



ALAGAPPA UNIVERSITY

[Accredited with A+ Grade by NAAC (CGPA:3.64) in the Third Cycle &
Graded as Category – I University by MHRD-UGC]

(A State University Established by the Government of Tamilnadu)



KARAIKUDI – 630 003

DIRECTORATE OF DISTANCE EDUCATION

B.B.A

III - SEMESTER

10432

BUSINESS STATISTICS

Copy Right Reserved

For Private Use Only

Author :

Dr. S.Gopalsamy

Assistant Professor ,
Department of International Business
Alagappa University,
Karaikudi.

“The Copyright shall be vested with Alagappa University”

All rights reserved. No part of this publication which is material protected by this copyright notice may be reproduced or transmitted or utilized or stored in any form or by any means now known or hereinafter invented, electronic, digital or mechanical, including photocopying, scanning, recording or by any information storage or retrieval system, without prior written permission from the Alagappa University, Karaikudi, Tamil Nadu.

Reviewer :

Dr. R.Perumal

Professor of Management,
Directorate of Distance Education,
Alagappa University,
Karaikudi.

Work Order No.AU/DDE/DE12/III Sem/ PCM /2021 Dated 27.08.2021 Copies – 800

SYLLABI – BOOK MAPPING TABLE BUSINESS STATISTICS

| | Syllabi | Mapping in Book |
|--------|---|------------------------|
| Unit 1 | Basics of Statistics Introduction Statistics Data Data collection techniques Presentation of data | 1 - 15 |
| Unit 2 | Data Condensation and Graphical Methods Data condensation Diagram Graphs | 16 - 32 |
| Unit 3 | Measures of Central Tendency Measures of Central Tendency Mean Median Mode Partition values | 33 - 60 |
| Unit 4 | Measures of Dispersion Measures of Dispersion Range Quartile deviation Mean Deviation Standard Deviation Coefficient of Variable | 61 - 79 |
| Unit 5 | Moments, Skewness And Kurtosis Moments Skewness Kurtosis | 80 - 94 |
| Unit 6 | Correlation Analysis Correlation Linear Correlation Types of Correlation Scatter Diagram One – Way table Two – Way table Pearson’s Correlation Coefficient Spearman’s rank Correlation Coefficient Properties of Correlation Coefficient | 95 - 104 |

| | | |
|---------|--|------------------|
| Unit 7 | Regression Analysis Regression Linear Regression Types of Regression Curve fitting by the Method of Least square Derivations of Regression Equation Properties of Correlation Coefficient | 105 – 115 |
| Unit 8 | Index Number Index Number Cost of living Index Numbers Uses of Index Numbers Limitations of Index Numbers | 116 – 136 |
| Unit 9 | Analysis of Time Series Time series Measurement of trends Measurement of seasonal variation Forecasting Deseasonalisation | 137 - 153 |
| Unit 10 | Sampling Basic Concept of Sampling Sampling Methods Sampling and Non Sampling Errors Sampling Distribution Procedure for Hypothesis Null and Alternative Hypothesis Errors in Hypothesis testing One Tailed and Two Tailed Test | 154 – 171 |
| Unit 11 | Test of Hypothesis Hypothesis Testing on Population Mean Difference Between Mean of Two Populations Test of Hypothesis for Population Proportion Difference Between Two Proportion | 172 – 188 |
| Unit 12 | Chi – Square Test Characteristics of Chi –Square Test Uses of Chi –Square Test Steps of Chi –Square Test Analysis of Variance (ANOVA) Assumptions in Analysis of Variance Basic steps in Analysis of Variance | 189 – 197 |
| Unit 13 | Probability Importance Terms Types of Probability Basic relationship of Probability Addition Theorem of Probability | 198 – 213 |

| | | |
|---------|---------------------------------------|------------------|
| | Multiplication Theorem of Probability | |
| | Condition Probability | |
| | Baye's Theorem and its application | |
| Unit 14 | Probability Distribution | 214 - 232 |
| | Random Variable | |
| | Types of Random Variable | |
| | Binomial Distribution | |
| | Poisson Distribution | |
| | Normal Distribution | |

CONTENTS

UNIT 1 BASICS OF STATISTICS **1 - 15**

- 1.0 Introduction
- 1.1 Objectives
- 1.2 Statistics
 - 1.2.1 Definition of statistics
 - 1.2.2 Importance of statistics
 - 1.2.3 Limitations of statistics
 - 1.2.4 Functions of statistics
 - 1.2.5 Scope of statistics
- 1.3 Data
 - 1.3.1 Types of data
- 1.4 Data collection techniques
 - 1.4.1 Primary data
 - 1.4.2 Secondary data
- 1.5 Presentation of data
 - 1.5.1 Textual or descriptive presentation
 - 1.5.2 Tabular presentation
 - 1.5.3 Diagrammatic presentation
- 1.6 Summary
- 1.7 Key Words
- 1.8 Answers to Check Your Progress
- 1.9 Questions and Exercise
- 1.10 Further Readings

UNIT 2 DATA CONDENSATION AND GRAPHICAL METHODS **16 - 32**

- 2.0 Introduction
- 2.1 Objectives
- 2.2 Data condensation
 - 2.2.1 Raw data
 - 2.2.2 Attributes and variables
 - 2.2.3 Classification of Data
- 2.3 Diagram
 - 2.3.1 One dimensional diagram
 - 2.3.2 Two dimensional diagram
 - 2.3.3 Three dimensional diagram
 - 2.3.4 Pictogram and Cartogram
- 2.4 Graphs
 - 2.4.1 Histogram
 - 2.4.2 Frequency Polygon
 - 2.4.3 Frequency Curve

- 2.4.4 Ogive
- 2.5 Summary
- 2.6 Key Words
- 2.7 Answers to Check Your Progress
- 2.8 Questions and Exercise
- 2.9 Further Readings

Unit 3 MEASURES OF CENTRAL TENDENCY

33 - 60

- 3.0 Introduction
- 3.1 Objectives
- 3.2 Measures of Central Tendency
- 3.3 Mean
 - 3.3.1 Arithmetic mean
 - 3.3.2 Geometric mean
 - 3.3.3 Harmonic mean
- 3.4 Median
- 3.5 Mode
- 3.6 Partition values
 - 3.6.1 Quartiles
 - 3.6.2 Deciles
 - 3.6.3 Percentile
- 3.7 Summary
- 3.8 Key Words
- 3.9 Answers to Check Your Progress
- 3.10 Question and Exercise
- 3.11 Further Readings

UNIT 4 MEASURES OF DISPERSION

61 - 79

- 4.0 Introduction
- 4.1 Objectives
- 4.2 Measures of Dispersion
 - 4.2.1 Properties of a good measure of Dispersion
 - 4.2.2 Characteristics of Measures of Dispersion
 - 4.2.3 Classification of Measures of Dispersion
- 4.3 Range
- 4.4 Quartile deviation
- 4.5 Mean Deviation
- 4.6 Standard Deviation
 - 4.6.1 Calculation of Standard Deviation
- 4.7 Coefficient of Variable
- 4.8 Summary
- 4.9 Key Words

- 7.4.1 Regression Equation of Y on X
- 7.4.2 Regression Equation of X on Y
- 7.5 Curve fitting by the Method of Least square
- 7.6 Derivations of Regression Equation
- 7.7 Properties of Correlation Coefficient
- 7.8 Summary
- 7.9 Key Words
- 7.10 Answer to Check Your Progress
- 7.11 Questions and Exercise
- 7.12 Further Readings

UNIT 8 INDEX NUMBER

116 – 136

- 8.0 Introduction
- 8.1 Objectives
- 8.2 Index Numbers
 - 8.2.1 Types of Index Numbers
 - 8.2.2 Problems in construction of Index Numbers
 - 8.2.3 Methods of Constructing Index Numbers
 - 8.2.4 Quantity or Volume Index Numbers
 - 8.2.5 Test for Index Numbers
 - 8.2.6 Chain Base Index Numbers
- 8.3 Cost of living Index Numbers
 - 8.3.1 Construction of cost of living Index Numbers
 - 8.3.2 Methods to construct cost of living Index Numbers
 - 8.3.3 Uses of cost of living Index Numbers
- 8.4 Uses of Index Numbers
- 8.5 Limitations of Index Numbers
- 8.6 Summary
- 8.7 Key Words
- 8.8 Answers to Check Your Progress
- 8.9 Questions and Exercise
- 8.10 Further Readings

UNIT 9 ANALYSIS OF TIME SERIES

137 - 153

- 9.1 Time series
 - 9.1.1 Components of time series
 - 9.1.2 Analysis of time series
- 9.2 Measurement of trends

- 9.2.1 Moving average method
- 9.2.2 Least square method
- 9.3 Measurement of seasonal variation
 - 9.3.1 Methods of constructing seasonal indices
- 9.4 Forecasting
- 9.5 Deseasonalisation
- 9.6 Summary
- 9.7 Key Words
- 9.8 Answers to Check Your Progress
- 9.9 Questions and Exercise
- 9.10 Further Readings

UNIT 10 SAMPLING

154 – 171

- 10.0 Introduction
- 10.1 Objectives
- 10.2 Basic Concept of Sampling
- 10.3 Sampling Methods
 - 10.3.1 Random Sampling
 - 10.3.2 Non Random Sampling
- 10.4 Sampling and Non Sampling Errors
- 10.5 Sampling Distribution
- 10.6 Procedure for Hypothesis
- 10.7 Null and Alternative Hypothesis
- 10.8 Errors in Hypothesis testing
- 10.9 One Tailed and Two Tailed Test
- 10.10 Summary
- 10.11 Key Words
- 10.12 Answers to Check Your Progress
- 10.13 Questions and Exercise
- 10.14 Further Readings

UNIT 11 TEST OF HYPOTHESIS

172 – 188

- 11.0 Introduction
- 11.1 Objectives
- 11.2 Hypothesis Testing on Population Mean
 - 11.2.1 Population is Known
 - 11.2.2 Population is Unknown
- 11.3 Difference Between Mean of Two Populations

- 11.3.1 Population Variance Known
- 11.3.2 Population Variance Unknown
- 11.4 Test of Hypothesis for Population Proportion
- 11.5 Difference Between Two Proportion
- 11.6 Summary
- 11.7 Key Words
- 11.8 Answers to Check Your Progress
- 11.9 Questions and Exercise
- 11.10 Further Readings

UNIT 12 CHI – SQUARE TEST

189 – 197

- 12.0 Introduction
- 12.1 Objectives
- 12.2 Characteristics of Chi –Square Test
- 12.3 Uses of Chi –Square Test
- 12.4 Steps of Chi –Square Test
- 12.5 Analysis of Variance (ANOVA)
- 12.6 Assumptions in Analysis of Variance
- 12.7. Basic steps in Analysis of Variance
 - 12.7.1 One Way ANOVA
 - 12.7.2 Two Way ANOVA
- 12.8 Summary
- 12.9 Key Words
- 12.10 Answer to check your progress
- 12.11 Questions and Exercise
- 12.12 Further Reading

UNIT 13 PROBABILITY

198 – 213

- 13.0 Introduction
- 13.1 Objectives
- 13.2 Importance Terms
- 13.3 Types of Probability
- 13.4 Basic relationship of Probability
- 13.5 Addition Theorem of Probability
- 13.6 Multiplication Theorem of Probability
- 13.7. Condition Probability
 - 13.7.1 Combined Use Of Addition And Multiplication Theorem
- 13.8 Baye’s Theorem and its application

- 13.9 Summary
- 13.10 Key Words
- 13.11 Answer to Check your progress
- 13.12 Questions and Exercise
- 13.13 Further Readings

UNIT 14 PROBABILITY DISTRIBUTION

214 - 232

- 14.0 Introduction
- 14.1 Objectives
- 14.2 Random Variable
- 14.3 Types of Random Variable
- 14.4 Binomial Distribution
- 14.5 Poisson Distribution
- 14.6 Normal Distribution
- 14.7 Summary
- 14.8 Key Words
- 14.9 Answer to Check your progress
- 14.10 Questions and Exercise
- 14.11 Further Reading

Model Question Paper

233-234

UNIT 1 -BASICS OF STATISTICS

Basics of Statistics

NOTES

Structure

- 1.0 Introduction
 - 1.1 Objectives
 - 1.2 Statistics
 - 1.2.1 Definition of statistics
 - 1.2.2 Importance of statistics
 - 1.2.3 Limitations of statistics
 - 1.2.4 Functions of statistics
 - 1.2.5 Scope of statistics
 - 1.3 Data
 - 1.3.1 Types of data
 - 1.4 Data collection techniques
 - 1.4.1 Primary data
 - 1.4.2 Secondary data
 - 1.5 Presentation of data
 - 1.5.1 Textual or descriptive presentation
 - 1.5.2 Tabular presentation
 - 1.5.3 Diagrammatic presentation
 - 1.6 Summary
 - 1.7 Key Words
 - 1.8 Answers to Check Your Progress
 - 1.9 Questions and Exercise
 - 1.10 Further Readings

1.0 INTRODUCTION

Statistics is an area of study that deals with collecting, organising, analysing, interpreting and presenting data. The study of statistics has a lot of applications in industries, agriculture, medicine etc. In this unit, you will learn about the various importance and scope of statistics. You will also come to know about types of data, ways of collecting them and also how to present data.

1.1 OBJECTIVES

After going through this unit, you will

- Understand the meaning , importance and functions of statistics
- Learn the different types of data and how to collect them.
- Know how the data collected can be presented

Self-Instructional Material

NOTES

1.2 STATISTICS

The word statistics of English language have been derived from the Latin word status or Italian word 'statista' or German word 'statistik'. In each case it means "an organised political state". Although, in the past, statistics was considered as the "science of statecraft" as it was used by the government of various States to collect data regarding population, births, deaths, taxes etc.,. Statistics, nowadays, have experienced a modern development. Statistics play a crucial role in enriching a specific domain by collecting data in that field, analyse the data by applying various statistical techniques and making inferences about the same. For example, knowing the average height of the students will enable the engineer to know about the size of the door.

1.2.1 DEFINITION OF STATISTICS

The definition of statistics can be expressed in two ways to cover two different concepts. They are

1. Statistics as numerical data
2. Statistics for statistical method

1. Statistics as numerical data

When the word 'statistics' is used in plural sense, it refers to the collection of numerical data.

For example: - Export or Import quantity, Foreign Direct Investment, etc.,.

According to **Webster**," statistics are classified facts representing the conditions of the people in a state especially those facts which can be stated in number or in table of numbers or in any tabular or classified arrangements"

This definition of Webster reveals that only numerical facts can be termed statistics. This is an old, narrow and inadequate definition for modern times.

According to **Bawley** "Statistics are numerical statement of facts in any department of inquiry placed relation to each other"

Here, Bowley says that statistics is the science of counting and ignores other aspects such as analysis, interpretations etc.,.

According to **Yule and Kendall**," By statistics we mean quantitative data affected to a market extent by multiplicity of cause"

Yule and Kendall's definition tells us that numerical data is affected by multiplicity of cause. For example, the cost of production is affected by wage cost, exchange rate, raw material etc.,.

According to **Professor Horace Secrist**," It is the aggregate of facts affected to mark extent by multiplicity of causes, numerically

expressed, enumerated or estimated according to a reasonable standard of accuracy, connecting in a systematic manner for the predetermined purpose and placed in relation to each other"

Secrist's definition for statistics is more complete. The vital point that the definition covers are

- 1) Aggregate of facts
- 2) Affected by multiplicity of cause
- 3) Numerically expressed
- 4) Estimated according to standard of accuracy
- 5) Systematic Collection of data
- 6) Data collected for a predetermined purpose
- 7) Comparable

2. Statistics as Statistical Methods

According to **Bowley**," Statistics the science of measurement of social organism, regarded as a whole in all its manifestation"

This definition of Bowley is insufficient

According to **Wallis and Roberts**," Statistics is a body of methods for making wise decision on the face of uncertainty"

This definition is modern as it conveys statistical methods enable us to arrive at valid decisions.

According to **Croxton and Cowden**"statistics must be defined as the science of collection, presentation, analysis and interpretation of numerical data"

This definition gives a more elaborate meaning to statistics as statistical tools.

1.2.2 IMPORTANCE OF STATISTICS

Statistics can be used to various areas of business operations for effective results. Some prominent areas are given below.

- 1) **Startups** - While opening a new business or acquire one, we need to study the market from a statistical point of view to get accuracy in the market demand and supply .A businessman must do proper research by collecting data, analysing and interpreting them regarding market trends before starting his business.
- 2) **Production** - The production of the commodity depends upon various factors such as demand, supply of capital etc..., These factors must be analysed statistically to get a precise and accurate view of the same.
- 3) **Marketing** - An ideal marketing strategy requires statistical analysis on population, income of consumers, availability of the product ect.,.

NOTES

4) **Investment** - Statistics play a vital role in making decisions regarding buying shares, debentures or real estate. Using this statistical data, an investor will buy investments at a lesser price and sell when the price increases.

5) **Banking** - Banking sector is highly influenced by economic and market conditions. Bank have separate research department which collect and analyse information regarding inflation rate, interest rates, bank rates etc....

1.2.3 LIMITATIONS OF STATISTICS

1) Statistics does not analyse qualitative phenomenon

As statistics is a science which deals with numerical, it cannot be applied in data that cannot be measured in terms of quantitative measurements. However statistical techniques can be used to convert the qualitative data to quantitative data.

