016

$$\begin{aligned} & 13 + 0.75 (2) \\ &= 13 + 1.5 \\ &= 14.5 \end{aligned}$$

For 2017

$$\begin{aligned} & 13 + 0.75 (3) \\ &= 13 + 2.25 \\ &= 15.25 \end{aligned}$$

ii) Find the Trend values and show the trend Line on a graph Paper.





## Short Question and Answers

### 1. Linear Regression.

*Ans :*

Linear regression is a form of regression which is used for modeling the relationship between scalar variables like X and F under linear regression, linear functions are used to model the data and the unknown parameters, of models are estimated from the data. Hence, these models are known as linear models.

Linear models more commonly refers to those models, where the conditional mean of variable 'F for a given value of variable X will be an affine function of X. A linear regression may also refer to a model, where median or other quantile of the conditional distribution of 'F for a given value of 'X is termed as linear function of X. Similar, to all types of regression analysis, linear regression also aims on the conditional probability distribution of 'F for a given 'X, instead of joint probability distribution of 'F and X.

### 2. What is Rank correlation ?

*Ans :*

The Karl Pearson's method is based on the assumption that the population being studied is normally distributed. When it is known that the population is not normal or when the shape of the distribution is not known, there is need for a measure of correlation that involves no assumption about the parameter of the population.

It is possible to avoid making any assumptions about the populations being studied by ranking the observations according to size and basing the calculations on the ranks rather than upon the original observations. It does not matter which way the items are ranked, item number one may be the largest or it may be the smallest. Using ranks rather than actual observations gives the coefficient of rank correlation.

This method of finding out covariability or the lack of it between two variables was developed by the British Psychologist Charles Edward Spearman in 1904.

### Properties

1. It is based on subjective ranking of variables.
2. It lies between -1
3. It measures both direction as well as degree of change. If  $r$  is less than 0, there is negative correlation, which means the direction of change of the two variables will be opposite. If  $r$  is more than 0, there is positive correlation, which means the direction of change of the two variables will be same. Higher the value of  $r$ , greater is the degree of correlation.
4. It is independent of direction. In other words, Correlation of X and Y is same as Correlation of Y and X.
5. It is a pure number without any units. In other words, it is independent of the unit of measurement of the 2 variables.
6. It is not impacted by extreme values as only ranking matters.

### 3. Significance of Measuring Correlation.

*Ans :*

1. Correction is very useful to economists to study the relationship between variables, like price and quantity demanded. To businessmen, it helps to estimate costs, sales, price and other related variables.
2. In economic theory we come across several types of variables which show some kind of relationship. For example, there exists a relationship between price, supply and quantity demanded; convenience, amenities, and service standards are related to customer retention; yield a crop related to quantity of fertilizer applied, type of soil, quality of seeds, rainfall and so on. Correlation analysis helps in measuring the degree of association and direction of such relationship.
3. The relation between variables can be verified and tested for significance, with the help of the correlation analysis. The effect of correlation is to reduce the range of uncertainty of our prediction.

**4. What are the differences between Correlation and Regression.***Ans :*

S.No.	Basis for Comparison	Correlation	Regression
1.	Meaning	Correlation is a statistical measure which determines co-relationship or association of two variables.	Regression describe how an independent variable is numerically related to the dependent variable.
2.	Usage	To represent linear relationship between two variables.	To fit a best line and estimate one variable on the basis of another variable.
3.	Dependent and Independent variables	No difference	Both variables are different
4.	Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (Y).

**5. What do you mean by correlation?***Ans :***Meaning**

Correlation is the study of the linear relationship between two variables. When there is a relationship of 'quantitative measure between two set of variables, the appropriate statistical tool for measuring the relationship and expressing each in a precise way is known as correlation.

**For example**, there is a relationship between the heights and weights of persons, demand and prices of commodities etc.

Correlation analysis is the statistical tool we can use to describe the degree to which one variable is linearly related to another.

**Definitions**

- (i) **According to L.R. Connor** "If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in other(s) then they are said to be correlated."
- (ii) **According to A.M. Tuttle** "Correlation is an analysis of covariation between two or more variables".
- (iii) **According to Croxton and Cowden** "When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation."
- (iv) **According to Ya Lun Chou** "Correlation analysis attempts to determine the 'degree of relationship' between variable".

**6. Scatter Diagram.***Ans :*

Scatter diagram method is the simplest way of diagrammatic representation of a bivariate distribution and helps in ascertaining the correlation between the two variables under study i.e., it portrays the relationship between these two variables graphically.

**Procedure of Scatter Diagram**

Given pair of values  $(x_1, y_1)$   $(x_2, y_2)$ ...  $(x_n, y_n)$  of two variables X and Y. Take the independent variables on the X axis and the dependent variable on the Y-axis. The V points denoted by the pair of values are plotted on the graph. The diagram of dots thus obtained is the scatter diagram. Regarding the correlation between the two variables, the scatter diagram can be interpreted as follows,

- (i) If the points reveal any upward or downward trend, the variables are said to be correlated, otherwise uncorrelated.
- (ii) If the points are very close to each other, a good amount of correlation exists, else poor correlation exists.
- (iii) Upward trend indicates positive correlation and downward trend indicates negative correlation.

**7. State the Merits and Demerits of Scatter Diagram.***Ans :***Merits**

- Scatter diagram is a simple and attractive method of finding out the nature of correlation between two variables.
- It is a non-mathematical method of studying correlation. It is easy to understand.
- We can get a rough idea at a glance whether it is a positive or negative correlation.
- It is not influenced by extreme items.
- It is a first step in finding out the relationship between two variables.

**Demerits**

- The major limitation of the method is that it only gives a visual picture of the relationship of two variables. It only tells us whether there is correlation between the variables, and if so, then in which direction, positive (or) negative.
- It does not give an idea about the precise degree of relationship as it is not amenable to mathematical treatment.

**8. State the Merits and Demerits of rank correlation.***Ans :***Merits**

1. It is easy to understand and calculate
2. It is not impacted by extreme values.
3. It is a numerical measure and provides objectivity to subjective ranking.
4. It is the only method of finding correlation with respect to qualitative factors such as honesty, beauty, etc.
5. It measures both direction as well as degree of change.
6. It facilitates comparisons between two series.
7. It can be applied to irregular data also.
8. It is ideal when the number of observations is very small.

**Demerits**

1. It cannot be applied to grouped data
2. It lacks the precision of Karl Pearson's Coefficient of Correlation.
3. All the information concerning the variable is not used.

**9. Multiple Correlation***Ans :*

In multiple correlations we study together the relationship between three or more factors like production, rainfall and use of fertilizers.

**10. Partial Correlation***Ans :*

In partial correlation though more than two factors are involved but correlation is studied only between two factors and the other factors are assumed to be constant.

**For example,** If out of the three related variables, say, marks in statistics, marks in accountancy, and marks in English, we study the correlation between the two variables, viz., marks in statistics, and marks in Accountancy ignoring the effect of the other variable, i.e., marks in English, it will be a case of partial correlation. On the other hand, when the relationship between any two or more variables is studied at a time, it is a case of multiple correlation. If the relationship between the volume of profits, volume of sales, and the volume of cost of sales at a time are studied it will be a case of multiple correlation. In actual practice, however, the study of multiple correlation is not popular.

**11. Define Regression?***Ans :*

Regression analysis which confines itself to a study of only two variables is called simple regression. The regression analysis which studies more than two variables at a time is called multiple regression. In the simple regression analysis there are two variables-one of which is known as 'independent variable' or 'regressor' or 'predictor'. On the basis of the values of this variable the values of the other variable are predicted. The other variable whose values are predicted is called the 'dependent' or 'regressed' variable.

**Definitions**

1. "Regression is the measure of the average relationship between two or more variables in terms of the original units of the data."
2. **According to Morris Hamburg** The term 'regression analysis' refers to the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process."

3. **According to Taro Yamane** "One of the most frequently used techniques in economics and business research, to find a relation between two or more variables that are related causally, is regression analysis."

4. **According to YaLum Chou** "Regression analysis attempts to establish the 'nature of the relationship between variables that is, to study the functional relationship between the variables and thereby provide a mechanism for prediction, or forecasting."

**12. What is the importance of regression analysis ?***Ans :*

1. Regression analysis helps in establishing a functional relationship between two or more variables. Once this is established it can be used for various advanced analytical purposes.
2. Since most of the problems of economic analysis are based on cause and effect relationship, the regression analysis is a highly valuable tool in economics and business research.
3. This can be used for prediction or estimation of future production, prices, sales, investments, income, profits and population which are indispensable for efficient planning of an economy and are of paramount importance to a businessman or an economist.
4. Regression analysis is widely used in statistical estimation of demand curves, supply curves, production functions, cost functions, consumption functions, etc. Economists have discovered many types of production functions by fitting regression lines to input and output data.

**13. What are the limitations of regression analysis ?***Ans :*

1. It assumes a linear relationship between two variables which need not be the case always.
2. It assumes a static relationship between the two variables over a period of time. However,

relationships between variables can change with a change in other factors. For example, the change in demand for a given change in price can be estimated using regression. However, the impact of price on demand will be different when a family or a nation is poor and when such a family or nation has abundance of wealth or resources.

**14. What are the properties of regression coefficient?**

*Ans :*

- (i)  $r^2 = b_{xy} * b_{yx}$  In other words,  $r$  is the Geometric mean between the two regression coefficients  $b_{xy}$  and  $b_{yx}$
- (ii) Both the regression coefficients will have the same sign, i.e. either they will be positive or negative. Also, the coefficient of correlation will have the same sign as that of regression coefficients
- (iii) The arithmetic mean of the two regression coefficients is greater than the correlation coefficient. In other words,  $(b_{xy} + b_{yx})/2 > r$ .
- (iv) If one regression coefficient is greater than 1, the other has to be less than one. This is an extension of the first property as the product of the two coefficients is equal to square of the correlation coefficient  $r$ . Since  $r$  lies between  $-1$  and  $+1$ ,  $r^2$  cannot be greater than 1. Thus,  $b_{xy} * b_{yx}$  cannot be greater than 1. Thus, if one regression coefficient is greater than 1, the other has to be less than one.

**15. Define time series.**

*Ans :*

**Meaning**

A time series is a statistical data that are collected, observed or recorded at regular intervals of time. The term time series applies, for example, to the data recorded periodically showing the total annual sales of retail stores, the total quarterly value of construction contracts awarded, the total amount of unfilled orders in durable goods industries at the end of each month, weekly earnings of workers in an industrial town, hourly temperature in a particular city.

**Definitions**

Some of the important definitions of time series, given by different experts are as under:

- i) **According to Morris Hamburg**, "A time series is a set of statistical observations arranged in chronological order."
- ii) **According to Patterson**, "A time series consists of statistical data which are collected, recorded observed over successive increments."
- iii) **According to Ya-Lun-Chou**, "A time series may be defined as a collection of magnitudes belonging to different time periods, of some variable or composite of variables, such as production of steel, per capita income, gross national product, price of tobacco, or index of industrial production."
- iv) **According to Wessel and Wellet**, "When quantitative data are arranged in the order of their occurrence, the resulting statistical series is called a time series."
- v) **According to Spiegel**, "A time series is a set of observations taken at specified times, usually at 'equal intervals'. Mathematically, a time series is defined by the values  $Y_1, Y_2, \dots$  of a variable  $Y$  (temperature, closing price of a share, etc.) at times  $t_1, t_2, \dots$ . Thus  $Y$  is a function of  $t$ , symbolized by  $Y = F(t)$ ."

**16. What are the characteristics of time series?**

*Ans :*

The essentials of Time Series are

- i) It must consist of a set of values that are homogeneous. For example, production data for a year and sales data for the next year will not be a time series.
- ii) The values must be with reference to time. In other words, in a time series, we have at least 2 variables, with one variable necessarily being time.
- iii) The data must be available for a reasonably long period of time.
- iv) The gaps between various time values should as far as possible be equal.

**17. What is Trend Analysis?***Ans :*

The trend analysis refers to the concept of collecting information and attempting to spot a pattern, or trend, in the information. In some fields of study, the term "trend analysis" has more formally-defined meanings.

Although trend analysis is often used to predict future events, it could be used to estimate uncertain events in the past, such as how many ancient kings probably ruled between two dates, based on data such as the average years which other known kings reigned.

Trend analysis uses a technique called least squares to fit a trend line to a set of time series data and then project the line into the future for a forecast.

It is a special case of regression analysis, where the dependent variable is the variable to be forecasted and the independent variable is time.

**18. Least Square Methods***Ans :*

Least square method is the most widely used method and provides us with a mathematical device to obtain an objective fit to the trend of a given time series. This method is so called because a trend line computed by this method is such that the sum of the squares of the deviation between the original data and the corresponding computed trend values is minimum. This method can be used to fit either a straight line trend or a parabolic trend.

The straight line trend equation is in the form of  $Y = a + bX$

Where, Y denotes the trend value of the dependent variable

X denotes the independent variable.

a and b are constants.

The values of a and b are obtained by solving the following normal equations.

$$\Sigma Y = Na + b \Sigma X$$

$$\Sigma XY = a \Sigma X + b \Sigma X^2$$

Where, N represents the number of years in the series.

When  $\Sigma X = 0$  the above normal equations are simplified to

$$a = \frac{\Sigma Y}{N}$$

$$b = \frac{\Sigma XY}{\Sigma X^2}$$

By substituting a and b values in straight line trend equation  $Y = a + bX$ , we get the straight line equation which can be used for estimation of future values.

**19. Merits of Least Squares Method.***Ans :*

1. The method of least squares is a mathematical method of measuring trend and is free from subjectiveness.
2. This method provides the line of best fit since it is this line from where the sum of positive and negative deviations is zero and the sum of square of deviations is the least.
3. This method enables us to compute the trend values for all the given time periods in the series.
4. The trend equation can be used to estimate the values of the variable for any given time period 'f' in future and the forecasted values are quite reliable.
5. This method is the only technique which enables us to obtain the rate of growth per annum for yearly data in case of linear trend.

**20. Demerits of Least Squares Method***Ans :*

Some of the demerits of least squares are as follows,

1. Fresh calculations become necessary even if a single new observation is added.
2. Calculations required in this method are quite tedious and time consuming as compared with other methods.
3. Future predictions based on this method completely ignore the cyclical, seasonal and erratic fluctuations.
4. This method cannot be used to fit growth curves, gomper  $t_2$  curve, logistic curve etc. to which most of the business and economic time series conform.

## Exercise Problems

1. Given the two regression Coefficient X on Y = +0.542 and Y on X = +0.905. Calculate the coefficient of correlation between X and Y.

**[Ans : +0.70]**

2. If the regression Coefficient of X on Y = 0.847 and Y on X = +0.732, find coefficient of correlation.

**[Ans : +0.787]**

3. If Regression equation of X on Y = 0.268 and of Y on X = 0.5, find coefficient of correlation.

**[Ans : +0.259]**

4. Fit a straight line regression equation of Y on X from the following data.

X	10	12	13	16	17	20	25	29
Y	10	12	24	27	29	33	37	42

**[Ans : Y = 1.6 X - 1.65]**

5. Find the two regression equations from the following data :

X	1	2	3	4	5
Y	2	3	5	4	6

**[Ans : X = 0.9 Y - 0.6; Y = 0.9 X + 1.3]**

6. From the following data, obtain the two regression lines.

X	2	6	8	11	13	13	13	14
Y	8	6	10	12	12	14	14	20

**[Ans : X = 0.8125Y + 0.25; Y = 0.8125 X + 3.875]**

7. From the following data, find out two regression equations.

X	32	46	57	65	55	72	80	67	75
Y	100	95	87	110	95	92	88	90	85

**[Ans : X = -0.8134 Y + 137.0981; Y = -0.21X + 106.40]**

8. Calculate the 3 yearly moving averages to determine trend component of the following time series.

Year :	1973	1974	1975	1976	1977	1978	1979
Annual sales :	2	6	1	5	3	7	2

9. Using three year moving averages, determine the trend and short term fluctuations.

Plot the original and trend values on the same graph paper:

Year :	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977
Production:	21	22	23	25	24	22	25	26	27	26

10. Calculate the trend values by the method of least squares from the data given below and estimate the sales for the year 1985.

Year:	1976	1977	1978	1979	1980
Sales of T.V. ('000):	9	12	14	16	20

**[Ans:  $Y_c = 20 + 5x$ , 44.5]**

11. Fit a straight line to the following data and compute the trend values

Year:	1960	1961	1962	1963	1964	1965	1966
Sales:	3	7	6	8	9	7	10

(tons in thousands)

**[Ans :  $Y_c = 7.14 + 0.85x$ ]**

12. Fit a straight line trend by method of least squares to the following data:

Year:	1981	1982	1983	1984	1985	1986	1987
Profit (Rs. '000):	57	65	63	72	69	78	82

**[Ans:  $Y_c = 69.43 + 3.82x$ ]**



## *Choose the Correct Answer*

1. The regression equation of X on Y gives [ a ]  
(a) The most probable values of X for given values of Y  
(b) The most probable values of Y for given values of X  
(c) Either of the two  
(d) None
2. The concept of regression was given by [ b ]  
(a) Sir Francis Galton in 1807 (b) Sir Francis Galton in 1877  
(c) Sir Hanrey Fayol in 1854 (d) Charles Babbage in 1867.
3. Regression is : [ c ]  
(a) Measures of average relationship between two more variables  
(b) To find a relation between two or more variables that are related casually.  
(c) Both of the above  
(d) None of these.
4. Number of observations in regression analysis is considered as [ d ]  
(a) Degree of possibility (b) Degree of average  
(c) Degree of variance (e) Degree of freedom
5. If all conditions or assumptions of regression analysis simple regression can give [ c ]  
(a) Dependent estimation (b) Independent estimation  
(c) Reliable estimates (d) Unreliable estimates
6. In Regression Analysis, testing of assumptions if these are true or not is classified as [ d ]  
(a) Weighted analysis (b) Average analysis  
(c) Significance analysis (d) Specification analysis
7. A process by which we estimate the value of dependent variable on the basis of one or more independent variables is called: [ b ]  
(a) Correlation (b) Regression  
(c) Residual (d) Slope
8. The method of least squares dictates that we choose a regression line where the sum of the square of deviations of the points from the lie is: [ b ]  
(a) Maximum (b) Minimum  
(c) Zero (d) Positive