2) Statistics does study individuals

Statistics deals with aggregate quantities and doesn't give importance to individual data. This is because individual data is not useful for statistical analysis.

3) Statistical laws are not exact

Statistical interpretations are based on averages and hence are only approximations can be made

4) Statistics may be misused

Statistical data when used by an inexperienced person or illiterate person can lead to wrong interpretations. Hence it must be used only by experts.

1.2.4 FUNCTIONS OF STATISTICS

1) Consolidation

Statistics enables you to consolidate and understand huge data by providing only significant observations.

For example, instead of observing the marks of each and every individual with class average will enable you to know the class's performance as a whole.

2) Comparison

Classification and tabulation of data are used to compare the data. Various statistical tools such as graph, measure of depression dispersion, correlation gives us huge scope for comparison.

For example, the market demand for a product can be compared among the states. This enables the company to identify and analyse the target market.

3) Forecasting

Forecasting means predicting the future prospects. Statistics plays a huge role in forecasting the future.

For example, with the data of the sales value for the past 10 years, we will be able to predict the sales of the coming year approximately. Time series analysis and regression analysis are important for forecasting.

4) Estimation

One of the main aims of statistics is to draw conclusions on a huge population based on the analysis from a sample group.

For example, from a sample height of 10 students will be able to estimate the average height of all the students from the class.

5) Test of hypothesis

Statistical hypothesis is portraying a huge population from the inferences of a sample observation.

For example, if a particular fertilizer helps in increasing the crop yield in a particular area then it will be used in other areas based on this sample.

1.2.5 SCOPE OF STATISTICS

1) Statistics in Industries

Statistics is extensively used in huge number of industries. Statistics may be used in sales forecasting, consumer preference, quality control, inventory control, risk management etc. Sampling is vital for inspection plans.

2) Statistics in Education

Statistics plays an important role in education. Statistics help in measuring and evaluating the progress of the student, formulating policies and also helps to predict the future performance of the students to help them improve in the same.

3) Statistics in Economics

Statistics helps us to understand and analyse economic theories. Right from analysing microeconomic factors like the demand for the product, research regarding different markets to macroeconomic concept like inflation, unemployment can be done easily using statistics.

4) Statistics in Medicine

Statistics helps in researching and analysing medical experiments and investigations. Biostatic enables researchers to identify if a particular treatment or drug is working and how effective it is.

5) Statistics in Modern Application

A lot of software's are developed day to day for experimentation, forecasting and estimation.

NOTES

For example, SYSTAT is one such software which provides with scientific and technical graphical options.

6) Statistics in Agriculture

Statistics can be applied in agriculture by analysing the effectiveness of fertilizers. It can be used in taking decisions regarding inputs and outputs, inventories etc.,.

1.3 DATA

Data are pieces of factual information that are recorded and applied for analysis. Data is a tool which helps us to understand certain problems by providing us with information. They are a set of values with qualitative and quantitative variable.

1.3.1 TYPES OF DATA

Data are broadly classified into two based upon who collected the data

Primary data

Primary data is the data collected by investigator himself for the first time for his own research and analysis. It is also known as first-hand information. Primary data is collected using method such as personal interview, survey etc.,.

Secondary data

Secondary data is the data which is already been collected and processed by the person for the purpose of his research. Journals, internal sources, journals, book etc.,. are sources of secondary data.

CHECK YOUR PROGRESS -1

1. What are the two ways in which statistics can be defined?
2. What is the definition of statistics according to Professor Horace Secrist?
3. How does statistics help in comparing data?
4. What is the role of statistics in medicine?
5. What is secondary data?

1.4 DATA COLLECTING TECHNIQUES

1.4.1 PRIMARY DATA

1) Direct Personal Investigation

Direct personal investigation is the method in which the investigator directly goes to the source to collect information.

Merits

- (i) Information collected in this method is more authentic and accurate
- (ii) There is high degree of accuracy in qualitative information
- (iii) The original opinion or data shall be obtained.

Demerits

- (i) This is a time consuming process
- (ii) If the investigator is not intelligent enough to understand the mental state of the source it may lead to wrong interpretation.
- (iii) It may result in personal bias.

2) Indirect Oral Investigation

Indirect oral investigation is when the investigator investigates a person close to the source. This is done due to the reluctance of the original person.

Merits

- (i) It saves time and labour
- (ii) It is easy and convenient
- (iii) It covers a wide range of area.

Demerits

- (i) Information received may not be reliable
- (ii) Person chosen for this purpose may not be suitable
- (iii) It may be expensive as information is collected from various sources.

3) Information collected from local agencies

In this method investigator appoints a few agencies in various regions to cover various fields of inquiry. This method is generally used by newspaper companies to get information from various places in various topics such as sports, economics etc.,.

Merits

- (i) Avoid area can be easily covered
- (ii) This is a time saving method of collecting data
- (iii) The cost of collecting data is less

NOTES

Demerits

- (i) Sometimes the information collected may contradict one another
- (ii) The information can be less accurate
- (iii) This method will be expensive and a full-time agent is hired in different places

4) Questionnaire method

Questionnaire method is the most famous method of collecting primary data. A questionnaire is a set of questions device for conducting survey. The questionnaire is sent to the respondent with the request to fill it and send it back within a specific time.

Merits

- (i) This method is cheaper
- (ii) The time consumed for this process is very less
- (iii) This is an unbiased method of collecting data

Demerits

- (i) Sometimes the respondent may provide wrong information
- (ii) There is no type of personal motivation in this method
- (iii) There are chances of ignorance or late reply from the respondents

General principles of framing a questionnaire

1) The questionnaire must not be very long

We must try to give the questions as minimum as possible. Long questionnaire may lead to boredom or discontentment among the respondents.

2) The question must move from general to specific

When the question moves from general to specific respondent become more comfortable in answering the questions

3) The question should be ambiguous

The questions must be in such ways that the respondents are able to give clear and quick answers to the questions

4) The person should not contain double negatives

Words like don't you or wouldn't you must not be used in the questions as they might tempt the respondent to give a biased answer.

5) The question should not be leading questions

The questions should not give clues to the respondent on how they must answer it.

6) The question must not provide alternators for the answer

For example, instead of asking would you like to do engineering or medicine after class 12, the correct way of asking the question is would

you like to do engineering?

1.4.2 SECONDARY DATA

1) Published sources

Certain government and non-government organisations publish various journals, research papers, surveys etc which are very helpful and reliable. Some of them are mentioned below

- (i) Publications of international bodies like UNO, WTO and WHO etc.,.
- (ii) Publications of research institutes like ISI, NCERT, ICAR etc.,.
- (iii) Government publications
- (iv) Publications of commercial and financial institutions
- (v) Publications of governmental organisations
- (vi) Newspaper, journals and periodicals.

2) Unpublished sources

Unpublished sources cover all the sources where data is maintained privately by certain private agencies or companies. The data collected by universities, research institutions also come under unpublished sources.

1.5 PRESENTATION OF DATA

In the previous topic we saw how data can be collected .As the data collected is generally huge we need to comprise and deliver it in a presentable form. Generally there are three ways of presenting presentation of data. They are

- 1) Textual or Descriptive Presentation
- 2) Tabular Presentation
- 3) Diagrammatic Presentation

1.5.1 Textual or Descriptive Presentation

When the data collected is presented in the form of a text it is called textual or descriptive presentation. Generally this method cannot be used to present large data.

For example, in the 2011 census, the population of India was 1,21,08,54,977 comprising of 58, 64, 69,174 females and 62, 37, 24,248 males. The literacy rate is 74.04 percentage and density of population is 382 person per square kilometer.

From the above example, we can see that the data is represented textually. One of the major limitations of this method is that the readers must go through the entire text and get the required information.

1.5.2 Tabular Presentation of Data

When the data is presented in the form of rows and columns it is called tabular presentation of data.

NOTES

Example:

| AREA | FEMALE | MALE | TOTAL |
|-------|--------|-------|-------|
| URBAN | 90% | 89% | 89.5% |
| RURAL | 87% | 88% | 87.5% |
| TOTAL | 88.5% | 88.5% | 88.5% |

The above table represents the pass percentage of the examination conducted in Tamilnadu it has three rows (urban, rural, total) and three columns (female, male, total). It is a 3×3 table where each small box is called the cell which gives information regarding the pass percentage. This method is very significant as it enables us to use it for further statistical treatment. This tabular representation is further classified into four

(i) Qualitative Classification

Qualitative classification is when the collected information is classified in the form of attributes such as gender, nationality etc...,. The table given above is an example of qualitative classification where the information is classified in the form of gender and location.

(ii) Quantitative Classification

When information can be measured quantitatively like age, income, marks etc...,.then, such classifications are called quantitative classification

Example

| MARKS | FREQUENCY |
|-------|-----------|
| 0-10 | 5 |
| 10-20 | 10 |
| 20-30 | 20 |
| 30-40 | 15 |
| 40-50 | 10 |

(iii) Temporal Classification

Temporal classification is when classification is based on the basis of time like year, months, days etc...,.

NOTES**Example**

| DAYS OF A WEEK | PRODUCTION (no of pairs of shoes) |
|----------------|-----------------------------------|
| MONDAY | 2000 |
| TUESDAY | 1750 |
| WEDNESDAY | 3000 |
| THURSDAY | 2250 |
| FRIDAY | 1550 |

(iv) Spatial Classification

Spatial classification is when the data classification is based on place like town, city, district, state, country etc...,.

Example

| STATE | LITERACY RATE |
|----------------|---------------|
| TAMIL NADU | 80.09% |
| ANDHRA PRADESH | 67.02% |
| KARNATAKA | 75.36% |
| KERALA | 93.91% |

1.5.3 Diagrammatic Presentation

In this method the data is represented diagrammatically and is very easy to understand generally data is represented diagrammatically in three ways.

1) Geometric Diagram

This category consists of bar diagrams and pie charts

(i) Bar diagram

Bar diagram is a diagrammatic representation of data in equal spaced and equalwidth rectangular bars for each class of data .The height or length of the bar tells us about the magnitude of the class. Bar diagrams can be easily used for comparison of data. Both qualitative and quantitative data can be represented in bar diagram. They can be further divided into two broad categories.

a) Multiple bar diagram

When there is a need to compare two set of data multiple bar diagram is used. For example import and export, production

NOTES

and sale etc....

b) Component bar diagram

Component bar diagram also known as Sub diagrams are used to compare different components of a particular class. For example, the various components such as rent, medicine, education on which the monthly salary spend can be easily understood from a component bar diagram.

(ii) Pie diagram

A pie diagram is similar to that of a component bar diagram but it is represented in circle proportionally instead of bars. The values given in each class is converted into percentage and then each figure is multiplied by 3.6 degree. (360/100 - 360 degree of a circle divided into 100 parts) the values are then divided accordingly in the circle.

2) Frequency diagram

When the data is in the form of grouped frequency are usually represented by frequency diagrams. Histogram, frequency polygon, frequency curve and ogive are types of frequency diagram.

(i) Histogram

Histogram is a diagram which consists of rectangular bars whose area is proportional to the frequency of a variable and whose width is equal to the class interval.

(ii) Frequency polygon

A frequency polygon is another type of frequency distribution graph. In a frequency polygon, the number of observations is marked with a single point at the midpoint of each and every interval. Then the points are connected using a straight line.

(iii) Frequency curve

The frequency curve is obtained by drawing a smooth freehand curve that passes through the points of a frequency polygon closely as possible.

(iv) Ogive

Ogive also known as the cumulative frequencies are of two types. When the cumulative frequencies are plotted against their upper limits respectively, then it is less than ogive. When the cumulative frequencies are plotted against their lower limits respectively, then it is more than ogive.

3) Arithmetic line graph

An arithmetic line graph also known as time series graph is a graph where the time (months, years, weeks) are plotted in the x axis and their

respective values are plotted in the y axis. It helps us in analysing trends and periodicity of data.

CHECK YOUR PROGRESS - 2

6. What is indirect oral investigation?
7. State two merits of questionnaire method
8. Give some examples of published sources
9. What is component bar graph?
10. What is spatial classification?

1.6 SUMMARY

- The word ‘statistics’ is used in plural sense refers to the collection of numerical data and in singular sense it means the science of collecting, classifying and using statistics
- Statistics can be used to various areas of business operations such as start-ups, production, and marketing for effective results.
- Data is a tool which helps us to understand certain problems by providing us with information. It can be further divided into primary and secondary data.
- Direct personal investigation, indirect oral investigation, questionnaire methods are some of the methods of collecting primary data. Publications of international bodies, research institutions are methods of collecting secondary data.
- Data can be presented in three ways. They are Textual or descriptive presentation, Tabular presentation, Diagrammatic presentation.

1.7 KEY WORDS

Statistics, data, Primary data, Secondary data, Direct personal interview, Indirect oral investigation, Questionnaire, Qualitative, Quantitative, Temporal, Spatial, Bar diagram, Pie diagram, Histogram, Frequency Polygon, Frequency curve, Ogive, Arithmetic line graph.

1.8 ANSWERS TO CHECK YOUR PROGRESS

1. The word ‘statistics’ is used in plural sense refers to the collection of numerical data and when in singular sense it means the science of collecting, classifying and using statistics
2. According to Professor Horace Secrist, " It is the aggregate of facts affected to mark extent by multiplicity of causes, numerically expressed, enumerated or estimated according to a

NOTES

reasonable standard of accuracy, connecting in a systematic manner for the predetermined purpose and placed in relation to each other

3. Classification and tabulation of data are used to compare the data. Various statistical tools such as graph, measure of dispersion, correlation gives us huge scope for comparison.
4. Statistics helps in researching and analysing medical experiments and investigations. Biostatic enables researchers to identify if a particular treatment or drug is working and how effective it is.
5. Secondary data is the data which is already been collected and process by the person for the purpose of his research.
6. Indirect oral investigation is when the investigator investigates a person close to the source. This is done due to the reluctance of the original person.
7. Questionnaire method
 - (i) This method is cheaper
 - (ii) The time consumed for this process is very less.
8. Publications of international bodies like UNO, WTO and WHO, Publications of research institutes like ISI, NCERT, ICAR, and Government publications.
9. Component bar diagram also known as Sub diagrams are used to compare different components of a particular class.
10. Spatial classification is when the data classification is based on place like town, city, district, state, country etc.,.

1.9 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTIONS

1. Write short notes about the types of data
2. List the merits and demerits of direct personal interview
3. What are the general principles followed while framing a questionnaire?
4. Write about the classification of tabular presentation of data.
5. What is a bar diagram? What are its types?

LONG ANSWER QUESTIONS

1. Analyse the importance and scope of statistics
2. Explain in detail about the data collection techniques used in primary data.
3. Discuss about the functions and limitations of statistics.
4. Explain the various methods used for presentation of data.

1.10 FURTHER READINGS

1. Gupta, S. P. : Statistical Methods, Sultan Chand and Sons, New Delhi.
2. Hooda, R. P.: Statistics for Business and Economics, Macmillan, New Delhi.

NOTES

3. Hein, L. W. Quantitative Approach to Managerial Decisions, Prentice Hall,NJ.
4. Levin, Richard I. and David S. Rubin: Statistics for Management, Prentice Hall, New Delhi.
5. Lawrance B. Moore: Statistics for Business & Economics, Harper Collins, NY.
6. Watsman Terry J. and Keith Parramor: Quantitative Methods in Finance International, Thompson Business Press, London.

UNIT 2- DATA CONDENSATION AND GRAPHICAL METHODS

Structure

- 2.0 Introduction
- 2.1 Objectives
- 2.2 Data condensation
 - 2.2.1 Raw data
 - 2.2.2 Attributes and variables
 - 2.2.3 Classification of Data
- 2.3 Diagram
 - 2.3.1 One dimensional diagram
 - 2.3.2 Two dimensional diagram
 - 2.3.3 Three dimensional diagram
 - 2.3.4 Pictogram and Cartogram
- 2.4 Graphs
 - 2.4.1 Histogram
 - 2.4.2 Frequency Polygon
 - 2.4.3 Frequency Curve
 - 2.4.4 Ogive
- 2.5 Summary
- 2.6 Key Words
- 2.7 Answers to Check Your Progress
- 2.8 Questions and Exercise
- 2.9 Further Readings

2.0 INTRODUCTION

After collecting data from various sources, we need to organize them in order to understand the data and arrive at conclusions from it. The data must be sorted and condensed to put it through further statistical treatment. In this chapter we will learn about raw data and how to organize them in the form of various diagrams and graphs.

2.1 OBJECTIVES

From this unit you will

- Learn about raw data, attributes and variables
- Come to know about various forms of diagrams such as bar diagram, pie charts etc.,.
- Know how to construct a histogram, frequency polygon, ogive etc.,.

2.2 DATA CONDENSATION

Data condensation literally means to reduce the data, that is, to organize

the data for easy understanding and interpretation.

2.2.1 RAW DATA

Raw data refers to those data that are disorganized. These are the data that has not been processed for further use. There is a need for organizing and presenting such data to apply for the Statistical Techniques.

For example, the marks scored in Statistics by 50 students are given below

57 55 20 70 61 70 69 66 65 72

52 79 74 67 72 74 40 47 85 87

72 61 65 24 35 80 57 59 92 50

59 64 74 92 50 59 64 74 79 80

77 63 56 80 53 55 54 67 86 92

From the above data it is difficult to find out how many students have passed or failed, how many students have scored above 80 etc.,. When the data is classified according to the similarities, it enables us to easily identify, compare and arrive at conclusions without any difficulty.