9. If one regression coefficient is greater than one, then other will be: [ c ]  
(a) More than one (b) Equal to one  
(c) Less than one (d) Equal to minus one
10. The dependent variable is also called: [ d ]  
(a) Regress and variable (b) Predict and variable  
(c) Explained variable (d) All of these
11. The first step in time-series analysis is to [ c ]  
(a) Perform preliminary regression calculations  
(b) Calculate a moving average  
(c) Plot the data on a graph  
(d) Identify relevant correlated variables
12. Time-series analysis is based on the assumption that [ a ]  
(a) Random error terms are normally distributed  
(b) There are dependable correlations between the variable to be forecast and other independent variables  
(c) Past patterns in the variable to be forecast will continue unchanged into the future  
(d) The data do not exhibit a trend
13. The cyclical component of time-series data is usually estimated using [ d ]  
(a) Linear regression analysis (b) Moving averages  
(c) Exponential smoothing (d) Qualitative methods
14. In time-series analysis, which source of variation can be estimated by the ratio-to-trend method? [ c ]  
(a) Cyclical (b) Trend  
(c) Seasonal (d) Irregular
15. If regression analysis is used to estimate the linear relationship between the natural logarithm of the variable to be forecast and time, then the slope estimate is equal to [ c ]  
(a) The linear trend  
(b) The natural logarithm of the rate of growth  
(c) The natural logarithm of one plus the rate of growth  
(d) The natural logarithm of the square root of the rate of growth

## *Fill in the Blanks*

1. The sign of regression coefficient is \_\_\_\_\_ as that of correlation coefficient.
2. The regression analysis measures \_\_\_\_\_ between X and Y.
3. The purpose of regression analysis is to study \_\_\_\_\_ between variation.
4. When one regression coefficient is positive the other would be \_\_\_\_\_
5. Lines of regression are \_\_\_\_\_ if  $r = 0$ , and they are \_\_\_\_\_ if  $r = \pm 1$
6. The farther the two regression lines cut each other the ..... be the degree of correlation.
7. Trend projection is an example of \_\_\_\_\_
8. In time series seasonal variations can occur within a period of \_\_\_\_\_
9. A single-equation econometric model of the demand for a product is a \_\_\_\_\_ equation in which the quantity demanded of the product is an \_\_\_\_\_ variable
10. Barometric methods are used to forecast \_\_\_\_\_
11. A \_\_\_\_\_ is a statistical data that are collected, observed or recorded at regular intervals of time.

### ANSWERS

1. Same
2. Average relationship
3. Dependence
4. Positive
5. Perpendicular, same
6. Lesser
7. Time-series
8. One year
9. Structural, endogenous
10. Cyclical variation
11. Time series

## LOGARITHMS

	0	1	2	3	4	5	6	7	8	9	Mean Difference								
											1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	6435	6445	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1	2	2	3	4	5	6	7	7
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

## LOGARITHMS

	0	1	2	3	4	5	6	7	8	9	Mean Difference								
											1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	7782	7789	7769	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1	1	2	2	3	4	4	5	6
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	5
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1	2	2	3	4	4	5	5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	5
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0	1	1	2	2	3	3	4	4
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0	1	1	2	2	3	3	4	4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	9912	9917	9921	9926	9930	9934	9939	9843	9948	9952	0	1	1	2	2	3	3	4	4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0	1	1	2	2	3	3	3	4
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

## ANTILOGARITHMS

	0	1	2	3	4	5	6	7	8	9	Mean Difference								
											1	2	3	4	5	6	7	8	9
.00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021	0	0	1	1	1	1	2	2	2
.01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045	0	0	1	1	1	1	2	2	2
.02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069	0	0	1	1	1	1	2	2	2
.03	1072	1074	1076	1079	1081	1084	1086	1089	1091	1094	0	0	1	1	1	1	2	2	2
.04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119	0	1	1	1	1	2	2	2	2
.05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146	0	1	1	1	1	2	2	2	2
.06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172	0	1	1	1	1	2	2	2	2
.07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199	0	1	1	1	1	2	2	2	2
.08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227	0	1	1	1	1	2	2	2	3
.09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256	0	1	1	1	1	2	2	2	3
.10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285	0	1	1	1	1	2	2	2	3
.11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315	0	1	1	1	2	2	2	2	3
.12	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	0	1	1	1	2	2	2	2	3
.13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	0	1	1	1	2	2	2	3	3
.14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	0	1	1	1	2	2	2	3	3
.15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442	0	1	1	1	2	2	2	3	3
.16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476	0	1	1	1	2	2	2	3	3
.17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510	0	1	1	1	2	2	2	3	3
.18	1514	1517	1521	1524	1528	1531	1535	1538	1542	1545	0	1	1	1	2	2	2	3	3
.19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581	0	1	1	1	2	2	3	3	3
.20	1585	1589	1592	1596	1600	1603	1607	1611	1614	1618	0	1	1	1	2	2	3	3	3
.21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656	0	1	1	2	2	2	3	3	3
.22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694	0	1	1	2	2	2	3	3	3
.23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734	0	1	1	2	2	2	3	3	4
.24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774	0	1	1	2	2	2	3	3	4
.25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816	0	1	1	2	2	2	3	3	4
.26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858	0	1	1	2	2	3	3	3	4
.27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901	0	1	1	2	2	3	3	3	4
.28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945	0	1	1	2	2	3	3	4	4
.29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991	0	1	1	2	2	3	3	4	4
.30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037	0	1	1	2	2	3	3	4	4
.31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084	0	1	1	2	2	3	3	4	4
.32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133	0	1	1	2	2	3	3	4	4
.33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183	0	1	1	2	2	3	3	4	4
.34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234	1	1	2	2	3	3	4	4	5
.35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286	1	1	2	2	3	3	4	4	5
.36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339	1	1	2	2	3	3	4	4	5
.37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393	1	1	2	2	3	3	4	4	5
.38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449	1	1	2	2	3	3	4	4	5
.39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506	1	1	2	2	3	3	4	5	5
.40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564	1	1	2	2	3	4	4	5	5
.41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624	1	1	2	2	3	4	4	5	5
.42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685	1	1	2	2	3	4	4	5	6
.43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748	1	1	2	3	3	4	4	5	6
.44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812	1	1	2	3	3	4	4	5	6
.45	2818	2825	2831	2838	2844	2851	2858	2864	2871	2877	1	1	2	3	3	4	5	5	6
.46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944	1	1	2	3	3	4	5	5	6
.47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013	1	1	2	3	3	4	5	5	6
.48	3020	3027	3034	3041	3048	3055	3062	3069	3076	3083	1	1	2	3	3	4	5	6	6
.49	3090	3097	3105	3112	3119	3126	3133	3141	3148	3155	1	1	2	3	4	4	5	6	6
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

## ANTILOGARITHMS

	0	1	2	3	4	5	6	7	8	9	Mean Difference								
											1	2	3	4	5	6	7	8	9
.50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228	1	1	2	3	4	4	5	6	7
.51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304	1	2	2	3	4	5	5	6	7
.52	3311	3319	3327	3334	3342	3350	3357	3365	3373	3381	1	2	2	3	4	5	5	6	7
.53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459	1	2	2	3	4	5	6	6	7
.54	3467	3475	3483	3491	3499	3508	3516	3524	3532	3540	1	2	2	3	4	5	6	6	7
.55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622	1	2	2	3	4	5	6	7	7
.56	3631	3639	3648	3656	3664	3673	3681	3690	3698	3707	1	2	3	3	4	5	6	7	8
.57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793	1	2	3	3	4	5	6	7	8
.58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882	1	2	3	4	4	5	6	7	8
.59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972	1	2	3	4	5	5	6	7	8
.60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064	1	2	3	4	5	6	6	7	8
.61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159	1	2	3	4	5	6	7	8	9
.62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256	1	2	3	4	5	6	7	8	9
.63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355	1	2	3	4	5	6	7	8	9
.64	4365	4375	4385	4395	4406	4416	4426	4436	4446	4457	1	2	3	4	5	6	7	8	9
.65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560	1	2	3	4	5	6	7	8	9
.66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667	1	2	3	4	5	6	7	9	10
.67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775	1	2	3	4	5	7	8	9	10
.68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887	1	2	3	4	6	7	8	9	10
.69	4898	4909	4920	4932	4943	4955	4966	4977	4989	5000	1	2	3	5	6	7	8	9	10
.70	5012	5023	5035	5047	5058	5070	5082	5093	5105	5117	1	2	4	5	6	7	8	9	11
.71	5129	5140	5152	5164	5176	5188	5200	5212	5224	5236	1	2	4	5	6	7	8	10	11
.72	5248	5260	5272	5284	5297	5309	5321	5333	5346	5358	1	2	4	5	6	7	9	10	11
.73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483	1	3	4	5	6	8	9	10	11
.74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610	1	3	4	5	6	8	9	10	12
.75	5623	5636	5649	5662	5675	5689	5702	5715	5728	5741	1	3	4	5	7	8	9	10	12
.76	5754	5768	5781	5794	5808	5821	5834	5848	5861	5875	1	3	4	5	7	8	9	11	12
.77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6012	1	3	4	5	7	8	10	11	12
.78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6152	1	3	4	6	7	8	10	11	13
.79	6166	6180	6194	6209	6223	6237	6252	6266	6281	6295	1	3	4	6	7	9	10	11	13
.80	6310	6324	6339	6353	6368	6383	6397	6412	6427	6442	1	3	4	6	7	9	10	12	13
.81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592	2	3	5	6	8	9	11	12	14
.82	6607	6622	6637	6653	6668	6683	6699	6715	6730	6745	2	3	5	6	8	9	11	12	14
.83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902	2	3	5	6	8	9	11	13	14
.84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063	2	3	5	6	8	10	11	13	15
.85	7079	7096	7112	7129	7145	7161	7178	7194	7211	7228	2	3	5	7	8	10	12	13	15
.86	7244	7261	7278	7295	7311	7328	7345	7362	7379	7396	2	3	5	7	8	10	12	13	15
.87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568	2	3	5	7	9	10	12	14	16
.88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745	2	4	5	7	9	11	12	14	16
.89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925	2	4	5	7	9	11	12	14	16
.90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110	2	4	6	7	9	11	13	15	17
.91	8128	8147	8166	8185	8204	8222	8241	8260	8279	8299	2	4	6	8	9	11	13	15	17
.92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492	2	4	6	8	10	12	14	15	17
.93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690	2	4	6	8	10	12	14	16	18
.94	8710	8730	8750	8770	8790	8810	8831	8851	8872	8892	2	4	6	8	10	12	14	16	18
.95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099	2	4	6	8	10	12	15	17	19
.96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311	2	4	6	8	11	13	15	17	19
.97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528	2	4	7	9	11	13	15	17	20
.98	9550	9572	9594	9616	9638	9661	9683	9705	9727	9750	2	4	7	9	11	13	16	18	20
.99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977	2	5	7	9	11	14	16	18	20
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

## BINOMIAL COEFFICIENTS

$n$	$\binom{n}{0}$	$\binom{n}{1}$	$\binom{n}{2}$	$\binom{n}{3}$	$\binom{n}{4}$	$\binom{n}{5}$	$\binom{n}{6}$	$\binom{n}{7}$	$\binom{n}{8}$	$\binom{n}{9}$	$\binom{n}{10}$
0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1
11	1	11	55	165	330	462	462	330	165	55	11
12	1	12	66	220	495	792	924	792	495	220	66
13	1	13	78	286	715	1287	1716	1716	1287	715	286
14	1	14	91	364	1001	2002	3003	3432	3003	2002	1001
15	1	15	105	455	1365	3003	5005	6435	6435	3005	3003
16	1	16	120	560	1820	4368	8008	11440	12870	11440	8008
17	1	17	136	680	2380	6188	12376	19448	24310	24310	19448
18	1	18	153	816	3060	8568	18564	31824	43758	48620	43758
19	1	19	171	969	3876	11628	27132	50388	75582	92378	92378
20	1	20	190	1140	4845	15504	38760	77520	125970	167960	184756



### VALUES OF $e^{-m}$ (For Computing Poisson Probabilities) ( $0 < m < 1$ )

m	0	1	2	3	4	5	6	7	8	9
0.0	1.0000	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
0.1	0.9048	.8958	.8860	.8781	.8694	.8607	.8521	.8437	.8353	.8270
0.2	0.8187	.8106	.8025	.7945	.7866	.7788	.7711	.7634	.7558	.7483
0.3	0.7408	.7334	.7261	.7189	.7118	.7047	.6977	.6907	.6839	.6771
0.4	0.6703	.6636	.6570	.6505	.6440	.6376	.6313	.6250	.6188	.6126
0.5	0.6065	.6005	.5945	.5886	.5827	.5770	.5712	.5655	.5599	.5543
0.6	0.5488	.5434	.5379	.5326	.5278	.5220	.5160	.5117	.5066	.5016
0.7	0.4966	.4916	.4868	.4810	.4771	.4724	.4670	.4630	.4584	.4538
0.8	0.4493	.4449	.4404	.4360	.4317	.4274	.4232	.4190	.4148	.4107
0.9	0.4066	.4025	.3985	.3946	.3906	.3867	.3829	.3791	.3753	.3716

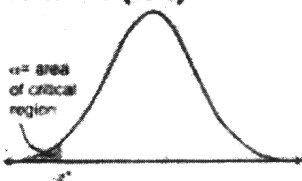
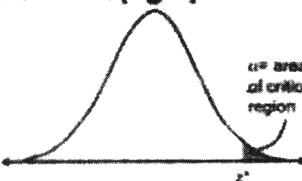
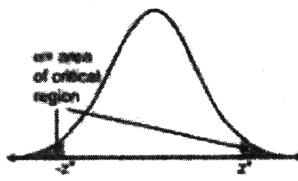
( $m = 1, 2, 3, \dots, 10$ )

m	1	2	3	4	5	6	7	8	9	10
$e^{-m}$	.36788	.13534	.04979	.01832	.00698	.00279	.00092	.000395	.000123	.000045

**Note:** To obtain values of  $e^{-m}$  for other values of  $m$ , use the laws of exponents.

**Example.**  $e^{-2.35} = (e^{-2.00}) (e^{-0.35}) = (.13534) (.7047) = .095374$

### Z CRITICAL VALUES

$\alpha$ level	One-tailed (left)	One-tailed (right)	Two-tailed
			
$\alpha = 0.05$	$z = -1.64$	$z = 1.64$	$z = \pm 1.96$
$\alpha = 0.01$	$z = -2.33$	$z = 2.33$	$z = \pm 2.57$
$\alpha = 0.001$	$z = -3.08$	$z = 3.08$	$z = \pm 3.32$

## AREAS UNDER THE NORMAL CURVE

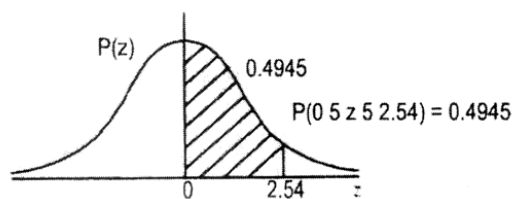
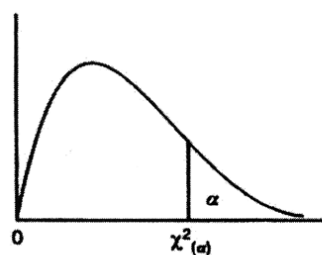


Table of Area

	0	1	2	3	4	5	6	7	8	9
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2223
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
.7	.2580	.2612	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4849	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4965	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4947	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4898	.4898
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