2.2.2 ATTRIBUTES AND VARIABLES

Attributes are data which focuses on numbers. They are usually something that defines the data. Variable refers to those data which gives us even more clear information regarding the data and involves calculation.

For example, when finding a defective machine for group of Machines attributes will enable us to identify if the machines is defective or not but variables will enable us to know the level of defectiveness, that is 20% defective or 10% defective etc.,.

Variables can be further classified into discrete and continuous. If a variable takes uncountable values, it is called continuous variable .For example the weight of the student increases from 40kg to 50kg, his or her weight may take any value between 40 to 50 kg, even fractions like 40.5 kg, 45.3 kg etc. Discrete variables can only take some values.For example, the strength of a class can only be whole number.

2.2.3 CLASSIFICATION OF DATA

Chronological Classification

When the variables are classified according to time such as weeks, months, years etc.,then it is chronological classification.

Spatial Classification

Variables are classified according to geographical locations like States, countries, towns etc., and then it is spatial classification.

NOTES

Qualitative Data

When the variables are classified according to attributes such as gender religion literacy nationality etc., they are called qualitative classification.

Quantitative Data

When the variables are classified according to characteristics that can be expressed numerically like height, weight, income etc., it is called quantitative classification.

2.3 DIAGRAM

A diagram is a visual form for presentation of statistical data. Diagrams are of different types they are bars, circles, maps, pictorial, and cartograms.

Advantages

- It is very simple to draw and read as well.
- It is the only form of diagram which can represent a large number of data on a piece of paper.
- It can be drawn both vertically and horizontally.
- It gives a better look and facilitates comparison.

Disadvantages

- It cannot exhibit a large number of aspects of the data.
- The bars are fixed arbitrarily by a drawer

Types of Diagrams

- One dimensional diagrams
- Two dimensional diagrams
- Three dimensional diagrams
- Pictograms and cartograms

2.3.1 ONE DIMENSIONAL DIAGRAM

These are the most commonly used diagrams. Usually horizontal or vertical lines or bars with their lengths proportional to the magnitudes of the observations corresponding to each category constitute this diagram.

Bar diagrams are of various types

- Simple bar diagrams
- Subdivided bar diagrams
- Percentage bar diagrams
- Multiple bar diagrams
- Deviation bar diagrams

Simple bar diagrams

Horizontal or vertical bars (fully shaded rectangles) with the same width, drawn with their bases on the same horizontal or vertical line with equal gaps in between and lengths proportional to the magnitudes of the

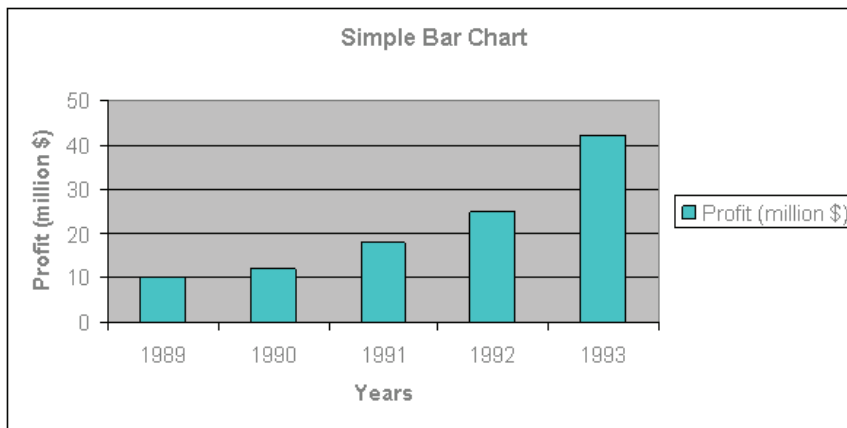
observations constitute a bar diagram.

Example:-

Draw simple bar diagram to represent the profits of a bank for 5 years.

| Years | 1989 | 1990 | 1991 | 1992 | 1993 |
|-------------------|------|------|------|------|------|
| Profits (million) | 10 | 12 | 18 | 25 | 42 |

A simple bar chart showing the profits of a bank for 5 years:



Subdivided bar diagram or component bar diagrams

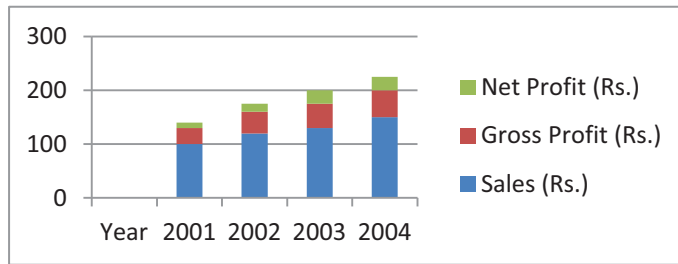
This type of diagram is used when the observations corresponding to the various categories have different components and it is felt that a comparison of the component parts is important. Here a simple bar diagram is first drawn with the length of the bars proportional to the totals of the component parts and then it is sub divided into parts of length proportional to the component magnitudes and each part given a different color or shading.

Example:-

Draw a component bar diagram for the following data

| Year | Sales (Rs.) | Gross Profit (Rs.) | Net Profit (Rs.) |
|------|-------------|--------------------|------------------|
| 2001 | 100 | 30 | 10 |
| 2002 | 120 | 40 | 15 |
| 2003 | 130 | 45 | 25 |
| 2004 | 150 | 50 | 25 |

NOTES



Percentage bar diagrams

In this, the component parts are expressed as the percentages of the total and a component bar diagram is drawn with all bars having equal length.

Sometimes when the volumes of different attributes may be greatly different for making meaningful comparisons, the attributes are reduced to percentages. In that case each attribute will have 100 as its maximum volume. This sort of component bar chart is known as percentage bar diagram.

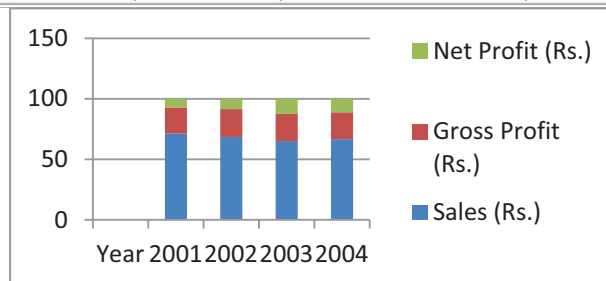
$$\text{Percentage} = \frac{\text{Actualvalue}}{\text{Totaloftheactualvalue}} \times 100$$

Example:-

Draw a Percentage bar diagram for the following data

Using the formula $\text{Percentage} = \frac{\text{Actualvalue}}{\text{Totaloftheactualvalue}} \times 100$, the above table is converted.

| Year | Sales (Rs.) | Gross Profit (Rs.) | Net Profit (Rs.) |
|------|-------------|--------------------|------------------|
| 2001 | 71.43 | 21.43 | 7.14 |
| 2002 | 68.57 | 22.86 | 8.57 |
| 2003 | 65 | 22.5 | 12.5 |
| 2004 | 66.67 | 22.22 | 11.11 |



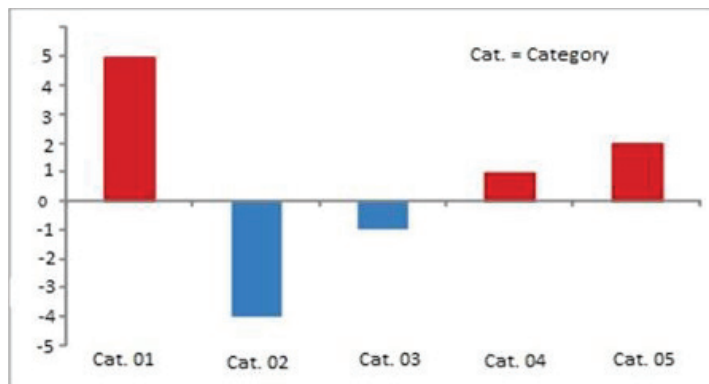
Deviation bar diagrams

This diagram is usually used to represent net quantities like net profit, balance payable, deficit or excess etc. as the observations may be positive or negative, the base line is usually drawn in the middle of the paper horizontally and positive values are indicated by bars of proportional length, drawn above the horizontal line and negative values by bars of proportional length drawn below the horizontal line.

Example:-

Represent the following data in a suitable bar diagram

| Year | Sales (Rs in '0000) | Profit / loss (Rs in '0000) |
|------|---------------------|-----------------------------|
| 2001 | 24 | 10 |
| 2002 | 35 | -3 |
| 2003 | 45 | 7 |
| 2004 | 59 | -5 |



2.3.2 TWO DIMENSIONAL DIAGRAMS

In two dimensional diagrams, areas of the diagrams are used to represent themagnitudes. Rectangles, squares and circles with area proportional to the observationsare used to represent each category. Of these, circles are most commonly used. Suchdiagrams are called pie-diagrams. Circles drawn with areas proportional to themagnitudes of the observations constitute a pie-diagram.

Pie Diagram or Circular Diagram:

Another way of preparing a two-dimensional diagram is in the form of circles. In suchdiagrams, both the total and the component parts or sectors can be shown. The area of acircle is proportional to the square of its radius. While making comparisons, pie diagramsshould be used on a percentage basis and not on an absolute basis.

- In constructing a pie diagram the first step is to prepare the data

NOTES

so that various components values can be transposed into corresponding degrees on the circle.

- The second step is to draw a circle of appropriate size with a compass. The size of the radius depends upon the available space and is proportional to the square root of total frequency.
- The third step is to measure points on the circle and representing the size of each sector with the help of a protractor. Since there are 360 degrees in a circle, a class with a relative frequency of .25 would consume $.25(360) = 90$ degrees of the circle.

Example:-

Given are the areas of cultivable land in four southern states of India. Construct a pie diagram for the following data.

| State | Cultivable area(in hectares) |
|----------------|-------------------------------|
| Andhra Pradesh | 663 |
| Karnataka | 448 |
| Kerala | 290 |
| Tamil Nadu | 556 |
| Total | 1957 |

Using the formula,

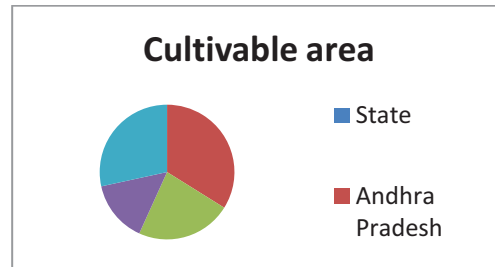
$$\text{Angle} = \frac{\text{Actual value}}{\text{Total of the actual value}} \times 360^\circ$$

(or)

$$\text{Angle} = \frac{\text{Percentage}}{100} \times 360^\circ$$

The table value becomes

| State | Cultivable area |
|----------------|-----------------|
| Andhra Pradesh | 121.96 |
| Karnataka | 82.41 |
| Kerala | 53.35 |
| Tamil Nadu | 102.28 |



2.3.3 THREE DIMENSIONAL DIAGRAMS

Cubes, cylinders, blocks etc. with volumes proportional to the magnitudes of the observations are drawn in this case to represent them.

Pictograms and Cartograms

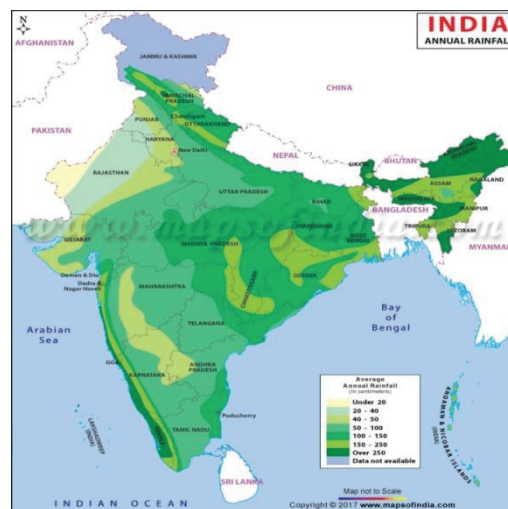
A pictogram or pictograph represents the frequency of data as pictures or symbols. Each picture or symbol may represent one or more units of the data. Cartograms are used to give quantitative information on a geographical basis. The map of a country with regions receiving the same annual rainfall shaded in the same manner is a cartogram. The magnitude in this case is the annual rainfall and it can be indicated by a foot note giving the rainfall corresponding to each type of shading.

Example:-

The pictograph shows the number of varieties of apples stored at a supermarket:

| Varieties of Apples in a food store | |
|-------------------------------------|--|
| Red Delicious | |
| Golden Delicious | |
| Red Rome | |
| McIntosh | |
| Jonathan | |

= 10 apples = 5 apples



Pictograms

Cartograms

CHECK YOUR PROGRESS - 1

1. What is a raw data?
2. Explain attributes and variables using an example
3. What is the formula used to find the angle in a pie chart?
4. What is subdivided bar diagram?
5. What are two dimensional diagrams used for?

6. Mention 2 merits of a diagram.

*Data Condensation
and Graphical
Methods*

NOTES

2.4 GRAPHS

A graph is a visual form of presentation of statistical data. A graph is more attractive than a table of figure. Even a common man can understand the message of data from the graph. Comparisons can be made between two or more phenomena very easily with the help of a graph. Most important types of graphs are

- Histogram
- Frequency Polygon
- Frequency Curve
- Ogive

A graph is a visual form of presentation of statistical data. A graph is more attractive than a table of figure. Even a common man can understand the message of data from the graph. Comparisons can be made between two or more phenomena very easily with the help of a graph. Most important types of graphs are

- Histogram
- Frequency Polygon
- Frequency Curve
- Ogive

2.4.1 HISTOGRAM

A histogram is a bar chart or graph showing the frequency of occurrence of each value of the variable being analyzed. In histogram, data are plotted as a series of rectangles. Class intervals are shown on the 'X-axis' and the frequencies on the 'Y-axis' if the classes are of equal width and frequency density (f/c) on 'Y-axis' if the classes are of unequal width. The height of each rectangle represents the frequency or frequency density of the class interval. Each rectangle is formed with the other so as to give a continuous picture. Such a graph is also called staircase or block diagram. However, we cannot construct a histogram for distribution with open-end classes.

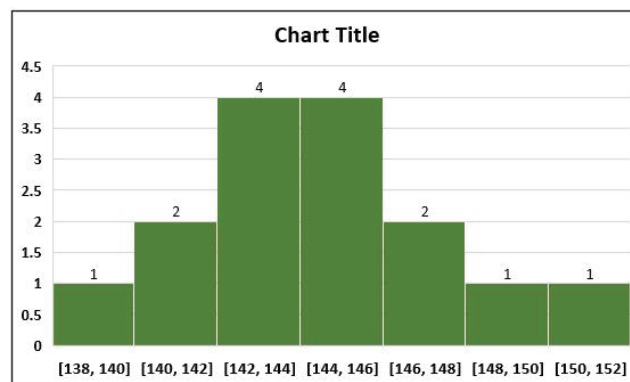
Example:-

Mr. Larry a famous doctor is conducting research on the height of the students studying in 8th standard. He has gathered a sample of 15 students but wants to know which the maximum category is where they belong.

| Sr No | Height (in cms) |
|-------|-----------------|
| 1 | 141 |
| 2 | 143 |
| 3 | 145 |
| 4 | 145 |
| 5 | 147 |
| 6 | 152 |
| 7 | 143 |
| 8 | 144 |
| 9 | 149 |
| 10 | 141 |
| 11 | 138 |
| 12 | 143 |
| 13 | 145 |
| 14 | 148 |
| 15 | 145 |

Solution:

We have created a histogram using 6 bins with 6 different frequencies as seen below in the chart. In Y axis it's the average number of students falling in that particular category. In X-axis we have the range of height, for example, the 1st bin range is 138 cms to 140 cms. And we can note that the count is 1 for that category from the table and as seen in the below graph.



Here we can see the heights of the students on an average are in range of 142 cm to 146 cm for 8th standard. And also, one can note that one side of the average also falls on the other side of the average which is the sign of normal distribution.

2.4.2 FREQUENCY POLYGON

If we mark the midpoints of the top horizontal sides of the rectangles in a histogram and join them by a straight line, the figure so formed is called a Frequency Polygon. This is done under the assumption that the frequencies in a class interval are evenly distributed throughout the class. The area of the polygon is equal to the area of the histogram, because the area left outside is just equal to the area included in it. Another method of drawing frequency polygon is to draw the midpoints on the X axis and the frequency density (f/c) on the Y axis. Join the points by straight line to obtain frequency polygon.

Example:-

In a batch of 400 students, the height of students is given in the

NOTES

following table. Represent it through a frequency polygon.

| Height (in cm) | Number of Students(Frequency) |
|----------------|-------------------------------|
| 140 – 150 | 74 |
| 150 – 160 | 163 |
| 160 – 170 | 135 |
| 170 – 180 | 28 |
| Total | 400 |

Solution:

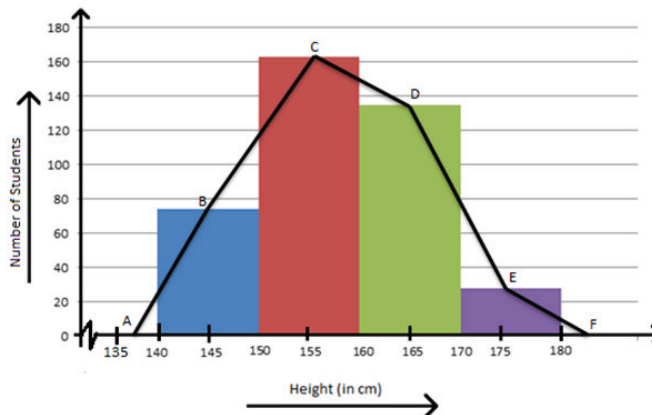
Following steps are to be followed to construct a histogram from the given data:

- The heights are represented on the horizontal axes on a suitable scale as shown.