## VALUES OF CHI-SQUARE



$\alpha$	.100	.050	.025	.010	.005	.001
$df$						
1	2.71	3.84	5.02	6.63	7.88	10.8
2	4.61	5.99	7.38	9.21	10.6	13.8
3	6.25	7.81	9.35	11.3	12.8	16.3
4	7.78	9.49	11.1	13.3	14.9	18.5
5	9.24	11.1	12.8	15.1	16.7	20.5
6	10.6	12.6	14.4	16.8	18.5	22.5
7	12.0	14.1	16.0	18.5	20.3	24.3
8	13.4	15.5	17.5	20.1	22.0	26.1
9	14.7	16.9	19.0	21.7	23.6	27.9
10	16.0	18.3	20.5	23.2	25.2	29.6
11	17.3	19.7	21.9	24.7	26.8	31.3
12	18.5	21.0	23.3	26.2	28.3	32.9
13	19.8	22.4	24.7	27.7	29.8	34.5
14	21.1	23.7	26.1	29.1	31.3	36.1
15	22.3	25.0	27.5	30.6	32.8	37.7
16	23.5	26.3	28.8	32.0	34.3	39.3
17	24.8	27.6	30.2	33.4	35.7	40.8
18	26.0	28.9	31.5	34.8	37.2	42.3
19	27.2	30.1	32.9	36.2	38.6	43.8
20	28.4	31.4	34.2	37.6	40.0	45.3
21	29.6	32.7	35.5	38.9	41.4	46.8
22	30.8	33.9	36.8	40.3	42.8	48.3
23	32.0	35.2	38.1	41.6	44.2	49.7
24	33.2	36.4	39.4	43.0	45.6	51.2
25	34.4	37.7	40.6	44.3	46.9	52.6
26	35.6	38.9	41.9	45.6	48.3	54.1
27	36.7	40.1	43.2	47.0	49.6	55.5
28	37.9	41.3	44.5	48.3	51.0	56.9
29	39.1	42.6	45.7	49.6	52.3	58.3
30	40.3	43.8	47.0	50.9	53.7	59.7
35	46.1	49.8	53.2	57.3	60.3	66.6
40	51.8	55.8	59.3	63.7	66.8	73.4
45	57.5	61.7	65.4	70.0	73.2	80.1
50	63.2	67.5	71.4	76.2	79.5	86.7
55	68.8	73.3	77.4	82.3	85.7	93.2
60	74.4	79.1	83.3	88.4	92.0	99.6
65	80.0	84.8	89.2	94.4	98.1	106.0
70	85.5	90.5	95.0	100.4	104.0	112.3
75	91.1	96.2	100.8	106.4	110.3	118.6
80	96.6	101.9	106.6	112.3	116.3	124.8
85	102.1	107.5	112.4	118.2	122.3	131.0
90	107.6	113.1	118.1	124.1	128.3	137.2
95	113.0	118.8	123.9	130.0	134.2	143.3
100	118.5	124.3	129.6	135.8	140.2	149.4

## 5% POINTS OF FISHER'S F-DISTRIBUTION

$\frac{m}{n}$	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	$\infty$
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	250.09	252.20	254.32
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396	19.413	19.420	19.446	19.462	19.479	19.496
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8688	8.8452	8.8123	8.7855	8.7446	8.7029	8.6602	8.6166	8.5720	8.5265
4	7.7044	6.9443	6.5914	6.3883	6.2560	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117	5.8578	5.8025	5.7459	5.6878	5.6281
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8753	4.8183	4.7725	4.7351	4.6777	4.6188	4.5681	4.4957	4.4314	4.3650
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2066	4.1468	4.0990	4.0600	3.9999	3.9381	3.8742	3.8082	3.7398	3.6688
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747	3.5108	3.4445	3.3758	3.3043	3.2298
8	5.3177	4.4590	4.0662	3.8378	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2840	3.2184	3.1503	3.0794	3.0053	2.9276
9	5.1174	4.2565	3.8626	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729	3.0001	2.9365	2.8637	2.7872	2.7007
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130	2.8450	2.7740	2.6996	2.6211	2.5379
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876	2.7186	2.6464	2.5705	2.4901	2.4045
12	4.7272	3.8653	3.4703	3.2392	3.0861	2.9761	2.8934	2.8386	2.7864	2.7434	2.6866	2.6169	2.5436	2.4663	2.3842	2.2962
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.6037	2.5330	2.4589	2.3820	2.2966	2.2064
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6021	2.5342	2.4630	2.3879	2.3082	2.2230	2.1307
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753	2.4035	2.3275	2.2468	2.1601	2.0658
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935	2.4247	2.3522	2.2756	2.1938	2.1058	2.0096
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499	2.3807	2.3077	2.2304	2.1477	2.0584	1.9604
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117	2.3421	2.2686	2.1906	2.1071	2.0166	1.9168
19	4.3808	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779	2.3080	2.2341	2.1555	2.0712	1.9796	1.8780
20	4.3513	3.4928	3.0984	2.8661	2.7100	2.5990	2.5140	2.4471	2.3928	2.3479	2.2776	2.2033	2.1242	2.0391	1.9464	1.8432
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3661	2.3210	2.2504	2.1757	2.0960	2.0102	1.9165	1.8117
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967	2.2258	2.1508	2.0707	1.9842	1.8895	1.7831
23	4.2793	3.4221	3.0280	2.7955	2.6500	2.5377	2.4522	2.3848	2.3301	2.2847	2.2136	2.1382	2.0576	1.9705	1.8749	1.7570
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547	2.1834	2.1077	2.0267	1.9390	1.8424	1.7331
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365	2.1649	2.0889	2.0075	1.9192	1.8217	1.7110
26	4.2252	3.3690	2.9751	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197	2.1479	2.0716	1.9898	1.9010	1.8027	1.6906
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	2.2043	2.1323	2.0558	1.9736	1.8842	1.7851	1.6717
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	2.1900	2.1179	2.0411	1.9586	1.8687	1.7689	1.6541
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2782	2.2229	2.1768	2.1045	2.0275	1.9446	1.8543	1.7537	1.6377
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646	2.0921	2.0148	1.9317	1.8409	1.7396	1.6223
40	4.0848	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772	2.0035	1.9245	1.8389	1.7444	1.6373	1.5089
60	4.0012	3.1504	2.7581	2.5252	2.3688	2.2540	2.1665	2.0970	2.0401	1.9926	1.9174	1.8364	1.7480	1.6491	1.5343	1.3893
120	3.9201	3.0718	2.6802	2.4472	2.2900	2.1750	2.0867	2.0164	1.9588	1.9105	1.8337	1.7505	1.6587	1.5543	1.4290	1.2539
$\infty$	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307	1.7522	1.6664	1.5705	1.4591	1.3180	1.0000

For  $m > 10$  interpolate using 60/ $m$   
 For  $n > 30$  interpolate using 120/ $n$

## 1% POINTS OF FISHER'S F-DISTRIBUTION

$\frac{m}{n}$	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	$\alpha$
1	4052.2	4999.5	5403.3	5624.6	5763.7	5859.0	5928.3	5981.6	6022.5	6055.8	6106.3	6157.3	6208.7	6260.7	6313.0	6366.0
2	98.503	99.000	99.166	99.249	99.299	99.332	99.356	99.374	99.388	99.399	99.416	99.432	99.449	99.466	99.483	99.501
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229	27.052	26.872	26.690	26.505	26.316	26.125
4	21.198	18.000	16.684	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.374	14.198	14.020	13.838	13.652	13.463
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.883	9.722	9.5527	9.3793	9.2020	9.0204
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1016	7.9761	7.8741	7.7183	7.5590	7.3958	7.2285	7.0568	6.8801
7	12.246	9.5466	8.4513	7.8467	7.4604	7.1914	6.9928	6.8401	6.7188	6.6201	6.4691	6.3143	6.1554	5.9921	5.8236	5.6495
8	11.259	8.6491	7.5910	7.0060	6.6318	6.3707	6.1776	6.0289	5.9106	5.8143	5.6668	5.5151	5.3591	5.1981	5.0316	4.8588
9	10.561	8.0215	6.9919	6.4221	6.0569	5.8018	5.6129	5.4671	5.3511	5.2565	5.1114	4.9621	4.8080	4.6486	4.4831	4.3105
10	10.044	7.5594	6.5523	5.9943	5.6363	5.3858	5.2001	5.0567	4.9424	4.8492	4.7059	4.5582	4.4054	4.2469	4.0819	3.9090
11	9.6460	7.2057	6.2167	5.6683	5.3180	5.0692	4.8861	4.7445	4.6315	4.5393	4.3974	4.2509	4.0990	3.9411	3.7761	3.6025
12	9.3302	6.9265	5.9526	5.4119	5.0643	4.8206	4.6395	4.4994	4.3875	4.2961	4.1553	4.0096	3.8584	3.7008	3.5355	3.3608
13	9.0738	6.7010	5.7394	5.2053	4.8616	4.6204	4.4410	4.3021	4.1911	4.1003	3.9603	3.8154	3.6646	3.5070	3.3413	3.1664
14	8.8616	6.5149	5.5639	5.0354	4.6950	4.4558	4.2778	4.1399	4.0297	3.9394	3.8001	3.6557	3.5052	3.3476	3.1813	3.0040
15	8.6831	6.3589	5.4170	4.8932	4.5556	4.3183	4.1415	4.0045	3.8948	3.8049	3.6662	3.5222	3.3719	3.2141	3.0471	2.8684
16	8.5310	6.2262	5.2922	4.7726	4.4374	4.2016	4.0259	3.8896	3.7804	3.6909	3.5527	3.4089	3.2588	3.1007	2.9330	2.7528
17	8.3997	6.1121	5.1850	4.6690	4.3359	4.1015	3.9267	3.7910	3.6822	3.5931	3.4552	3.3117	3.1615	3.0032	2.8348	2.6530
18	8.2854	6.0129	5.0919	4.5790	4.2479	4.0146	3.8406	3.7054	3.5971	3.5082	3.3706	3.2273	3.0771	2.9185	2.7493	2.5660
19	8.1850	5.9259	5.0103	4.5003	4.1708	3.9386	3.7653	3.6305	3.5225	3.4338	3.2965	3.1533	3.0031	2.8442	2.6742	2.4893
20	8.0960	5.8489	4.9382	4.4307	4.1027	3.8714	3.6987	3.5644	3.4567	3.3682	3.2311	3.0880	2.9377	2.7785	2.6077	2.4212
21	8.0166	5.7804	4.8740	4.3688	4.0421	3.8117	3.6396	3.5056	3.3981	3.3098	3.1729	3.0299	2.8796	2.7200	2.5484	2.3603
22	7.9454	5.7190	4.8166	4.3134	3.9880	3.7583	3.5867	3.4530	3.3458	3.2576	3.1209	2.9709	2.8274	2.6675	2.4951	2.3055
23	7.8811	5.6637	4.7649	4.2635	3.9392	3.7102	3.5390	3.4057	3.2986	3.2106	3.0740	2.9311	2.7805	2.6202	2.4471	2.2559
24	7.8229	5.6136	4.7181	4.2184	3.8951	3.6667	3.4959	3.3629	3.2560	3.1681	3.0316	2.8887	2.7390	2.5773	2.4035	2.2107
25	7.7698	5.5680	4.6755	4.1774	3.8550	3.6272	3.4568	3.3239	3.2172	3.1294	2.9931	2.8502	2.6993	2.5383	2.3637	2.1694
26	7.7213	5.5263	4.6366	4.1400	3.8183	3.5911	3.4210	3.2884	3.1818	3.0941	2.9579	2.8150	2.6640	2.5026	2.3273	2.1315
27	7.6767	5.4881	4.6009	4.1056	3.7848	3.5580	3.3882	3.2558	3.1494	3.0618	2.9256	2.7827	2.6316	2.4699	2.2938	2.0965
28	7.6356	5.4529	4.5681	4.0740	3.7539	3.5276	3.3581	3.2259	3.1195	3.0320	2.8959	2.7530	2.6017	2.4397	2.2629	2.0642
29	7.5976	5.4205	4.5378	4.0449	3.7254	3.4995	3.3302	3.1982	3.0920	3.0045	2.8685	2.7256	2.5742	2.4118	2.2344	2.0342
30	7.5625	5.3904	4.5097	4.0179	3.6990	3.4735	3.3045	3.1726	3.0665	2.9791	2.8431	2.7002	2.5487	2.3860	2.2079	2.0062
40	7.3141	5.1785	4.3126	3.8283	3.5138	3.2910	3.1238	2.9930	2.8876	2.8005	2.6648	2.5216	2.3689	2.2034	2.0194	1.8047
60	7.0771	4.9774	4.1259	3.6491	3.3389	3.1187	2.9530	2.8233	2.7185	2.6318	2.4961	2.3523	2.1978	2.0285	1.8363	1.6006
120	6.8510	4.7865	3.9493	3.4796	3.1735	2.9559	2.7918	2.6629	2.5586	2.4721	2.3363	2.1915	2.0346	1.8600	1.6557	1.3805
$\infty$	6.6349	4.6052	3.7816	3.3192	3.0173	2.8020	2.6393	2.5113	2.4073	2.3209	2.1848	2.0385	1.8783	1.6964	1.4730	1.0000