The number of students is represented on the vertical axes on a suitable scale as shown.

Now rectangular bars of widths equal to the class- size and the length of the bars corresponding to a frequency of the class interval is drawn.

ABCDEF represents the given data graphically in form of frequency polygon as:



Frequency polygons can also be drawn independently without drawing histograms. For this, the midpoints of the class intervals known as class marks are used to plot the points.

$$\text{Class Mark} = \frac{\text{Upper Limit} + \text{Lower Limit}}{2}$$

2.4.3 FREQUENCY CURVE

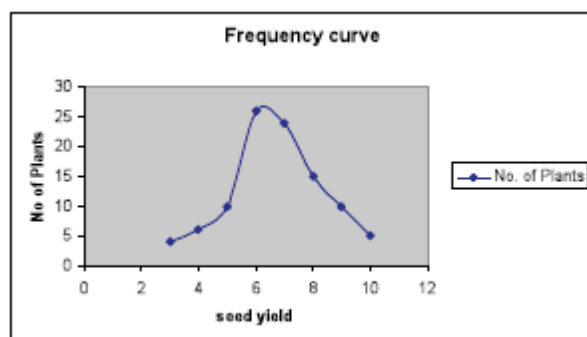
If the middle point of the upper boundaries of the rectangles of a histogram is corrected by a smooth freehand curve, then that diagram is called frequency curve. The curve should begin and end at the base line.

Example:-

Draw frequency curve for the following data

| Seed yield (gms) | No.of.Plants |
|------------------|--------------|
| 2.5-3.5 | 4 |
| 3.5-4.5 | 6 |
| 4.5-5.5 | 10 |
| 5.5-6.5 | 26 |
| 6.5-7.5 | 24 |
| 7.5-8.5 | 15 |
| 8.5-9.5 | 10 |
| 9.5-10.5 | 5 |

Solution:



2.4.4 OGIVES

The cumulative frequency gives the cumulative frequency of each of the class. The curve table obtained by plotting cumulative frequencies is called a cumulative frequency curve or an Ogive.

There are two type of Ogive namely:

1. The 'less than Ogive'
2. The 'more than Ogive'.

In less than Ogive method we start with the upper limits of the classes and go adding the frequencies. When these frequencies are plotted, we get a rising curve. In more than Ogive method, we start with the lower limits of the classes and from the total frequencies we subtract the frequency of each class. When these frequencies are plotted we get a declining curve.

Example:-

For the data given below, construct a less than cumulative frequency table and plot its Ogive.

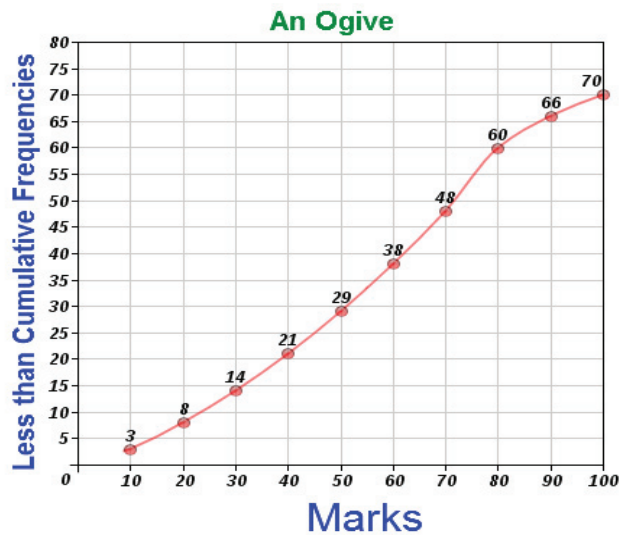
NOTES

| | | | | | | | | | | |
|------------------|--------|---------|---------|---------|---------|---------|---------|---------|---------|----------|
| Marks | 0 - 10 | 10 - 20 | 20 - 30 | 30 - 40 | 40 - 50 | 50 - 60 | 60 - 70 | 70 - 80 | 80 - 90 | 90 - 100 |
| Frequency | 3 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 6 | 4 |

Solution:

| Marks | Frequency | Less than cumulative frequency |
|----------|-----------|--------------------------------|
| 0 - 10 | 3 | 3 |
| 10 - 20 | 5 | 8 |
| 20 - 30 | 6 | 14 |
| 30 - 40 | 7 | 21 |
| 40 - 50 | 8 | 29 |
| 50 - 60 | 9 | 38 |
| 60 - 70 | 10 | 48 |
| 70 - 80 | 12 | 60 |
| 80 - 90 | 6 | 66 |
| 90 - 100 | 4 | 70 |

Plot the points having abscissa as upper limits and ordinates as the cumulative frequencies (10, 3), (20, 8), (30, 14), (40, 21), (50, 29), (60,38), (70, 48), (80, 60), (90, 66), (100, 70) and join the points by a smooth curve.



Example:-

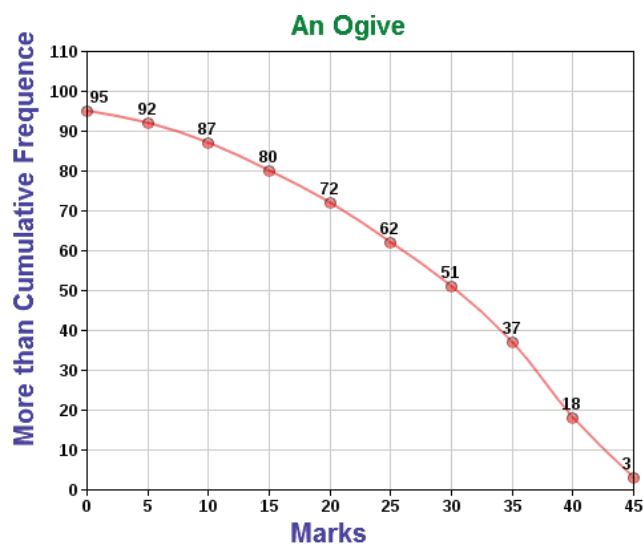
For the data given below, construct a more than cumulative frequency table and plot its Ogive.

| | | | | | | | | | | |
|------------------|-------|--------|---------|---------|---------|---------|---------|---------|---------|---------|
| Marks | 0 - 5 | 5 - 10 | 10 - 15 | 15 - 20 | 20 - 25 | 25 - 30 | 30 - 35 | 35 - 40 | 40 - 45 | 45 - 50 |
| Frequency | 3 | 5 | 7 | 8 | 10 | 11 | 14 | 19 | 15 | 13 |

Solution:

| Marks | Frequency | More than cumulative frequency |
|--------------|------------------|---------------------------------------|
| 0 - 5 | 3 | 95 |
| 5 - 10 | 5 | $95 - 3 = 92$ |
| 10 - 15 | 7 | $92 - 5 = 87$ |
| 15 - 20 | 8 | $87 - 7 = 80$ |
| 20 - 25 | 10 | $80 - 8 = 72$ |
| 25 - 30 | 11 | $72 - 10 = 62$ |
| 30 - 35 | 14 | $62 - 11 = 51$ |
| 35 - 40 | 19 | $51 - 14 = 37$ |
| 40 - 45 | 15 | $37 - 19 = 18$ |
| 45 - 50 | 13 | $18 - 15 = 3$ |

On the graph, plot the points (0, 95), (5, 92), (10, 87), (15, 80), (20, 72), (25, 62), (30, 51), (35, 37), (40, 18), (45, 3) and join the points by a smooth curve.



CHECK YOUR PROGRESS - 2

7. state whether the following statement is true or false
- In less than Ogive method we start with the upper limits of the classes and go adding the frequencies.
 - If we mark the midpoints of the top horizontal sides of the rectangles in a histogram and join them by a straight line, the figure so formed is called a Frequency curve.
 - A common man cannot understand the message of data from the graph.
 - Frequency polygons can also be drawn independently without drawing histograms

2.5 SUMMARY

- Data condensation literally means to reduce the data, that is, to organize the data for easy understanding and interpretation. Raw data refers to those data that are disorganized. These are the data that has not been processed for further use.
- Attributes are data which focuses on numbers. They are usually something that defines the data. Variable refers to those data which gives us even more clear information regarding the data and involves calculation.
- A diagram is a visual form for presentation of statistical data. Diagrams are of different types they are bars, circles, maps, pictorial, and cartograms. they are further classified as one dimensional diagrams, two dimensional diagrams, three dimensional diagrams, Pictograms and cartograms.
- A graph is a visual form of presentation of statistical data. A graph is more attractive than a table of figure. . Most important types of graphs are Histogram, Frequency Polygon, Frequency Curve, and Ogive.

2.6 ANSWERS TO CHECK YOUR PROGRESS

- Raw data refers to those data that are disorganized. These are the data that has not been processed for further use.
- For example, when finding a defective machine for group of Machines attributes will enable us to identify if the machine is defective or not but variables will enable us to know the level of defectiveness, that is 20% defective or 10% defective etc.,.
- $$\text{Angle} = \frac{\text{Actual value}}{\text{Total of the actual value}} \times 360^\circ$$
- Here a simple bar diagram is first drawn with the length of the bars proportional to the totals of the component parts and then it is sub divided into parts of length proportional to the component

- magnitudes and each part given a different color or shading.
- In two dimensional diagrams, areas of the diagrams are used to represent themagnitudes.
 - It is very simple to draw and read as well.
It is the only form of diagram which can represent a large number of data on a piece of paper.
 - a) True
b) False
c) False
d) True

2.7 KEY WORDS

Raw data, attributes, variables, diagram, One dimensional diagrams, Two dimensional diagrams, Three dimensional diagrams, Pictograms and cartograms, Simple bar diagrams, Subdivided bar diagrams, Percentage bar diagram, Multiple bar diagrams, Deviation bar diagrams, Histogram, Frequency Polygon, Frequency Curve, and Ogive.

2.8 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTIONS

- Write short notes on variables and attributes.
- Mention the merits and demerits of a diagram
- Explain about one dimensional diagrams and their classification
- Describe how to draw a pie chart
- Give a detailed account on Ogive.
- Draw a bar diagram from the following data

| Subject | English | Tamil | Maths | Science | Social Science |
|---------|---------|-------|-------|---------|----------------|
| Marks | 76 | 58 | 98 | 86 | 77 |

LONG ANSWER QUESTIONS

- From the following information, draw a less than and more than ogive curve

| Marks | 0 - 10 | 10 - 20 | 20 - 30 | 30 - 40 | 40 - 50 | 50 - 60 | 60 - 70 | 70 - 80 | 80 - 90 | 90 - 100 |
|-----------|--------|---------|---------|---------|---------|---------|---------|---------|---------|----------|
| Frequency | 6 | 9 | 5 | 3 | 8 | 6 | 14 | 10 | 7 | 3 |

- From the following data, draw a frequency polygon and frequency curve.

| Marks | 0 - 5 | 5 - 10 | 10 - 15 | 15 - 20 | 20 - 25 | 25 - 30 | 30 - 35 | 35 - 40 | 40 - 45 | 45 - 50 |
|-------|-------|--------|---------|---------|---------|---------|---------|---------|---------|---------|
|-------|-------|--------|---------|---------|---------|---------|---------|---------|---------|---------|

| | | | | | | | | | | |
|------------------|---|----|----|----|----|----|----|----|----|----|
| Frequency | 6 | 10 | 14 | 16 | 20 | 22 | 28 | 38 | 30 | 26 |
|------------------|---|----|----|----|----|----|----|----|----|----|

- 3) Explain in detail about graphs and their classification
4) Construct a histogram from the following data

| MARKS | FREQUENCY |
|-------|-----------|
| 0-10 | 5 |
| 10-20 | 10 |
| 20-30 | 20 |
| 30-40 | 15 |
| 40-50 | 10 |

2.9 FURTHER READINGS

1. Business Statistics by Shenoy and Shenoy.
2. Statistical Methods by S.P. Gupta.
3. Statistics for Business and Economics by R.P. Hooda.

UNIT 3 MEASURES OF CENTRAL TENDENCY

Measures of Central Tendency

NOTES

Structure

- 3.0 Introduction
- 3.1 Objectives
- 3.2 Measures of Central Tendency
- 3.3 Mean
 - 3.3.1 Arithmetic mean
 - 3.3.2 Geometric mean
 - 3.3.3 Harmonic mean
- 3.4 Median
- 3.5 Mode
- 3.6 Partition values
 - 3.6.1 Quartiles
 - 3.6.2 Deciles
 - 3.6.3 Percentile
- 3.7 Summary
- 3.8 Key Words
- 3.9 Answers to Check Your Progress
- 3.10 Question and Exercise
- 3.11 Further Readings

3.0 INTRODUCTION

Measures of central tendency are a statistical tool used to summarize data that depicts the central value of the given data. These measures enable us to identify where most of the values fall. The three most commonly used measures of central tendency are mean, median and mode. In this unit you will learn about them extensively and also learn about some other partition values.

3.1 OBJECTIVES

From this unit you will

- Learn about the measures of central tendency
- Come to know about the various methods of calculating mean, median and mode.
- Know about the partition values.

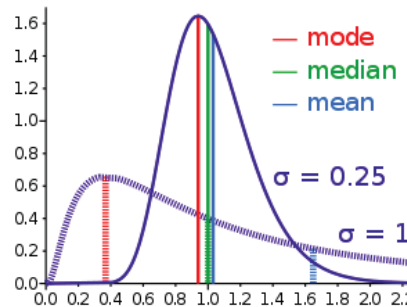
3.2 MEASURES OF CENTRAL TENDENCY

When working on a given set of data, it is not possible to remember all the values in that set. But we require inference of the data given to us. This problem is solved by mean, median and mode. Measures of Central Tendency, represent all the values of the data. As a result, they help us to draw an inference and an estimate of all the values. They are also known as statistical averages. Their simple

Measures of Central Tendency

NOTES

function is to mathematically represent all the values in a particular set of data. Hence, this representation shows the general trend and inclination of all the values.



An average provides a simple way of representation of all the individual data. It also aids in the comparison of different groups of data. In addition to this, an average in economic terms can represent the direction an economy is headed towards. Hence, it can be easily used to formulate policies and bring about a reform for a better economy.

3.3 MEAN

3.3.1 ARITHMETIC MEAN

The arithmetic mean of a series of numbers is sum of all observations divided by the total number of observations in the series.

Example:

There are two brothers, with different heights. The height of the younger brother is 138 cm and height of the elder brother is 154cm. The average height of the two brother is total height divided into two equal parts,

$$(138+154) \div 2 = 292 \div 2 = 146 \text{ cm}$$

So 146 cm is the average height of the brothers. Here $154 > 146 > 138$. The average value lies in between the minimum value and the maximum value.

Thus if x_1, x_2, \dots, x_n represent the values of n observations, then arithmetic mean (A.M.) for n observations is: (direct method)

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

There are two methods for computing the arithmetic mean: (i) Direct method (ii) Short cut method.

Direct Method:

Example:

The following data represent the number of books issued in a college library is selected from 7 different days 17, 19, 22, 25, 15, 40, 21 find the mean number of books.

Solution:

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

$$\bar{x} = \frac{20 + 39 + 22 + 25 + 45 + 40 + 54}{7} = \frac{245}{7} = 35$$

Hence the mean of the number of books is 35

Indirect Method:

In this method an assumed mean or an arbitrary value (A) is used as the basis of calculation of deviations (d_i) from individual values. If $d_i = x_i - A$

$$\bar{x} = A + \frac{\sum_{i=1}^n d_i}{n}$$

Example:

A student's marks in 5 subjects are 95, 78, 88, 72, 99. Find the average of his marks.

Let us take the assumed mean, $A = 88$

| x_i | $d_i = x_i - 88$ |
|-------|------------------|
| 95 | 7 |
| 78 | 10 |
| 88 | 0 |
| 72 | -16 |
| 99 | 10 |
| Total | 11 |

Solution:

$$\bar{x} = A + \frac{\sum_{i=1}^n d_i}{n}$$

$$= 88 + \frac{11}{5} = 88 + 5.5 = 93.5$$

The arithmetic mean of average marks is 93.5

*Measures of Central
Tendency*

NOTES

Discrete Grouped data

If x_1, x_2, \dots, x_n are discrete values with the corresponding frequencies f_1, f_2, \dots, f_n .

Then the mean for discrete grouped data is defined as (direct method)

$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{N}$$

In the short cut method the formula is modified as

$$\bar{x} = A + \frac{\sum_{i=1}^n f_i d_i}{N} \quad \text{where } d_i = x_i - A$$

Example:

Given the following frequency distribution, calculate the arithmetic mean

| | | | | | | |
|------------------|----|----|----|----|----|----|
| Marks | 64 | 63 | 62 | 61 | 60 | 59 |
| No. Of. Students | 8 | 18 | 12 | 9 | 7 | 6 |

Solution:

| x_i | f_i | $f_i x_i$ | $d_i = x_i - A$ ($A=62$) | $f_i d_i$ |
|-------|-------|-----------|-------------------------------|-----------|
| 64 | 8 | 512 | 2 | 16 |
| 63 | 18 | 1134 | 1 | 18 |
| 62 | 12 | 744 | 0 | 0 |
| 61 | 9 | 549 | -1 | -9 |
| 60 | 7 | 420 | -2 | -14 |
| 59 | 6 | 354 | -3 | -18 |
| | 60 | 3713 | | -7 |

Direct Method

$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{N}$$

$$\bar{x} = 3713 / 60 = 61.88$$

Short cut method

$$\bar{x} = A + \frac{\sum_{i=1}^n f_i d_i}{N} \times c$$

Here A = 62

$$\bar{x} = 62 - \frac{7}{60} = 61.88$$

The mean mark is 61.88

Mean of continuous Grouped data:

Direct method

$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{N}, x_i \text{ is the midpoint of the class interval}$$

Short cut method

$$\bar{x} = A + \frac{\sum_{i=1}^n f_i d_i}{N} \times c$$
$$d = \frac{x_i - A}{c}$$

Where A – any arbitrary value

c - Width of the class interval

x_i - is the midpoint of the class interval

Example:

For the frequency distribution of yield of tomato given in table calculate the mean yield per plot.

| | | | | |
|-------------------------|---------|----------|-----------|-----------|
| Yield per plot (in Kg) | 64 - 84 | 84 - 104 | 104 - 124 | 124 - 144 |
| No of plots | 3 | 5 | 7 | 20 |

NOTES

Solution:

| Yield (in Kg) | No of plots (f _i) | Mid x _i | f _i x _i | d = (x _i - A) / c | f _i d _i |
|----------------|--------------------------------|--------------------|-------------------------------|------------------------------|-------------------------------|
| 64 - 84 | 3 | 74 | 222 | -1 | -3 |
| 84 - 104 | 5 | 94 | 470 | 0 | 0 |
| 104 - 124 | 7 | 114 | 798 | 1 | 7 |
| 124 - 144 | 20 | 134 | 2680 | 2 | 40 |
| Total | 35 | | 4170 | | 44 |

Direct Method

$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{N}$$

$$\bar{x} = 4170 / 35 = 119.143$$

Short cut method

$$\bar{x} = A + \frac{\sum_{i=1}^n f_i d_i}{N} \times c$$

$$\bar{x} = 94 + \frac{44}{35} \times c = 119.143$$

3.3.2 WEIGHTED ARITHMETIC MEAN

For calculating simple mean, all the values or the sizes of items in the distribution have equal importance. But in practical life this may not be so, in case some items are more important than others, a simple average computed is not representative of the distribution. Proper weightage has to be given to the various items.

For example a student may use a weighted in order to calculate their percentage grade in a course, in this the student would multiply the weighing of all assessment items in the course(eg: assignment, exams, projects, etc.)by respective grade that was obtained in each of categories

The average whose component items are being multiplied by certain values known as “weights” and the aggregate of the multiplied results are divided by the total sum of their “weight”

Let x₁,x₂,...,x_n be the set of n values having weights w₁,w₂,...,w_n respectively,

then the weighted mean is

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + \dots + w_n x_n}{w_1 + w_2 + w_3 + \dots + w_n} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Example:

A student obtained the marks 40,50,60,80, and 45 in math, statistics, physics, chemistry and biology respectively. Assuming weights 5,2,4,3, and 1 respectively for the above mentioned subjects, find the weighted arithmetic mean per subject.

Solution

| Components | Marks scored (x_i) | Weightage (w_i) | $w_i x_i$ |
|--------------|------------------------|---------------------|------------|
| Maths | 40 | 5 | 200 |
| Statistics | 50 | 2 | 100 |
| Physics | 60 | 4 | 240 |
| Chemistry | 80 | 3 | 240 |
| Biology | 45 | 1 | 45 |
| Total | | 15 | 825 |

Weighted average:

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$
$$= 825 / 15 = 55 \text{ marks / subject}$$

Combined Mean:

In the arithmetic averages and the number of items in two or more related groups are known, the combined or the composite mean of the entire group can be obtained by

$$\bar{x}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

The advantage of combined arithmetic mean is that we can determine the overall mean of the combined data without going back to the original data

Example:

If a sample size of 22 items has a mean of 15 and another sample size of 18 items has a mean of 20. Find the mean of the combined sample?

Solution:

$$\bar{x}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$
$$= \frac{22 \times 15 + 18 \times 20}{22 + 18}$$
$$= \frac{330 + 360}{40} = \frac{690}{40} = 172.5$$

NOTES

Merits of AM

1. It can be calculated easily and is also easy to understand.
2. Fluctuation can be minimized
3. It can further be used for statistical treatment like median, mode etc.,.
4. This method is rigidly defined and hence can be used for comparison

Demerits of AM

1. It cannot be plotted in a graph.
2. It is not applicable in qualitative data.
3. AM cannot be calculated if the class intervals have open ends.
4. It is highly influenced by extreme observations.

3.3.2 GEOMETRIC MEAN (GM)

A geometric mean is a mean or average which shows the central tendency of a set of numbers by using the product of their values.

The geometric mean of two numbers, say x , and y is the square root of their product $x \times y$. For three numbers, it will be the cube root of their products i.e., $(x \cdot y \cdot z)^{1/3}$.

The geometric mean of a series containing n observations is the n th root of the product of the values. If x_1, x_2, \dots, x_n are observations then

$$\begin{aligned} \text{G. M.} &= \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n} \\ &= (x_1 \cdot x_2 \cdot \dots \cdot x_n)^{\frac{1}{n}} \\ \log \text{G.M.} &= \log (x_1 \cdot x_2 \cdot \dots \cdot x_n) \\ &= (\log x_1 + \log x_2 + \dots + \log x_n) \\ &= \frac{\sum_{i=1}^n \log x_i}{n} \\ \text{G.M.} &= \text{Antilog} \frac{\sum_{i=1}^n \log x_i}{n} \end{aligned}$$

Example:

Calculate the geometric mean of the following growth of price of onions per 100 Kg per annum is 180, 250, 490, 1400, and 1050

| | | | | | | |
|-------|--------|--------|--------|--------|--------|----------------|
| x | 180 | 250 | 490 | 1400 | 1050 | Total |
| log x | 2.2553 | 2.3979 | 2.6902 | 3.1461 | 3.0212 | 13.5107 |

*Measures of Central
Tendency*

NOTES

Solution:

$$\begin{aligned} \text{G.M.} &= \text{Antilog } \frac{\sum_{i=1}^n \log x_i}{n} \\ &= \text{Antilog } \frac{13.5107}{5} \\ &= \text{Antilog } 2.7021 = 503.6 \end{aligned}$$

Geometrical mean of onion rate is 503.6

Example:

Find the geometric mean for the following distribution of student's marks:

| | | | | |
|------------------|--------|---------|---------|----------|
| Marks | 0 – 30 | 30 – 50 | 50 – 80 | 80 - 100 |
| No . of students | 20 | 30 | 40 | 10 |

Solution:

| Marks | No of students f | Mid points x | f log x |
|--------------|------------------|--------------|-------------------------------------|
| 0 – 30 | 20 | 15 | 20 (log 15) = 20(1.1761) = 23.5218 |
| 30 – 50 | 30 | 40 | 30 (log 40) = 30 (1.6020) = 48.0168 |
| 50 – 80 | 40 | 65 | 40 (log 65) = 20(1.8129) = 72.5165 |
| 80 - 100 | 10 | 90 | 10 (log 90) = 20(1.9542) = 19.5424 |
| Total | 100 | | 163.6425 |

$$\begin{aligned} \text{G.M.} &= \text{Antilog } \frac{\sum_{i=1}^n \log x_i}{n} \\ &= \text{Antilog } \frac{163.6425}{100} \\ &= \text{Antilog } 1.6364 = 503.6 \end{aligned}$$

NOTES

Geometrical mean of onion rate is 43.29

Merits of Geometric mean:

1. It is strictly defined
2. It is based on all items
3. It is very suitable for averaging ratio, rates and percentages
4. It is capable of further mathematical treatment
5. Unlike AM, it is not affected much by the presence of extreme values

Demerits of geometric mean:

1. It cannot be used when the values are negative or if any of the observations is zero
2. It is difficult to calculate particularly when the items are very large or when there is a frequency distribution
3. It brings out the property of the ratio of the change and not the absolute difference of change as the case in arithmetic mean
4. The GM may not be the actual value of the series

3.3.3 HARMONIC MEAN

Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given values. If x_1, x_2, \dots, x_n are n observations.

A harmonic mean is used in averaging of ratios. The most common examples of ratios are that of speed and time, cost and unit of material, work and time etc. The harmonic mean (H.M.) of n observations is

H.M. for ungrouped data

$$H. M. = \frac{n}{\sum_{i=1}^n \left(\frac{1}{x_i}\right)}$$

Example:

Calculate the harmonic mean of the numbers 13.5, 14.5, 14.8, 15.2 and 16.1

Solution:

The harmonic mean is calculated as below:

| x | 1 / x |
|------|--------|
| 13.2 | 0.0758 |
| 14.2 | 0.0704 |
| 14.8 | 0.0676 |
| 15.2 | 0.0658 |

| | |
|--------------|---------------|
| 16.1 | 0.0621 |
| Total | 0.3417 |

$$\begin{aligned} \text{H. M.} &= \frac{n}{\sum \left(\frac{1}{x_i}\right)} \\ &= \frac{5}{0.3417} = 14.63 \end{aligned}$$

H.M. Discrete Grouped data:

For a frequency distribution

$$\text{H. M.} = \frac{N}{\sum_{i=1}^n f_i \left(\frac{1}{x_i}\right)}$$

Example:

The frequency distribution of first year students of a particular college, calculate the harmonic mean

| | | | | | |
|-------------|----|----|----|----|----|
| Age (years) | 17 | 18 | 19 | 20 | 21 |
| | 2 | 5 | 13 | 7 | 3 |

Solution:

| Age (years) x | Number of students f | f / x |
|----------------|----------------------|---------------|
| 17 | 2 | 0.1176 |
| 18 | 5 | 0.2778 |
| 19 | 13 | 0.6842 |
| 20 | 7 | 0.3500 |
| 21 | 3 | 0.1429 |
| Total | 30 | 1.5725 |

$$\begin{aligned} \text{H. M.} &= \frac{N}{\sum_{i=1}^n f_i \left(\frac{1}{x_i}\right)} \\ &= 30 / 1.5725 = 19.0779 \approx 19 \text{ years} \end{aligned}$$

Merits of H.M:

1. It is strictly defined
2. It is defined on all observations.
3. It is amenable to further algebraic actions
4. It is most suitable average when it is desired to give greater

NOTES

weight to smaller observations and less weight to larger observations.

Demerits of H.M:

1. It is not easily understood.
2. It is difficult to calculate.
3. It is only an abstract figure and may not be the action of the series.

CHECK YOUR PROGRESS – 1

1. What the 3 measures of central tendency?
2. What is the formula for arithmetic mean under direct method?
3. Mention 2 merits of geometric mean
4. What do you mean by harmonic mean?

3.4 MEDIAN

The number of students in your classroom, the money your parents earns, the temperature in your city is all important numbers. But how can you get the information of the number of students in your school or the amount earned by the citizen of your entire city?

The median is that value of the variable which divides the group into two equal parts, one part comprising all values greater and the other all values less than median.

Ungrouped data

Arrange the given values in the ascending or descending order.

If the number of value is odd, median is the middle value.

For example if we have the number of values 12, 15, 21, 27, 35. So the numbers are odd then taking the mean as the midpoint 21.

$$\text{Median} = \frac{(n+1)^{th}}{2} \text{ term if } n \text{ is odd}$$

If the number of values is even, median is the mean of the middle two values.

For example if we have 12, 15, 21, 27, 35, 40. So the numbers are even then taking the mean of the numbers,

$$\text{Median} = \text{Mean}\left(\frac{(n)^{th}}{2} \text{ and } \frac{(n+1)^{th}}{2} \text{ terms}\right)$$

So in the above example, take the mean of 21 and 27 and divide it by 2 which will give you 24.

Example:

The salaries of 8 employees who work for a small company are listed below. What is the median salary?

40,000; 29,000; 35,500; 31,000; 43,000; 30,000; 27,000; 32,000

Solution:

Arrange the data in ascending order

27,000; 29,000; 30,000; 31,000; 32,000; 35,500; 40,000; 43,000

Since there is an even number of items in the data set, we compute the median by taking the mean of the two middlemost numbers.

$$\begin{aligned} \text{Mean } \left(\frac{(n)^{\text{th}}}{2} \text{ and } \frac{(n+1)^{\text{th}}}{2} \text{ terms} \right) &= \frac{4^{\text{th}} + 5^{\text{th}} \text{ item}}{2} \\ &= \frac{31,000 + 32,000}{2} = \frac{63,000}{2} = 31,500 \end{aligned}$$

The median salary is 31,500

Example: 13

Find the median of the following set of points in a game: 15, 14, 10, 8, 12, 8, 16

Solution:

First arrange the values in an ascending order 8, 8, 10, 12, 14, 15, 16

The number of point values is 7, an odd number. Hence, the median is the value in the middle position.

$$\begin{aligned} \text{Median} &= \left(\frac{n+1}{2} \right)^{\text{th}} \text{ term} \\ &= \left(\frac{7+1}{2} \right)^{\text{th}} \text{ term} = 4^{\text{th}} \end{aligned}$$

The median is 12

Grouped data:

In grouped distribution, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or continuous frequency distribution. Whatever may be the distribution, cumulative frequencies have to be calculated the total number of items.

Cumulative frequency: (cf)

Cumulative frequency of each class is the sum of the frequency of the class and the frequencies of the previous classes, ie adding the frequencies successively, so that the last cumulative frequency gives the total number of items.

Measures of Central Tendency

NOTES

When the data follows a discrete set of values grouped by size, we use the formula $\frac{(n+1)^{th}}{2}$ item for finding the median. First we form a cumulative frequency distribution, and the median is that value which corresponds to the cumulative frequency in which $\frac{(n+1)^{th}}{2}$ item lies.

Example: 14

The following frequency distribution is classified according to the number of students on different branches. Calculate the median number of leaves per branch.

| | | | | | | | |
|--------------------|---|----|----|----|----|----|----|
| No of Students | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Number of Branches | 2 | 11 | 15 | 20 | 25 | 18 | 10 |

Solution:

| No of Students x | No of Branches f | Cumulative Frequency cf |
|---------------------|---------------------|----------------------------|
| 1 | 2 | 2 |
| 2 | 11 | 13 |
| 3 | 15 | 28 |
| 4 | 20 | 48 |
| 5 | 25 | 73 |
| 6 | 18 | 91 |
| 7 | 10 | 101 |
| Total | 101 | |

$$\begin{aligned} \text{Median} &= \text{size of } \frac{(N+1)^{th}}{2} \text{ item} \\ &= \text{size of } \frac{(101+1)^{th}}{2} \text{ item} \\ &= 51^{th} \text{ item} \end{aligned}$$

Median = 5 because 51th item corresponds to 5

Median for continuous grouped data

In case, the data is given in the form of a frequency table with class interval etc, then the following formula is used for calculating median in continuous grouped data

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

Where l = Lower limit of the median class

m = cumulative frequency preceding the median

c = width of the median class

f = frequency in the median class

N = total frequency

Example:

Calculate median from the following data

| | | | | | | | | |
|----------------|-----|-----|-------|-------|-------|-------|-------|-------|
| Class interval | 0-4 | 5-9 | 10-14 | 15-19 | 20-24 | 25-29 | 30-34 | 35-39 |
| Frequency | 5 | 8 | 10 | 12 | 7 | 6 | 3 | 2 |

Solution:

| Class interval | Frequency f | True interval class | Cumulative frequency cf |
|----------------|---------------|---------------------|---------------------------|
| 0-4 | 5 | 0.5 - 4.5 | 5 |
| 5-9 | 8 | 4.5 - 9.5 | 13 |
| 10-14 | 10 | 9.5 - 14.5 | 23 |
| 15-19 | 12 | 14.5 - 19.5 | 35 |
| 20-24 | 7 | 19.5 - 24.5 | 42 |
| 25-29 | 6 | 24.5 - 29.5 | 48 |
| 30-34 | 3 | 29.5 - 34.5 | 51 |
| 35-39 | 2 | 34.5 - 39.5 | 53 |
| | 53 | | |

$$\frac{N}{2} = \frac{53}{2} = 26.5$$

Here the cumulative frequency is greater than or equal to 26.5 is 14.5

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$$l = 14.5$$

$$N/2 = 26.5$$

$$m = 23$$

$$f = 12$$

NOTES

$$= 14.5 + \frac{(26.5 - 23) \times 5}{12} = 14.5 + 1.46 = 15.96$$

Merits of Median:

1. Median is not influenced by extreme values because it is a positional average.
2. Median can be calculated in case of distribution with open end intervals.
3. Median can be located even if the data are incomplete.
4. Median can be located even for qualitative factors such as ability, honesty etc.

Demerits of Median:

1. A slight change in the series may bring drastic change in median value
2. In case of even number of items or continuous series, median is an estimated value other than any value in the series.
3. It is not suitable for further mathematical treatment except its use in mean deviation.
4. It is not taken into account all the observation.

3.5 MODE

The mode is the most frequently occurring values or scores. The mode is useful when there are a lot of repeated values. There can be no mode, one mode, or multiple modes.

Its importance is very great in marketing studies where a manager is interested in knowing about the size, which has the highest concentration of items. For example, in placing an order foot shoes or ready-made garments the model size helps because the sizes and other sizes around in common demand.