For  $m > 10$  interpolate using  $60/m$   
 For  $n > 30$  interpolate using  $120/n$

**STUDENT'S T CRITICAL VALUES**

<b>Conf. Level</b>	<b>50%</b>	<b>80%</b>	<b>90%</b>	<b>95%</b>	<b>98%</b>	<b>99%</b>
<b>One Tail</b>	<b>0.250</b>	<b>0.100</b>	<b>0.050</b>	<b>0.025</b>	<b>0.010</b>	<b>0.005</b>
<b>Two Tail</b>	<b>0.500</b>	<b>0.200</b>	<b>0.100</b>	<b>0.050</b>	<b>0.020</b>	<b>0.010</b>
<b>df = 1</b>	1.000	3.078	6.314	12.706	31.821	63.657
<b>2</b>	0.816	1.886	2.920	4.303	6.965	9.925
<b>3</b>	0.765	1.638	2.353	3.182	4.541	5.841
<b>4</b>	0.741	1.533	2.132	2.776	3.747	4.604
<b>5</b>	0.727	1.476	2.015	2.571	3.365	4.032
<b>6</b>	0.718	1.440	1.943	2.447	3.143	3.707
<b>7</b>	0.711	1.415	1.895	2.365	2.998	3.499
<b>8</b>	0.706	1.397	1.860	2.306	2.896	3.355
<b>9</b>	0.703	1.383	1.833	2.262	2.821	3.250
<b>10</b>	0.700	1.372	1.812	2.228	2.764	3.169
<b>11</b>	0.697	1.363	1.796	2.201	2.718	3.106
<b>12</b>	0.695	1.356	1.782	2.179	2.681	3.055
<b>13</b>	0.694	1.350	1.771	2.160	2.650	3.012
<b>14</b>	0.692	1.345	1.761	2.145	2.624	2.977
<b>15</b>	0.691	1.341	1.753	2.131	2.602	2.947
<b>16</b>	0.690	1.337	1.746	2.120	2.583	2.921
<b>17</b>	0.689	1.333	1.740	2.110	2.567	2.898
<b>18</b>	0.688	1.330	1.734	2.101	2.552	2.878
<b>19</b>	0.688	1.328	1.729	2.093	2.539	2.861
<b>20</b>	0.687	1.325	1.725	2.086	2.528	2.845

**FACULTY OF MANAGEMENT**  
**MBA I-Semester (CBCS) Examination**  
**MODEL PAPER - I**  
**STATISTICS FOR MANAGEMENT**

Time: 3 Hours]

[Max. Marks: 80

**PART - A (5 × 4 = 20 Marks)**

**(Short Answer Type)**

**Note :** Answer **All** the questions in not more than one page each

**ANSWERS**

1. Marginal Probability (Unit - I, SQA-2)
2. Define Random Variable. (Unit - II, SQA-2)
3. Define Sampling. (Unit - III, SQA-5)
4. What is Small sample test ? (Unit - IV, SQA-3)
5. What do you mean by correlation? (Unit - V, SQA-5)

**PART - B (5 × 12 = 60 Marks)**

**(Essay Answer Type)**

**Note :** Answer **All** the questions by using internal choice in not exceeding four pages each

6. (a) What are the Managerial Applications of Statistics? (Unit - I, Q.No.8)

OR

- (b) From the following data calculate mean deviation from the median.

C.I.	16 - 20	21 - 25	26 - 30	31 - 35	36 - 40	41 - 45	46 - 50	51 - 55	56 - 60
Frequency	8	15	13	20	11	7	3	2	1

(Unit - I, Prob No. 24)

7. (a) What is Binomial Distribution? State the assumptions of Binomial Distribution.

(Unit - II, Q.No.9)

OR

- (b) Fit a Poisson distribution to the following data and calculate theoretical frequencies,

Deaths	0	1	2	3	4
Frequency	122	60	15	2	1

(Unit - II, Prob No.8)

8. (a) Explain various probabilistic sampling methods.

(Unit - III, Q.No. 6)

OR

- (b) Define Hypothesis. What are the characteristics of Hypothesis ?

(Unit - III, Q.No. 12)

9. (a) Explain the test concerning the significance of single and two mean

(Unit - IV, Q.No. 5)

OR

- (b) Four technicians analyzed three samples each of the moisture content in the sample. The results are given below :

Samples	Technicians			
	A	B	C	D
X	9	12	10	11
Y	12	11	15	12
Z	9	10	12	14

Analyze the data and comment. Use 5% significance level.

(Unit - IV, Prob No. 12)

10. (a) Calculate Karl Pearson's Coefficient of Correlation for the following data.

X	7	6	5	4	3	2	1
Y	18	16	14	12	10	6	8

(Unit - V, Prob No. 3)

OR

- (b) Obtain the straight line trend equation for the following data by the method of the least square. Tabulate the trend values.

Year :	2010	2011	2012	2013	2014	2015	2016
Sale (in '000 units)	140	144	160	152	168	176	180

(Unit - V, Q.No. 31)



**FACULTY OF MANAGEMENT**  
**MBA I-Semester (CBCS) Examination**  
**MODEL PAPER - II**  
**STATISTICS FOR MANAGEMENT**

Time: 3 Hours]

[Max. Marks: 80

**PART - A (5 × 4 = 20 Marks)****(Short Answer Type)****Note :** Answer **All** the questions in not more than one page each**ANSWERS**

1. Explain the various functions of statistics. (Unit - I, SQA-5)
2. What is Binomial Distribution? (Unit - II, SQA-6)
3. Systematic Sampling (Unit - III, SQA- 9)
4. What is ANOVA? (Unit - IV, SQA-6)
5. What are the differences between Correlation and Regression. (Unit - V, SQA-4)

**PART - B (5 × 12 = 60 Marks)****(Essay Answer Type)****Note :** Answer **All** the questions by using internal choice in not exceeding four pages each

6. (a) Explain the concept of independent event. Discuss the different types of probabilities under Statistical Independence. (Unit - I, Q.No.56)

OR

- (b) Calculate Coefficient of Skewness based on Quarties from the following data:

Class Interval	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80	80 - 90
f	6	10	18	30	12	10	6	2

(Unit - I, Prob No.32)

7. (a) Define Poisson distribution. State the assumptions of Poisson distribution. (Unit - II, Q.No. 12)

OR

- (b) Fit a Binomial distribution to the following data.

X	0	1	2	3	4
F(x)	122	60	15	2	1

(Unit - II, Prob No. 1)

8. (a) Differentiate between random and non-random sampling methods. (Unit - III, Q.No. 7)

OR

- (b) Explain the procedure generally followed in testing of hypothesis. (Unit - III, Q.No. 13)

9. (a) What is ANOVA? State the assumptions and applications of ANOVA. (Unit - IV, Q.No. 7)

OR

- (b) The number of automobile accidents per week in a certain community are as follows : 12, 8, 20, 2, 14, 10, 15, 6, 9, 4. Are these frequencies in agreement with the be that accident conditions were the same during this 10 week period. (Unit - IV, Prob No. 17)
10. (a) Find the equation of the line of regression of x on y for the following data:

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y	5.3	5.7	6.3	7.2	8.2	8.7	8.4

(Unit - V, Prob No. 25)

OR

- (b) What are the different Types of correlations. (Unit - V, Q.No. 5)

**FACULTY OF MANAGEMENT**  
**MBA I-Semester (CBCS) Examination**  
**MODEL PAPER - III**  
**STATISTICS FOR MANAGEMENT**

Time: 3 Hours]

[Max. Marks: 80

**PART - A (5 × 4 = 20 Marks)****(Short Answer Type)****Note :** Answer **All** the questions in not more than one page each**ANSWERS**

1. Measures of Central Tendency. (Unit - I, SQA-7)
2. Define Poisson distribution. (Unit - II, SQA-9)
3. Cluster Sampling. (Unit - III, SQA- 10)
4. Chi-Square Test for independence of attributes. (Unit - IV, SQA- 9)
5. What are the properties of regression coefficient? (Unit - V, SQA-14)

**PART - B (5 × 12 = 60 Marks)****(Essay Answer Type)****Note :** Answer **All** the questions by using internal choice in not exceeding four pages each

6. (a) State and explain Baye's probability theorem. (Unit - I, Q.No.57)

OR

- (b) For the following data calculate mean, median and mode and comment on the same.

Class Interval	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
Frequency	7	14	18	20	11	10

(Unit - I, Prob No.16)

7. (a) In a company, the average number of phone calls per minute coming into a switch board is 3.2. Find the probability that during one particular minute there will be,
  - (i) Atleast 5 calls
  - (ii) Exactly 4 calls
  - (iii) No phone calls. (Unit - II, Prob No.11)

OR

- (b) Describe briefly about Normal Distribution. (Unit - II, Q.No.17)
8. (a) What are the Properties of a Good Estimator? (Unit - III, Q.No. 16)

OR

- (b) Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal, are same against that they are not, at 5% level. **(Unit - III, Q.No. 12)**

9. (a) Explain briefly about Chi-square test as test of goodness of fit. **(Unit - IV, Q.No. 12)**

OR

- (b) A machine is designed to produce insulating washers for electrical device of average thickness of 0.025 cm. A random sample of 10 washers was found to have a thickness of 0.024 cm with a S.D of 0.002 cm. Test the significance of the deviation. Value of t for 9 degrees of freedom at 5% level is 2.262. **(Unit - IV, Prob No.3)**

10. (a) What are the components of time series analysis. **(Unit - V, Q.No. 33)**

OR

- (b) Find Karl Pearson's coefficient correlation to the following:

X	40	45	53	55	38	42	45	62
Y	71	78	87	73	74	71	76	75

**(Unit - V, Prob No. 6)**

**FACULTY OF MANAGEMENT**  
**M.B.A. (CBCS) II - Semester Examination**  
**December - 2020**  
**BUSINESS RESEARCH METHODS**

Time : 2 Hours]

[Max. Marks : 80

**PART – A (4 × 5 = 20 Marks)**  
**[Short Answer Type]**

**ANSWERS**

**Note: Answer any four questions.**

- |                          |                    |
|--------------------------|--------------------|
| 1. Central Limit Theorem | (Unit-III, SQA-1)  |
| 2. Joint probability     | (Unit-I, SQA-1)    |
| 3. Ordinal Scale         | (Out of Syllabus)  |
| 4. Type-I error          | (Unit-III, SQA-12) |
| 5. Linear Regression     | (Unit-V, SQA-1)    |

**PART – B (4 × 15 = 60 Marks)**  
**(Essay Answer Type)**

**Note: Answer any four questions.**

- |   |                                |
|---|--------------------------------|
| 6. Explain about the Research Process.  | (Out of Syllabus)              |
| 7. Describe measures of Central Tendency and Dispersion.  | (Unit-I, Q.No. 9, 25)          |
| 8. Define Probability and explain about additive, multiplicative and conditional rules of probability with suitable examples. | (Unit-I, Q.No. 46, 54, 55, 56) |
| 9. In a Research Methodology Book, the following frequency mistakes per page were observed. Fit a Poisson distribution.       | (Unit-II, Prob. 6)             |

No. of Mistakes	0	1	2	3	4	5
No. of Pages	620	180	80	60	40	80

- |   |                          |
|---|--------------------------|
| 10. Describe the basic approaches and methods for merging data from different sources.                      | (Out of Syllabus)        |
| 11. Explain the Source of Secondary Data.   | (Out of Syllabus)        |
| 12. What is Hypothesis Testing? Why is it done? Explain various steps in Hypothesis Testing.                | (Unit-III, Q.No. 12, 13) |
| 13. A sample of sales in similar shops in two towns are taken for a new product with the following results: |                          |

Town	Mean Sales	Variance	Sample Size
A	5	5	5
B	7	3	7

Is there any evidence of difference in sales in the two towns? Use 5 per cent level of significance to test this difference between the means of two samples. (Unit-IV, Q.No. 5)

14. Calculate coefficient of correlation from the following data:

(Unit-V, Prob. 5)

X	100	200	300	400	500	600	700
Y	0.3	0.5	0.6	0.8	1.0	1.1	1.3

15. The following table gives the age of cars of a certain make and annual maintenance costs. Obtain the regression equation for costs related to age.

(Unit-V, Prob. 22)

Age of cars in year	2	4	6	8
Maintenance Cost in ` ('00) Hundered of `	10	20	25	30

**FACULTY OF MANAGEMENT**  
**M.B.A. II – Semester (CBCS) Examination**  
**November - 2020**  
**BUSINESS RESEARCH METHODS**

Time : 2 Hours]

[Max. Marks : 80

**PART - A (4 × 5 = 20 Marks)**

**ANSWERS**

**Note :** Answer any Four questions.

- |                                |                   |
|--------------------------------|-------------------|
| 1. Descriptive Research        | (Out of Syllabus) |
| 2. Marginal Probability        | (Unit-I, SQA-2)   |
| 3. Interval Scale              | (Out of Syllabus) |
| 4. Goodness of Fit             | (Unit-IV, SQA-1)  |
| 5. Spearman's Rank Correlation | (Unit-V, SQA-2)   |

**PART - B (4 × 15 = 60 Marks)**

**Note :** Answer any Four questions.

- |  |                          |
|--|--------------------------|
| 6. Explain various types of Research and give a suitable examples.   | (Out of Syllabus)        |
| 7. Discuss the features of a Good Research.  | (Out of Syllabus)        |
| 8. Explain about Baye's theorem with the help of suitable example.   | (Unit-I, Q.No. 57)       |
| 9. In a company, the average number of phone calls per minute coming into a switch board is 3.2. Find the probability that during one particular minute there will be, |                          |
| (i) Atleast 5 calls  |                          |
| (ii) Exactly 4 calls   |                          |
| (iii) No phone calls.  | (Unit-II, Prob. 11)      |
| 10. Explain reliability and validity of a scale.   | (Out of Syllabus)        |
| 11. What are the methods of sampling? Explain in briefly.  | (Unit-III, Q.No. 6)      |
| 12. What is estimation? What are the properties of a good estimator?   | (Unit-III, Q.No. 15, 16) |
| 13. 200 Digits were choosen at random from a set of tables. The frequencies of the digits were:  |                          |

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	16	25	22	20	4	15

Calculate  $\chi^2$  and test whether data is uniformly distributed.

(Incomplete Problem)

14. Following figures gives the rain fall in inches and production in '00 mds for Rabi crops for number of years. Find the coefficient of correlation between rainfall and production.