Ungrouped Data:

For ungrouped values or a series of individual observation mode is often found by mere inspection

Example:

Find the mode for the following list of values:
13,18,13,14,13,16,14,21,13

Solution:

The mode is the number that is repeated more often than any other

Therefore the Mode = 13

In some cases the mode may be absent while in some cases there may be more than one mode.

Example:

Ms. Rossy asked students in her class how many siblings they each has.

Find the mode of the data : 0,0,0,1,1,1,1,2,2,2,2,3,3,4

Solution:

The modes are 1 and 2 siblings

Grouped Data

For Discrete distribution, the highest frequency and corresponding value of X is mode.

Continuous distribution:

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

- Where L is the lower class limit of the modal class
- f_1 is the frequency of the modal class
- f_0 is the frequency of the class preceding the modal class in the frequency table
- f_2 is the frequency of the class succeeding the modal class in the frequency table
- h is the class interval of the modal class

Example:18

Calculate mode for the following:

| C -I | 0-50 | 50-100 | 100-150 | 150-200 | 200-250 | 250-300 | 300-350 | 350-400 | 400 and above |
|---------|------|--------|---------|---------|---------|---------|---------|---------|---------------|
| f | 5 | 14 | 40 | 91 | 450 | 87 | 60 | 38 | 15 |

Solution:

The highest frequency is 450 and corresponding class interval in 200 – 250, which is the modal class

Here L = 200, $f_1 = 450$, $f_0=91$, $f_2=87$, $h=50$

$$\begin{aligned} \text{Mode} &= L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times h \\ &= 200 + \frac{450 - 91}{2 \times 450 - 91 - 87} \times 50 \\ &= \frac{2450}{122} = 200 + 24.18 = \mathbf{224.18} \end{aligned}$$

Example: 19

Find the modal class and the actual mode of the data set below

| | | | | | | | | | | |
|-----------|-----|-----|-----|-------|-------|-------|-------|-------|-------|-------|
| Number | 1-3 | 4-6 | 7-9 | 10-12 | 13-15 | 16-18 | 19-21 | 22-24 | 25-27 | 28-30 |
| Frequency | 7 | 6 | 4 | 2 | 2 | 8 | 1 | 2 | 3 | 2 |

Solution:

Modal class = 10 – 12

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

Here $L = 10$, $f_1 = 9$, $f_0 = 4$, $f_2 = 2$, $h = 3$

$$= 10 + \frac{9 - 4}{2 \times 9 - 2 - 4} \times 3$$

$$= 10 + \frac{5}{12} \times 3 = 10 + 1.25 = 11.25$$

Mode = 11.25

Merits of mode:

1. It is easy to calculate and in some cases it can be located mere inspection.
2. Mode is not at all affected by extreme values
3. It can be calculated for open-end classes
4. It is usually an actual value of an important part of the series
5. In some circumstances it is the best representative of data

Demerits of mode:

1. It is not based on all observation
2. It is not capable of further mathematical treatment
3. Mode is ill defined generally it is not possible to find mode in some cases.
4. As compared with mean, mode is affected to a great extent by sampling fluctuations

It is unsuitable in cases where relative importance of items has to be considered.

3.6 PARTITION MEASURES

3.6.1 QUARTILES

The quartiles divide the distribution in four parts. There are three quartiles denoted by Q1, Q2 and Q3 divides the frequency distribution in to four equal parts

That is 25% of data will lie below Q1, 50% of data below Q2 and 75percent below Q3. Here Q2 is called the Median. Quartiles are obtained in almost the same way as median.

Ungrouped Data:

If the data set consist of n items and arranged in ascending order then

$$Q_1 = \left(\frac{n+1}{4}\right)^{th} \text{ item, } Q_2 = \left(\frac{n+1}{2}\right)^{th} \text{ item and } Q_3 = 3 \left(\frac{n+1}{4}\right)^{th} \text{ item}$$

Example:20

Compute quartiles for the data 25, 18, 30, 8, 15, 5, 10, 35, 40, 45.

Solution:

$$\begin{aligned} Q_1 &= \frac{(n+1)^{th}}{4} \text{ item} = \frac{(10+1)^{th}}{4} \text{ item} = (2.75)^{th} \text{ item} \\ &= 2^{nd} \text{ item} + \frac{(3)^{rd}}{4} (3^{rd} \text{ item} - 2^{nd} \text{ item}) \\ &= 8 + \frac{(3)}{4} (10 - 8) = 8 + 1.5 \end{aligned}$$

Q1= 9.5

$$\begin{aligned} Q_3 &= 3 \frac{(n+1)^{th}}{4} \text{ item} = \frac{(10+1)^{th}}{4} \text{ item} = 3 \times (2.75)^{th} \text{ item} \\ &= (8.25)^{th} \text{ item} \\ &= 2^{nd} \text{ item} + \frac{(1)}{4} (9^{th} \text{ item} - 8^{th} \text{ item}) \\ &= 35 + \frac{(1)}{4} (40 - 35) = 35 + 1.25 \end{aligned}$$

Q3 = 36.25

Continuous series:

In the case of continuous series, find the cumulative frequency and then use the interpolation formula.

- Find Cumulative frequencies
- Find $N / 4$
- Q1 class is the class interval corresponding to the value of the cumulative frequency just greater than $N / 4$
- Q3 class is the class interval corresponding to the value of the cumulative frequency just greater than $3 N / 4$

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1 \quad \text{and} \quad Q_3 = l_3 + \frac{3\left(\frac{N}{4}\right) - m_3}{f_3} \times c_3$$

Where $N = \Sigma f$ = total of all frequency values

l_1 = lower limit of the first quartile class

f_1 = frequency of the first quartile class

c_1 = width of the first quartile class

m_1 = cumulative frequency preceding the first quartile

class

l_3 = lower limit of the 3rd quartile class

f_3 = frequency of the 3rd quartile class

m_3 = cumulative frequency preceding the 3rd quartile class

c_3 = width of the third quartile class

Example:

The marks secured by group of students in their internals.

| | | | | | |
|-----------|---------|---------|---------|---------|---------|
| Class | 10 - 20 | 20 - 30 | 30 - 40 | 40 - 50 | 50 - 60 |
| Frequency | 4 | 3 | 2 | 1 | 5 |

Solution:

| Class | Frequency f | Cumulative frequency cf |
|---------|-------------|-------------------------|
| 10 - 20 | 4 | 4 |
| 20 - 30 | 3 | 7 |
| 30 - 40 | 2 | 9 |
| 40 - 50 | 1 | 10 |
| 50 - 60 | 5 | 15 |

$N / 4 = 15 / 4 = 3.75$ which lies in 10 – 20

Lies in the group 10 – 20

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1$$

$$= 10 + \frac{(3.75 - 0)}{4} \times 10 = 10 + 9.38 = \mathbf{19.38}$$

$3N / 4 = 3 \times 15 / 4 = 11.25$ which lies in 50 - 60

Therefore Q3 lies in the group 50 – 60

$$Q_3 = l_3 + \frac{\frac{3N}{4} - m_3}{f_3} \times c_3$$

$$= 50 + \frac{(11.25 - 10)}{5} \times 10$$

$$= 50 + 2.5 = \mathbf{52.5}$$

3.6.2 DECILES

These are the values which divide the total number of observation

into 10 equal parts. They are $D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9$ and D_{10} .

Ungrouped Data:

Example:

Compute the D_7 for the data: 5, 24, 36, 12, 20, and 8.

Solution:

Arranging the given data in the ascending order 5,8,12,20,24,36

$$D_5 = \frac{(5(n+1))^{th}}{10} \text{ observation} = \frac{(5(6+1))^{th}}{10} \text{ observation} = (3.5)^{th} \text{ observation}$$

$$= 3^{rd} \text{ item} + \frac{1}{2} (4^{th} \text{ item} - 3^{rd} \text{ item})$$

$$= 12 + \frac{1}{2} (20-12) = 12 + 4 = \mathbf{16}$$

Grouped Data:

Example:

Calculate the D_1 and D_7 for the given data

| | | | | | | | |
|----------------|-------|-------|-------|-------|-------|-------|-------|
| Class interval | 0 -10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
| Frequency | 5 | 7 | 12 | 16 | 10 | 8 | 4 |

Solution:

| Class interval | Frequency f | Cumulative frequency cf |
|----------------|-------------|-------------------------|
| 0 -10 | 5 | 5 |
| 10-20 | 7 | 12 |
| 20-30 | 12 | 24 |
| 30-40 | 16 | 40 |
| 40-50 | 10 | 50 |
| 50-60 | 8 | 58 |
| 60-70 | 4 | 62 |

$$D_4 = (4N / 10)^{th} \text{ item} = (4 \times 62 / 10)^{th} \text{ item} = (24.8)^{th} \text{ item}$$

This lies in the interval 30 – 40

$$D_4 = 1 + \frac{(4N / 10 - m)}{f} \times c$$

$$= 30 + \frac{(24.8-24)}{16} \times 10 = 30 + \frac{(0.8)}{16} \times 10$$

$$= 30 + 0.5 = \mathbf{30.5}$$

$$D_7 = (7N / 10)^{th} \text{ item}$$

$$= (7 \times 62 / 10)^{th} \text{ item}$$

$$= (43.4)^{th} \text{ item}$$

NOTES

This lies in the interval 40 – 50

$$\begin{aligned} D_4 &= 1 + \frac{(7N / 10 - m)}{f} \times c \\ &= 40 + \frac{(43.4 - 40)}{10} \times 10 = 30 + \frac{(3.4)}{10} \times 10 \\ &= 40 + 3.4 = \mathbf{43.4} \end{aligned}$$

3.6.3 PERCENTILE

The percentile values divide the distribution into 100 parts each containing 1 percent of the cases. The percentile (P_k) is that value of the variable up to which lie exactly k% of the total number of observation

Relationship

$$P_{25} = Q_1$$

$$P_{50} = \text{Median} = Q_2$$

$$P_{75} = \text{3rd quartile} = Q_3$$

Ungrouped Data:

Example: 24

The monthly income (in ₹1000) of 8 persons working in a factory. Find P_{30} income value 17, 21,14,36,10,25,15,29

Solution:

Arrange the data in the increasing order : 10, 14, 15, 17, 21, 25, 29, 36

$$n = 8$$

$$\begin{aligned} P_{30} &= \left(\frac{30(n+1)}{100} \right)^{\text{th}} \text{ item} \\ &= \left(\frac{30(8+1)}{100} \right)^{\text{th}} \text{ item} \\ &= \left(\frac{30 \times 9}{100} \right)^{\text{th}} \text{ item} = 2.7^{\text{th}} \text{ item} \\ &= 2^{\text{nd}} \text{ item} + 0.7(3^{\text{rd}} \text{ items} - 2^{\text{nd}} \text{ items}) \\ &= 14 + 0.7(15 - 14) \\ &= 14 + 0.7 \end{aligned}$$

$$P_{30} = \mathbf{14.7}$$

Grouped Data:

Example: 25

Find P_{53} for the following frequency distribution.

| | | | | | | | | |
|----------------|-----|------|-------|-------|-------|-------|-------|-------|
| Class interval | 0-5 | 5-10 | 10-15 | 15-20 | 20-25 | 25-30 | 30-35 | 35-40 |
| Frequency | 5 | 8 | 12 | 16 | 20 | 10 | 4 | 3 |

Solution:

| Class interval | Frequency | Cumulative frequency |
|----------------|-----------|----------------------|
| 0-5 | 5 | 5 |
| 5-10 | 8 | 13 |
| 10-15 | 12 | 25 |
| 15-20 | 16 | 41 |
| 20-25 | 20 | 61 |
| 25-30 | 10 | 71 |
| 30-35 | 4 | 75 |
| 35 - 40 | 3 | 78 |
| Total | 78 | |

$$\begin{aligned}
 P_{53} &= 1 + \frac{(53N / 10 - m)}{f} \times c \\
 &= 20 + \frac{(41.34 - 41)}{20} \times 5 = 20 + 0.335 = \mathbf{20.335}
 \end{aligned}$$

CHECK YOUR PROGRESS - 2

- 5) What is meant by median
- 6) What does cumulative frequency mean?
- 7) What is the formula for calculating quartile under ungrouped data?
- 8) ----- refers to the values which divide the total number of observation into 10 equal parts.

Measures of Central Tendency

NOTES

NOTES

3.7 SUMMARY

- When working on a given set of data, it is not possible to remember all the values in that set. But we require inference of the data given to us. This problem is solved by mean median and mode. .

- **Arithmetic mean**

1. Direct method –

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

2. Indirect method

$$\bar{x} = A + \frac{\sum_{i=1}^n d_i}{n}$$

Discrete grouped data

- 1) Direct method

$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{N}$$

- 2) Short method

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

Continuous grouped data

- 1) Direct method

$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{N},$$

- 2) Short cut method

$$\bar{x} = A + \frac{\sum_{i=1}^n f_i d_i}{N} \times C$$

Weighted average mean

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

Combined mean

$$\bar{x}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

Geometric mean-

$$\text{G.M.} = \text{Antilog } \frac{\sum_{i=1}^n \log x_i}{n}$$

Harmonic mean-

1) Grouped data

$$\text{H. M.} = \frac{N}{\sum_{i=1}^n f_i \left(\frac{1}{x_i}\right)}$$

2) Ungrouped data

$$\text{H. M.} = \frac{n}{\sum_{i=1}^n \left(\frac{1}{x_i}\right)}$$

Median

1) Ungrouped data

$\left(\frac{n+1}{2}\right)^{\text{th}}$ term if n is odd

Median = Mean $\left(\frac{n}{2}^{\text{th}} \text{ and } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ terms if n is even}\right)$

2) Grouped data

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

Mode

Ungrouped data –

The mode is the number that is repeated more often than any other

Grouped data –

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

Quartile

ungrouped data

$$Q_1 = \left(\frac{n+1}{4}\right)^{\text{th}} \text{ item, } Q_2 = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item and } Q_3 = 3 \left(\frac{n+1}{4}\right)^{\text{th}} \text{ item}$$

grouped data

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1 \quad \text{and} \quad Q_3 = l_3 + \frac{3\left(\frac{N}{4}\right) - m_3}{f_3} \times C_3$$

Deciles –

These are the values which divide the total number of observation into 10 equal parts. They are $D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9$ and D_{10} .

Percentile

The percentile values divide the distribution into 100 parts each containing 1% of the cases. The percentile (P_k) is that value of the variable up to which lie exactly $k\%$ of the total number of observation.

3.8 KEY WORDS

Mean, Arithmetic Mean, Geometric Mean, Harmonic Mean, Mode, Median, Quartile, Percentile, Deciles.

3.9 ANSWERS TO CHECK YOUR PROGRESS

1) Three most commonly used measures of central tendency are mean, median and mode.

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

3) It is strictly defined

It is based on all items

4) Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given values. If x_1, x_2, \dots, x_n are n observations

5) The median is that value of the variable which divides the group into two equal parts, one part comprising all values greater and the other all values less than median.

6) Cumulative frequency of each class is the sum of the frequency of the class and the frequencies of the pervious classes, ie adding the frequencies successively, so that the last cumulative frequency gives the total number of items.

7)

$$Q_1 = \left(\frac{n+1}{4}\right)^{\text{th}} \text{ item, } Q_2 = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item and } Q_3 = 3 \left(\frac{n+1}{4}\right)^{\text{th}} \text{ item}$$

8) Deciles

3.10 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTIONS

1. What do you understand by measures of central tendency?
2. Give two examples where (i) G.M. and H.M. would be most suitable averages.
3. Define median. Discuss its advantages and disadvantages as an average.
4. Calculate the geometric and the harmonic mean of the following series of monthly expenditure of hostel students. 125, 130, 75, 10, 45, 50, 40, 500, 150.

LONG ANSWER QUESTIONS

1. Find the median, quartile, 7th decile and 85th percentile of the frequency distribution given below:

| Mark in Maths | 0-10 | 10-20 | 50-30 | 30-40 | 40-50 | 50-60 | 60-70 | Above 70 |
|----------------|------|-------|-------|-------|-------|-------|-------|----------|
| No of students | 8 | 12 | 20 | 32 | 30 | 28 | 12 | 4 |

2. Determine median from the following data: 25, 20, 15, 45, 18, 7, 10, 38, 12

In a class of 100 students, 20 have failed and their average of marks is 5. The total marks secured by the entire class were 562. Find the average marks of those who have passed.

3. Find P_{53} for the following frequency distribution

| Class interval | 0 - 10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
|----------------|--------|-------|-------|-------|-------|-------|-------|
| Frequency | 5 | 7 | 12 | 16 | 10 | 8 | 4 |

4. Find the mode of the following data

| Class | 10 - 20 | 20 - 30 | 30 - 40 | 40 - 50 | 50 - 60 |
|-----------|---------|---------|---------|---------|---------|
| Frequency | 4 | 3 | 2 | 1 | 5 |

NOTES

3.11 FURTHER READINGS

1. Levin, Richard I. and David S. Rubin: Statistics for Management, PrenticeHall, New Delhi.
 2. Watsman Terry J. and Keith Parramor: Quantitative Methods in FinanceInternational, Thompson Business Press, London.
 3. Hooda, R. P.: Statistics for Business and Economics, Macmillan, New Delhi.
 4. Hein, L. W. Quantitative Approach to Managerial Decisions, Prentice Hall,NJ.
-

UNIT 4 - MEASURES OF DISPERSION

Measures of Dispersion

Structure

- 4.0 Introduction
 - 4.1 Objectives
 - 4.2 Measures of Dispersion
 - 4.2.1 Properties of a good measure of Dispersion
 - 4.2.2 Characteristics of Measures of Dispersion
 - 4.2.3 Classification of Measures of Dispersion
 - 4.3 Range
 - 4.4 Quartile deviation
 - 4.5 Mean Deviation
 - 4.6 Standard Deviation
 - 4.6.1 Calculation of Standard Deviation
 - 4.7 Coefficient of Variable
 - 4.8 Summary
 - 4.9 Key Words
 - 4.10 Answers to Check Your Progress
 - 4.11 Questions and Exercise
 - 4.12 Further Readings

NOTES

4.0 INTRODUCTION

Measures of central tendency, Mean, Median, Mode, etc., indicate the central position of a series. They indicate the general magnitude of the data but fail to reveal all the peculiarities and characteristics of the series. In other words, they fail to reveal the degree of the spread out or the extent of the variability in individual items of the distribution. This can be explained by certain other measures, known as 'Measures of Dispersion' or Variation.

4.1 OBJECTIVES

After going through this unit, you will,

- Learn about the measures of dispersion.
- Understand about quartile deviation, mean deviation, standard deviation.
- Come to know about coefficient of variation.

4.2 MEASURES OF DISPERSION

Dispersion is the extent till which a distribution can be stretched or squeezed. We can understand variation with the help of the following example:

| Series I | Series II | Series III |
|----------|-----------|------------|
| 10 | 2 | 10 |
| 10 | 8 | 12 |
| 10 | 20 | 8 |

Self-Instructional Material

NOTES

| | | |
|-----------------|----|----|
| $\Sigma X = 30$ | 30 | 30 |
|-----------------|----|----|

In all three series, the value of arithmetic mean is 10. On the basis of this average, we can say that the series are alike. If we carefully examine the composition of three series, we find the following differences:

- (i) In case of 1st series, three items are equal; but in 2nd and 3rd series, the items are unequal and do not follow any specific order.
- (ii) The magnitude of deviation, item-wise, is different for the 1st, 2nd and 3rd series. But all these deviations cannot be ascertained if the value of simple mean is taken into consideration.
- (iii) In these three series, it is quite possible that the value of arithmetic mean is 10; but the value of median may differ from each other. This can be understood as follows;

| Series I | Series II | Series III |
|-----------------|-----------|------------|
| 10 | 2 | 8 |
| 10 median | 8 median | 10 median |
| 10 | 20 | 12 |
| $\Sigma X = 30$ | 30 | 30 |

The value of Median' in 1st series is 10, in 2nd series = 8 and in 3rd series = 10. Therefore, the value of the Mean and Median are not identical.

- (iv) As the average remains the same, the nature and extent of the distribution of the size of the items may vary. In other words, the structure of the frequency distributions may differ even though their means are identical.

4.2.1 PROPERTIES OF A GOOD MEASURE OF DISPERSION

There are certain pre-requisites for a good measure of dispersion:

1. It should be simple to understand.
2. It should be easy to compute.
3. It should be rigidly defined.
4. It should be based on each individual item of the distribution.
5. It should be capable of further algebraic treatment.

4.2.2 CHARACTERISTICS OF MEASURES OF DISPERSION

- A measure of dispersion should be rigidly defined
- It must be easy to calculate and understand
- Not affected much by the fluctuations of observations

- Based on all observations

4.2.3 CLASSIFICATION OF MEASURES OF DISPERSION

The measure of dispersion is categorized as:

(i) An absolute measure of dispersion:

It involves the units of measurements of the observations. For example, (i) the dispersion of salary of employees is expressed in rupees, and (ii) the variation of time required for workers is expressed in hours. Such measures are not suitable for comparing the variability of the two data sets which are expressed in different units of measurements

(ii) A relative measure of dispersion:

It is a pure number independent of the units of measurements. This measure is useful especially when the data sets are measured in different units of measurement

For example, a nutritionist would like to compare the obesity of school children in India and Africa. He collects data from some of the schools in these two countries. The weight is normally measured in kilograms in India and in pounds in Africa. It will be meaningless, if we compare the obesity of students using absolute measures. So it is sensible to compare them in relative measures.

4.3 RANGE

Raw Data: A range is the most common and easily understandable measure of dispersion. It is the difference between the largest and smallest observations in the data set

$$\text{Range (R)} = L - S$$

Grouped Data: The grouped frequency distribution of values in the data set, the range is the difference between the upper class limit of the last class interval and the lower class limit of the first class interval.

Coefficient of Range: The relative measure of range is called the coefficient of range

$$\text{Coefficient of range} = (L-S) / (L + S)$$

Example:

Find the value of range and its coefficient for the following data 49, 81, 36, 64, 121, 100.

Solution:

$$L = 121 \quad ; \quad S = 36$$

$$\text{Range : } L - S = 121 - 36 = 85$$

$$\text{Co-efficient of Range} = (L-S) / (L+S) = 121-36 / 121+36$$

$$= 85 / 157 = 0.5414$$

Measures of Dispersion

NOTES

Example:

Calculate range and its coefficient from the following distribution.

| | | | | |
|-----------|--------|---------|---------|---------|
| x | 10- 15 | 15 – 20 | 20 – 25 | 25 - 30 |
| Frequency | 4 | 10 | 16 | 8 |

Solution: L = 30, S = 10

$$\text{Range} = L - S = 30 - 10 = 20$$

$$\begin{aligned} \text{Coefficient of Range} &= (L-S) / (L+S) = 30 - 10 / 30 + 10 \\ &= 20 / 40 = \mathbf{0.5} \end{aligned}$$

Merits of Range

- It is the simplest of the measure of dispersion
- Easy to calculate
- Easy to understand
- Independent of change of origin

Demerits of Range

- It is based on two extreme observations. Hence, get affected by fluctuations
- A range is not a reliable measure of dispersion
- Dependent on change of scale

4.4 QUARTILE DEVIATION

The quartiles divide a data set into quarters. The first quartile, (Q1) is the middle number between the smallest number and the median of the data. The second quartile, (Q2) is the median of the data set. The third quartile, (Q3) is the middle number between the median and the largest number. Quartile deviation is half of the difference between the first and third quartiles. Hence it is called as Semi Inter Quartile Range

Quartile deviation or semi-inter-quartile deviation is

$$Q = \frac{1}{2} \times (Q3 - Q1)$$

Coefficient of Quartile Deviation

$$\text{Coefficient of Q.D} = \frac{Q3 - Q1}{Q3 + Q1}$$

Merits of Quartile Deviation

- All the drawbacks of Range are overcome by quartile deviation
- It uses half of the data
- Independent of change of origin

- The best measure of dispersion for open-end classification

Demerits of Quartile Deviation

- It ignores fifty percent of the data
- Dependent on change of scale
- Not a reliable measure of dispersion

Example:

Calculate the quartile deviation and its coefficient for the wheat production (in Kg) of 25 acres is given as : 1120, 1240, 1320, 1040, 1080, 1200, 1440, 1360, 1680, 1730, 1785, 1342, 1960, 1880, 1755, 1720, 1600, 1470, 1750 and 1885.

Solution: Arrange the observation in increasing order:

1040, 1080, 1120, 1200, 1240, 1320, 1342, 1360, 1440, 1470, 1600, 1680, 1720, 1730, 1750, 1755, 1785, 1880, 1885, 1960.

$$\begin{aligned} Q_1 &= \text{value of } (n+1) / 4 \text{ th item} \\ &= \text{value of } (20+1) / 4 \text{ th item} = \text{value of } (5.25)\text{th item} \\ &= 5\text{th item} + 0.25 (6\text{th item} - 5\text{th item}) \\ &= 1240 + 0.25 (1320 - 1240) \\ &= 1240 + 20 = 1260 \end{aligned}$$

$$Q_1 = 1260$$

$$\begin{aligned} Q_3 &= \text{value of } 3(n+1) / 4 \text{ th item} \\ &= \text{value of } 3(20+1) / 4 \text{ th item} = \text{value of } (15.75)\text{th item} \\ &= 15\text{th item} + 0.75 (16\text{th item} - 15\text{th item}) \\ &= 1750 + 0.75 (1755 - 1750) \\ &= 1750 + 3.75 = 1753.75 \end{aligned}$$

$$Q_3 = 1753.75$$

$$\begin{aligned} Q.D &= (Q_3 - Q_1) / 2 = (1753.75 - 1260) / 2 = 492.75 / 2 \\ &= 246.875 \end{aligned}$$

$$\begin{aligned} \text{Coefficient of QD} &= (Q_3 - Q_1) / (Q_3 + Q_1) \\ &= (1753.75 - 1260) / (1753.75 + 1260) \\ &= 0.164 \end{aligned}$$

CHECK YOUR PROGRESS - 1

1. What is dispersion?
2. How to find range in grouped data?

NOTES

- 3. Mention the formula to find coefficient of quartile deviation.
- 4. State 2 merits of quartile deviation.

4.5 MEAN DEVIATION

The average deviation, it is defined as the sum of the deviations from an average divided by the number of items in a distribution. The average can be mean, median or mode. Theoretically median is the best average of choice because the sum of deviations from the median is minimum, provided signs are ignored. However, practically speaking, the arithmetic mean is the most commonly used average for calculating mean deviation and is denoted by the symbol MD.

Mean Deviation is of three types of series:

- Individual Data Series
- Discrete Data Series
- Continuous Data Series

Individual Data Series: For individual series, the Mean Deviation can be calculated using the following formula.

$$MD = \frac{1}{N} \sum |X - A| = \frac{\sum |D|}{N}$$

Where

MD = Mean deviation.

X = Variable values

A = Average of choices

N = Number of observations

Coefficient of Mean Deviation:

Mean deviation calculated by any measure of central tendency is an absolute measure. The purpose of comparing variation among different series, a relative mean deviation is required. The relative mean deviation is obtained by dividing the mean deviation by the average used for calculating mean deviation.

The Coefficient of Mean Deviation can be calculated using

Coefficient of MD = $\frac{MD}{A}$

Example:

Calculate mean deviation and coefficient of mean deviation for the following individual data:

| | | | | | |
|-------|----|----|----|-----|-----|
| Items | 28 | 72 | 90 | 140 | 210 |
|-------|----|----|----|-----|-----|

NOTES

Solution:

$$A = \frac{28 + 72 + 90 + 140 + 210}{5} = \frac{540}{5} = 108$$

| Item X | Deviation D |
|--------|-------------------|
| 28 | 80 |
| 72 | 36 |
| 90 | 18 |
| 140 | 32 |
| 210 | 102 |
| | $\Sigma D = 268$ |

$$\text{Mean Deviation} = MD = \frac{1}{N} \sum |X - A| = \frac{\sum |D|}{N} = \frac{268}{5} = 53.6$$

$$\text{Coefficient of Mean Deviation} = \frac{MD}{A} = \frac{53.6}{108} = 0.4963$$

Discrete Data Series

For discrete series, the Mean Deviation can be calculated using

$$MD = \frac{\sum f |x - Me|}{N} = \frac{\sum f |D|}{N}$$

Where, N = Number of observations.

f = Different values of frequency f.

x = Different values of items.

Me = Median.

Coefficient of Mean Deviation

The Coefficient of Mean Deviation can be calculated using the following formula.

$$\text{Coefficient of MD} = \frac{MD}{Me}$$

Example: Calculate the mean deviation and for the following discrete data

| | | | | | |
|-------|----|-----|-----|-----|-----|
| Items | 42 | 108 | 135 | 150 | 210 |
|-------|----|-----|-----|-----|-----|

NOTES

| | | | | | |
|-----------|---|----|---|---|---|
| Frequency | 6 | 15 | 3 | 3 | 9 |
|-----------|---|----|---|---|---|

Solution:

| X_i | Frequency f_i | $f_i x_i$ | $ x_i - Me $ | $f_i x_i - Me $ |
|-------|-----------------|-----------|--------------|--------------------------------|
| 42 | 6 | 252 | 93 | 558 |
| 108 | 15 | 1620 | 27 | 405 |
| 135 | 3 | 405 | 0 | 0 |
| 150 | 3 | 550 | 15 | 45 |
| 210 | 9 | 1890 | 75 | 675 |
| | $N = 36$ | | | $\Sigma f_i x_i - Me = 1683$ |

$$\text{Median} = \frac{(N + 1)\text{th item}}{2} = \frac{(5 + 1)\text{th item}}{2} = \frac{6\text{th item}}{2} = 3\text{rd item} = 135$$

$$\text{Mean Deviation} = \frac{\Sigma f |x - Me|}{N} = \frac{\Sigma f |D|}{N} = \frac{1683}{36} = 46.75$$

$$\text{Coefficient of MD} = \frac{MD}{Me} = \frac{46.75}{135} = 0.3463$$

Continuous Data Series

The method of calculating mean deviation in a continuous series is same as the discrete series. In continuous series, find a midpoint of the various classes and take deviation of these points from the average selected

$$MD = \frac{\Sigma f |x - Me|}{N} = \frac{\Sigma f |D|}{N}$$

Where N = Number of observations.

f = Different values of frequency f.

x = Different values of items.

Me = Median.

Coefficient of Mean Deviation

The Coefficient of Mean Deviation can be calculated using the following formula.

$$\text{Coefficient of MD} = \frac{\text{MD}}{\text{Me}}$$

Example:

Find out the mean deviation from the given data

| | | | | | | | | |
|---------------|------|-------|-------|-------|-------|-------|-------|-------|
| Age in years | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 |
| No of persons | 40 | 50 | 64 | 80 | 82 | 70 | 20 | 16 |

Solution:

| Items | Mid point x_i | Frequency f_i | $f_i x_i$ | $ x_i - \text{Me} $ | $f_i x_i - \text{Me} $ | Items | Mid point x_i | Frequency f_i | $f_i x_i$ |
|-------|-----------------|-----------------|--|---------------------|-------------------------|-------|-----------------|-----------------|--|
| 0-10 | 5 | 40 | 200 | 31.47 | 1258.8 | 0-10 | 5 | 40 | 200 |
| 10-20 | 15 | 50 | 750 | 21.47 | 1073.5 | 10-20 | 15 | 50 | 750 |
| 20-30 | 25 | 64 | 1600 | 11.47 | 734.08 | 20-30 | 25 | 64 | 1600 |
| 30-40 | 35 | 80 | 2800 | 1.47 | 117.6 | 30-40 | 35 | 80 | 2800 |
| 40-50 | 45 | 82 | 3690 | 9.47 | 776.54 | 40-50 | 45 | 82 | 3690 |
| 50-60 | 55 | 70 | 3850 | 19.47 | 1362.9 | 50-60 | 55 | 70 | 3850 |
| 60-70 | 65 | 20 | 1300 | 29.47 | 589.4 | 60-70 | 65 | 20 | 1300 |
| 70-80 | 75 | 16 | 1200 | 39.47 | 631.52 | 70-80 | 75 | 16 | 1200 |
| | | N = 422 | $\Sigma f_i x_i = 15390$ | | | | | | $\Sigma f_i x_i - \text{Me} = 6544.34$ |

NOTES

$$\text{Median} = \frac{\sum \text{fixi}}{N} = \frac{15390}{422} = \mathbf{36.47}$$

$$\text{Mean Deviation} = \frac{\sum f|x-Me|}{N} = \frac{\sum f|D|}{N} = \frac{6544.34}{422} = 15.5079$$

$$\text{Coefficient of MD} = \frac{MD}{Me} = \frac{15.5079}{36.47} = 0.4252$$

Merits of Mean Deviation:

- It is simple to understand and easy to compute.
- It is based on each and every item of the data.
- MD is less affected by the values of extreme items than the Standard deviation.

Demerits of Mean Deviation:

- The greatest drawback of this method is that algebraic signs are ignored while taking the deviations of the items.
- It is not capable of further algebraic treatments.
- It is much less popular as compared to standard deviation.

4.6 STANDARD DEVIATION

The concept of Standard Deviation was introduced by Karl Pearson in 1893. It is by far the most important and widely used measure of dispersion. Its significance lies in the fact that it is free from those defects which afflicted earlier methods and satisfies most of the properties of a good measure of dispersion. Standard Deviation is also known as root-mean square deviation as it is the square root of means of the squared deviations from the arithmetic mean.

The standard deviation is defined as the positive square root of the mean of the square deviations taken from the arithmetic mean of the data

Ungrouped data

$x_1, x_2, x_3 \dots x_n$ are the ungrouped data then standard deviation is calculated by there are two methods of calculating standard deviation in an individual series

- Actual mean method
- Assumed mean method

Actual Mean Method:

$$\text{Standard deviation } \sigma = \frac{\sqrt{\sum(X - \bar{X})^2}}{n}$$

Example:

Calculate the standard deviation from the following data 28, 44, 18, 30, 40, 34, 24, 22.