Rainfall	20	22	24	26	28	30	32
Production	15	18	20	32	40	39	40

(Unit-V, Prob. 8)

15. Find the equation of the line of regression of x on y for the following data:

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y	5.3	5.7	6.3	7.2	8.2	8.7	8.4

Find the (i) x on y

(ii) y on x

(Unit-V, Prob. 25)



**FACULTY OF MANAGEMENT**  
**M.B.A. II - Semester (CBCS) Examination**  
**May / June - 2019**  
**BUSINESS RESEARCH METHODS**

Time : 3 Hours ]

[Max. Marks : 80

**Note :** Answer ALL the questions from Part-A and Part-B  
Each question carries 4 marks in Part-A and 12 marks in Part-B.

**PART – A (5 × 4 = 20 Marks)**  
**(Short Answer Type)**

**ANSWERS**

- |    |  |                   |
|----|--|-------------------|
| 1. | What are the requirements of a good research design? | (Out of Syllabus) |
| 2. | Discuss the addition theorem of probability.         | (Unit-I, SQA-3)   |
| 3. | What are the benefits of sampling?                   | (Unit-III, SQA-2) |
| 4. | State the properties of good estimator.              | (Unit-III, SQA-3) |
| 5. | Explain the importance of correlation in research.   | (Unit-V, SQA-3)   |

**PART – B (5 × 12 = 60 Marks)**  
**(Essay Answer Type)**

- |     |   |                            |
|-----|---|----------------------------|
| 6.  | (a) Discuss various types of business research.   | (Out of Syllabus)          |
|     | OR  |                            |
|     | (b) What is experimental design? Explain its meaning and significance.  | (Out of Syllabus)          |
| 7.  | (a) What is meant by theoretical frequency distribution? Discuss the salient features of Binomial and Poisson distribution. | (Unit-II, Q.No. 15, 9, 12) |
|     | OR  |                            |
|     | (b) Briefly explain the properties and application of normal distribution in business research.                             | (Unit-II, Q.No. 19, 20)    |
| 8.  | (a) Discuss the procedure for measurement of variables in research design.  | (Out of Syllabus)          |
|     | OR  |                            |
|     | (b) Explain different methods of data collection.   | (Out of Syllabus)          |
| 9.  | (a) What is hypothesis testing? Explain the process of formulation and testing of an hypothesis.                            | (Unit-III, Q.No. 12, 13)   |
|     | OR  |                            |
|     | (b) Explain the important methods of Multi-Variate Analysis.  | (Out of Syllabus)          |
| 10. | (a) Explain the role of regression and correlation in carrying out a business research.                                     | (Unit-V, Q.No. 2, 20)      |
|     | OR  |                            |
|     | (b) Explain the significance and process of business report.  | (Out of Syllabus)          |

**FACULTY OF MANAGEMENT**  
**M.B.A. (CBCS) II - Semester Examination**  
**July - 2018**  
**BUSINESS RESEARCH METHODS**

Time : 3 Hours ]

[Max. Marks : 80

**Note :** Answer ALL the questions from Part-A and Part-B  
Each question carries 4 marks in Part-A and 12 marks in Part-B.

**PART – A (5 × 4 = 20 Marks)**  
**(Short Answer Type)**

**Note :** Answer ALL the questions in not more than one page each.

**ANSWERS**

- |   |                   |
|---|-------------------|
| 1. What is exploratory research?                              | (Out of Syllabus) |
| 2. Explain normal probability distribution.                   | (Unit-II, SQA-1)  |
| 3. What is questionnaire design?                              | (Out of Syllabus) |
| 4. Explain paired t-test.                                     | (Unit-IV, SQA-2)  |
| 5. Discuss the difference between regression and correlation. | (Unit-V, SQA-4)   |

**PART – B (5 × 12 = 60 Marks)**  
**(Long Answer Type)**

**Note :** Answer ALL the questions by using internal choice.

6. (a) What do you understand by applied research? Why it is known as applied? (Out of Syllabus)

OR

- (b) For the following data calculate mean, median and mode and comment on the same.

Class Interval	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
Frequency	7	14	18	20	11	10

(Unit-I, Prob. 16)

7. (a) What is probability? Explain the theorems related to probability. (Unit-I, Q.No. 46, 54, 55)

OR

- (b) Fit a Binomial distribution to the following data.

X	0	1	2	3	4
F(x)	122	60	15	2	1

(Unit-II, Prob. 8)

8. (a) Differentiate between random and non-random sampling methods.  
Explain in detail any two random sampling methods. (Unit-III, Q.No. 8, 7)

OR

(b) Distinguish between:

- (i) Direct and Indirect questions
- (ii) Open ended and closed questions
- (iii) Research, Investigative and Measurement questions
- (iv) Alternative response strategies.

(Out of Syllabus)

9. (a) Two different types of drugs A and B were tried on certain patients for increasing weight, 5 persons were given drug A and 7 persons were given drug B. The increase in weight (in pounds) is given below

Drug A	8	12	13	9	3		
Drug B	10	8	12	15	6	8	11

Do the drugs differ significantly with regard to their effect in increasing weight?

(Unit-IV, Prob. 4)

OR

- (b) A manufacturing company wishes to test the average life of the four brands of electric bulbs. The company uses all brands in a randomly selected production plants. The records showing the lives (in "00" hours) of bulbs are as given in the table below:

Brand 1	Brand 2	Brand 3	Brand 4
22	21	23	17
25	17	21	19
20	19	22	18
19	22	19	20
	18	18	

Test the hypothesis that the average life for each brand of bulbs is the same. Assume alpha 1%.

(Unit-IV, Prob. 11)

10. (a) Find Karl Pearson's coefficient correlation to the following:

X	40	45	53	55	38	42	45	62
Y	71	78	87	73	74	71	76	75

(Unit-V, Prob. 6)

OR

- (b) Find both regression lines to the following:

Mean (X) = 15

Mean (Y) = 110

Variance (X) = 25

Variance (Y) = 62.5 and  $r = 0.81$ .

(Unit-V, Prob. 23)

**FACULTY OF MANAGEMENT**  
**M.B.A. (CBCS) II - Semester (New) Examination**  
**July / August - 2017**  
**BUSINESS RESEARCH METHODS**

Time : 3 Hrs]

[Max. Marks : 80

**Note :** Answer all questions from Part - A and Part - B.  
Each question carries 4 marks in Part - A and 12 marks in Part - B.

**PART - A (5 × 4 = 20 Marks)****[Short Answer Type]****ANSWERS**

- |   |                    |
|---|--------------------|
| 1. What is standard error ?                     | (Unit-III, SQA-4)  |
| 2. Explain Joint probability.                   | (Unit-I, SQA-1)    |
| 3. What is Stratified Random Sampling ?         | (Unit-III, SQA-8)  |
| 4. Explain Type-I Error.                        | (Unit-III, SQA-12) |
| 5. Discuss the concept of correlation analysis. | (Unit-V, SQA-5)    |

**PART - B (5 × 12 = 60 Marks)****[Essay Answer Type]****Note : Answer all the questions**

6. (a) Distinguish between exploratory research and descriptive research with suitable examples. (Out of Syllabus)

OR

- (b) Find the mean and median of the following distribution.

Wages (Rs.)	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of workers	3	5	20	10	5

(Unit-I, Prob. 17)

7. (a) A husband and wife appear in an interview for two vacancies in the same post. The probability of husband's selection is  $\frac{1}{7}$  and that of wife's selection is  $\frac{1}{5}$ . What is the probability that only one of them will be selected ? (Unit-I, Prob. 49)

OR

- (b) An automatic machine makes paper clips from coils of wire. On the average, 1 in 400 paper clips is defective. If the paper clips are packed in boxes of 100. What is the probability that any given box of clips will contain (i) no defective, (ii) one or more defective and (iii) less than two defectives. (Unit-II, Prob. 10)

8. (a) What are the different sources of primary data and secondary data ? (Out of Syllabus)

OR

- (b) Explain various probabilistic sampling methods. (Unit-III, Q.No. 6)

9. (a) A manager wants an estimate of average sales of salesman in his company. A random sample of 100 out of 500 salesman is selected and average sales is found to be Rs.750 (thousand). If population standard deviation is Rs. 150 (thousand), manager specifies a 98% level of confidence. What is the interval estimate for average sales of salesman?

(Unit-III, Prob. 2)

OR

- (b) Four technicians analyzed three samples each of the moisture content in the sample. The results are given below :

Samples	Technicians			
	A	B	C	D
X	9	12	10	11
Y	12	11	15	12
Z	9	10	12	14

Analyze the data and comment. Use 5% significance level.

(Unit-IV, Prob. 12)

10. (a) Find Karl Pearson's coefficient correlation to the following :

X	48	39	65	80	73	60	52	120
Y	10	50	120	225	90	60	55	25

(Unit-V, Prob. 7)

OR

- (b) Find both regression lines to the following. Estimate the coefficient correlation and comment.

X	13	48	88	42	22	10	6
Y	8	52	82	84	22	10	6

(Unit-V, Prob. 24)