NOTES

Solution:

Deviations from actual mean

| Values (X) | $X - \bar{X}$ | $(X - \bar{X})^2$ |
|------------|---------------|-------------------|
| 28 | -2 | 4 |
| 44 | -14 | 196 |
| 18 | -12 | 144 |
| 30 | 0 | 0 |
| 40 | 10 | 100 |
| 34 | 4 | 16 |
| 24 | -6 | 36 |
| 22 | -8 | 64 |
| 240 | | 560 |

$$\bar{X} = \frac{240}{8} = 30$$

$$\sigma = \frac{\sqrt{\Sigma(X - \bar{X})^2}}{n} = \frac{\sqrt{560}}{8} = \sqrt{70} = 8.3666$$

Assumed Mean Method

This method is used when the arithmetic mean is fractional value. Taking deviations from fractional value would be a very difficult and tedious task. To save time and labour a short cut method is used; deviations are taken from a assumed mean.

$$\text{Standard Deviation } \sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

Example:

The marks obtained by the college students in statistics. Using the following data calculate standard deviation.

| | | | | | | | | | | |
|--------------|----|----|----|----|----|----|----|----|----|----|
| Students No: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Marks | 53 | 58 | 46 | 67 | 32 | 70 | 35 | 68 | 88 | 99 |

NOTES

Solution: Deviations from assumed mean

| Students No | Marks (x) | d = X - A (A=67) | d ² |
|---------------|-------------|-------------------|-------------------------------|
| 1 | 53 | -14 | 196 |
| 2 | 58 | -9 | 81 |
| 3 | 75 | 8 | 64 |
| 4 | 67 | 0 | 0 |
| 5 | 32 | -35 | 1225 |
| 6 | 70 | 3 | 9 |
| 7 | 35 | -32 | 1024 |
| 8 | 68 | 1 | 1 |
| 9 | 88 | 21 | 441 |
| 10 | 69 | 2 | 4 |
| n = 10 | | Σd = -55 | Σ d² = 3045 |

$$\begin{aligned}\sigma &= \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2} \\ &= \sqrt{\frac{3045}{10} - \left(\frac{-55}{10}\right)^2} = \sqrt{304.5 - 30.25} = \sqrt{274.25} \\ &= \mathbf{16.5605}\end{aligned}$$

4.6.1 CALCULATION OF STANDARD DEVIATION

Discrete series: There are three methods for calculating standard deviation in discrete series. They are

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

Actual mean method

Calculate the mean of the series. Find the deviations for various items from the means and square the deviations and multiply by the respective frequency and total the product the formula to calculate actual mean method is

NOTES

$$\sigma = \frac{\sqrt{\sum fd^2}}{\sum f}$$

If the actual mean is fractions, the calculation takes lot of time and labour; and as such this method is rarely used in practice

Assumed mean method

Here deviation is taken not from an actual mean but from an assumed mean. Also this method is used, if the given variable values are not in equal intervals.

$$\sigma = \sqrt{\frac{\sum d^2}{f} - \left(\frac{\sum d}{f}\right)^2} \quad \text{where } d = X - A, \quad N = \sum f$$

Example:

Calculate standard deviation from the following data:

| | | | | | | | | |
|---|----|----|----|----|----|----|----|----|
| X | 20 | 22 | 25 | 31 | 35 | 40 | 42 | 45 |
| f | 5 | 12 | 15 | 20 | 25 | 14 | 10 | 6 |

Solution:

Deviation from assumed mean

| x | f | d = X-A (A=31) | d ² | fd | fd ² |
|----|---------------|-------------------|----------------|------------------|--------------------------------|
| 20 | 5 | -11 | 121 | -55 | 605 |
| 22 | 12 | -9 | 81 | -108 | 972 |
| 25 | 15 | -6 | 36 | -90 | 540 |
| 31 | 20 | 0 | 0 | 0 | 0 |
| 35 | 25 | 4 | 16 | 100 | 400 |
| 40 | 14 | 9 | 81 | 126 | 1134 |
| 42 | 10 | 11 | 121 | 110 | 1210 |
| 45 | 6 | 14 | 196 | 84 | 504 |
| | N= 107 | | | Σfd = 167 | Σ fd² = 5365 |

NOTES

$$\sigma = \sqrt{\frac{\sum fd^2}{f} - \left(\frac{\sum fd}{f}\right)^2} = \sqrt{\frac{5365}{107} - \left(\frac{167}{107}\right)^2} = \sqrt{50.16 - 2.44} = 6.91$$

Step – deviation method:

If the variable values are in equal intervals, then we adopt this method

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

Example:

The frequency distribution of marks in mathematics given in the table

| | | | | | | | |
|----------------|----|----|----|----|----|----|----|
| Marks | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
| No of students | 8 | 12 | 20 | 10 | 7 | 3 | 2 |

Solution:

| Marks x | f | d= (x-50)/ 10 | fd | fd ² |
|---------|---------------|---------------|-----------------|------------------------------|
| 30 | 8 | -2 | -16 | 32 |
| 40 | 12 | -1 | -12 | 12 |
| 50 | 20 | 0 | 0 | 0 |
| 60 | 10 | 1 | 10 | 10 |
| 70 | 7 | 2 | 14 | 28 |
| 80 | 3 | 3 | 9 | 27 |
| 90 | 2 | 4 | 8 | 32 |
| | N = 62 | | Σfd = 13 | Σfd² = 141 |

$$\begin{aligned} \text{Standard Deviation } \sigma &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times C \\ &= \sqrt{\frac{141}{62} - \left(\frac{13}{62}\right)^2} \times 10 = 1.4934 \times 10 = 14.934 \end{aligned}$$

NOTES

Combined Mean and Combined Standard Deviation

Combined arithmetic mean can be computed if we know the mean and number of items in each group of the data.

$\bar{x}_1, \bar{x}_2, \sigma_1, \sigma_2$ are mean and standard deviation of two data sets having n_1 and n_2 as number of elements respectively.

$$\text{combined mean } \bar{x}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} \quad (\text{if two data sets})$$

$$\bar{x}_m = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + n_3 \bar{x}_3}{n_1 + n_2 + n_3} \quad (\text{if three data sets})$$

Example:

Particulars regarding income of two company are given below:

| | Company | |
|------------------------------|---------|------|
| | A | B |
| No.of Employees | 600 | 500 |
| Average income | 1500 | 1750 |
| Standard deviation of income | 10 | 9 |

Compute combined mean and combined standard deviation.

Solution:

Given $n_1 = 600$; $\bar{x}_1 = 1500$; $\sigma_1 = 10$

$n_2 = 500$; $\bar{x}_2 = 1750$; $\sigma_2 = 9$

$$\begin{aligned} &= \frac{600 \times 1500 + 500 \times 1750}{600 + 500} = \frac{900000 + 875000}{1100} \\ &= \mathbf{1613.6363} \end{aligned}$$

Combined Standard Deviation:

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

$$d_1 = \bar{x}_{12} - \bar{x}_1 = 1613.6363 - 1500 = 113.6363$$

$$d_2 = \bar{x}_{12} - \bar{x}_2 = 1613.6363 - 1750 = -136.3637$$

$$\begin{aligned} \sigma_{12} &= \sqrt{\frac{600(100 + 12913.209) + 500(81 + 18595.0587)}{600 + 500}} \\ &= \mathbf{124.8488} \end{aligned}$$

NOTES

Merits of Standard Deviation:

Among all measures of dispersion Standard Deviation is considered superior because it possesses almost all the requisite characteristics of a good measure of dispersion. It has the following merits:

- It is rigidly defined.
- It is based on all the observations of the series and hence it is representative.
- It is amenable to further algebraic treatment.
- It is least affected by fluctuations of sampling.

Demerits:

- It is more affected by extreme items.
- It cannot be exactly calculated for a distribution with open-ended classes.
- It is relatively difficult to calculate and understand.

4.7 COEFFICIENT OF VARIATION

The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another.

Coefficient of Variation = (Standard Deviation / Mean) * 100.

$$CV = \left(\frac{\sigma}{\bar{x}}\right) \times 100$$

The coefficient of variation (CV) is a measure of relative variability. It is the ratio of the standard deviation to the mean (average). For example, the expression “The standard deviation is 15% of the mean is a CV.

The CV is particularly useful when you want to compare results from two different surveys or tests that have different measures or values. For example, if you are comparing the results from two tests that have different scoring mechanisms. If sample A has a CV of 12% and sample B has a CV of 25%, you would say that sample B has more variation, relative to its mean.

Example:

Price of car in five years in two cities is given below:

| Price in city A | Price in city B |
|-----------------|-----------------|
| 20,00000 | 10,00000 |

NOTES

| | |
|----------|----------|
| 22,00000 | 20,00000 |
| 19,00000 | 18,00000 |
| 23,00000 | 12,00000 |
| 16,00000 | 15,00000 |

Which city has more stable prices?

Solution:

| City A | | | City B | | |
|------------------------------------|-----------------------------------|--------------------------------------|-----------------------------------|-----------------------------------|--------------------------------------|
| Price X (in lakhs) | Deviation $\bar{x} = 20$ dx | dx ² | Price Y (in lakhs) | Deviation $\bar{y} = 15$ dx | dy ² |
| 20 | 0 | 0 | 10 | -5 | 25 |
| 22 | 2 | 4 | 20 | 5 | 25 |
| 19 | -1 | 1 | 18 | 3 | 9 |
| 23 | 3 | 9 | 12 | -3 | 9 |
| 16 | -4 | 16 | 15 | 0 | 0 |
| $\Sigma x = 100$ | $\Sigma dx = 0$ | $\Sigma dx^2 = 30$ | $\Sigma y = 75$ | $\Sigma dy = 0$ | $\Sigma dy^2 = 68$ |

City A: $\bar{x} = \Sigma x / n = 100 / 5 = 20$

$$\sigma_x = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n}} = \sqrt{\frac{\Sigma dx^2}{n}} = \sqrt{\frac{30}{5}} = 2.45$$

$$C.V. (X) = \left(\frac{\sigma}{\bar{x}}\right) \times 100 = \frac{2.45}{20} \times 100 = 12.25\%$$

City B: $\bar{x} = \Sigma x / n = 75 / 5 = 15$

$$\sigma_y = \sqrt{\frac{\Sigma(y - \bar{y})^2}{n}} = \sqrt{\frac{\Sigma dy^2}{n}} = \sqrt{\frac{68}{5}} = 3.69$$

$$C.V. (Y) = \left(\frac{\sigma}{\bar{y}}\right) \times 100 = \frac{3.69}{15} \times 100 = 24.6\%$$

City A had more stable prices than City B, because the coefficient of variation is less in City A.

CHECK YOUR PROGRESS - 2

NOTES

4.8 SUMMARY

- Measures of central tendency fail to reveal the degree of the spread out or the extent of the variability in individual items of the distribution. Dispersion is the extent till which a distribution can be stretched or squeezed
- A range is the most common and easily understandable measure of dispersion. It is the difference between the largest and smallest observations in the data set.

$$\text{Coefficient of range} = (L-S) / (L + S)$$

- The quartiles divide a data set into quarters. The first quartile, (Q1) is the middle number between the smallest number and the median of the data. The third quartile, (Q3) is the middle number between the median and the largest number. Quartile deviation is half of the difference between the first and third quartiles. Hence it is called as Semi Inter Quartile Range
- The average deviation, it is defined as the sum of the deviations from an average divided by the number of items in a distribution. The average can be mean, median or mode.
- The standard deviation is defined as the positive square root of the mean of the square deviations taken from the arithmetic mean of the data.
- The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean.

4.9 KEY WORDS

Measures of dispersion, range quartile deviation, mean deviation, standard deviation, and coefficient of variable.

4.10 ANSWERS TO CHECK YOUR PROGRESS

1. Dispersion is the extent till which a distribution can be stretched or squeezed.
2. The grouped frequency distribution of values in the data set, the range is the difference between the upper class limit of the last class interval and the lower class limit of the first class interval.
3. Coefficient of Q.D = $(Q_3 - Q_1) / (Q_3 + Q_1)$
4. All the drawbacks of Range are overcome by quartile deviation
5. Coefficient of MD = $\frac{MD}{Me}$
6. The standard deviation is defined as the positive square root of the mean of the square deviations taken from the arithmetic mean of the data.
7. Coefficient of variation (CV) $CV = \left(\frac{\sigma}{\bar{x}}\right) \times 100$

4.11 QUESTIONS AND EXERCISE

NOTES

SHORT ANSWER QUESTIONS

- What is dispersion? How is it advantageous over the measures of central tendency?
- Write short notes on range
- What is Coefficient of variation? Explain
- Write about mean deviation

LONG ANSWER QUESTIONS

- Calculate mean deviation under assumed mean method

| | | | | | | | |
|----------------|----|----|----|----|----|----|----|
| Marks | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
| No of students | 16 | 24 | 40 | 20 | 14 | 6 | 4 |

- Give a detained account on standard deviation.
- Calculate the quartile deviation and its coefficient for the corn production (in Kg) of 25 acres is given as: 1100, 1340, 1370, 1050, 1780, 1200, 2440, 1390, 1480, 1780, 1783, 1542, 1970, 1680, 1775, 1320, 1680, 1770, 1780 and 1889.

4.12 FURTHER READINGS

1. Levin, Richard I. and David S. Rubin: Statistics for Management, Prentice Hall, New Delhi.
2. Watsman terry J. and Keith Parramor: Quantitative Methods in Finance International, Thompson Business Press, London.
3. Hooda, R. P.: Statistics for Business and Economics, Macmillan, New Delhi.
4. Hein, L. W. Quantitative Approach to Managerial Decisions, Prentice Hall, NJ

UNIT 5 - MOMENTS, SKWENESS AND KURTOSIS

Structure

- 5.0 Introduction
- 5.1 Objectives
- 5.2 Moments
- 5.3 Skewness
- 5.4 Kurtosis
- 5.5 Summary
- 5.6 Key Words
- 5.7 Answers to Check Your Progress
- 5.8 Questions and Exercise.
- 5.9 Further Readings

5.0 INTRODUCTION

Moments are the arithmetic means of first; second, third and so on, i.e. r^{th} power of the deviation taken from either mean or an arbitrary point of a distribution. In other words, moments are statistical measures that give certain characteristics of the distribution. Skewness means the symmetry or the lack of symmetry of a data. Skewness can be easily observed from the frequency curve. Kurtosis is all about the tails of the distribution — not the peakedness or flatness. It is used to describe the extreme values in one versus the other tail. We will learn about these in detail.

5.1 OBJECTIVES

From this unit you will

- Learn about moments, skewness and kurtosis.
- Understand the various methods to calculate them
- Know how they are used in statistics

5.2 MOMENTS

Moment word is very popular in mechanical sciences. In science moment is a measure of energy which generates the frequency. In Statistics, moments are the arithmetic means of first; second, third and so on, i.e. r^{th} power of the deviation taken from either mean or an arbitrary point of a distribution. In other words, moments are statistical measures that give certain characteristics of the distribution. In statistics, some moments are very important. Generally, in any frequency distribution, four moments are obtained which are known as first, second, third and fourth moments. These four moments describe the information about

NOTES

mean, variance, skewness and kurtosis of a frequency distribution.

Calculation of moments gives some features of a distribution which are of statistical importance. Moments can be classified in raw and central moment. Raw moments are measured about any arbitrary point A. If A is taken to be zero then raw moments are called moments about origin. When A is taken to be Arithmetic mean we get central moments. The first raw moment about origin is mean whereas the first central moment is zero. The second raw and central moments are mean square deviation and variance, respectively. The third and fourth moments are useful in measuring Skewness and Kurtosis.

Raw Moments:

Raw moments can be defined as the arithmetic mean of various powers of deviations taken from origin. The r^{th} raw moment is denoted by μ_r' , $r = 1,2,3,\dots$. then the first raw moments are given by

| Raw moments | Raw data ($d=x - A$) | Discrete data ($d=x - A$) | Continuous data ($d = (x - A) / c$) |
|-------------|------------------------|-----------------------------|---------------------------------------|
| μ_1' | $\frac{\sum d}{n}$ | $\frac{\sum fd}{N}$ | $\frac{\sum fd}{N} \times c$ |
| μ_2' | $\frac{\sum d^2}{n}$ | $\frac{\sum fd^2}{N}$ | $\frac{\sum fd^2}{N} \times c^2$ |
| μ_3' | $\frac{\sum d^3}{n}$ | $\frac{\sum fd^3}{N}$ | $\frac{\sum fd^3}{N} \times c^3$ |
| μ_4' | $\frac{\sum d^4}{n}$ | $\frac{\sum fd^4}{N}$ | $\frac{\sum fd^4}{N} \times c^4$ |

Central Moments:

Central moments can be defined as the arithmetic mean of various powers of deviation taken from the mean of the distribution. The r^{th} central moment is denoted by μ_r , $r = 1,2,3,\dots$.

| Central moments | Raw data | Discrete data | Continuous data $d' = \frac{(x - \bar{x})}{c}$ |
|-----------------|---------------------------------------|--|--|
| μ_1 | $\frac{\sum (x - \bar{x})}{n} = 0$ | $\frac{\sum f(x - \bar{x})}{N} = 0$ | $\frac{\sum fd}{N} \times c$ |
| μ_2 | $\frac{\sum f(x - \bar{x})^2}{N} = 0$ | $\frac{\sum f(x - \bar{x})^2}{N} = \sigma^2$ | $\frac{\sum fd^2}{N} \times c^2$ |
| μ_3 | $\frac{\sum (x - \bar{x})^3}{n}$ | $\frac{\sum f(x - \bar{x})^3}{N}$ | $\frac{\sum fd^3}{N} \times c^3$ |
| μ_4 | $\frac{\sum (x - \bar{x})^4}{n}$ | $\frac{\sum f(x - \bar{x})^4}{N}$ | $\frac{\sum fd^4}{N} \times c^4$ |

In general, given n observation $x_1, x_2, x_3, \dots, x_n$ the r^{th} order raw moments $r = 0, 1, 2, 3, \dots$.

$$\mu_r' = \frac{1}{N} \sum f(x - A)^r \text{ about } A$$

**Moments, Skewness and
Kurtosis**

NOTES

$$\mu'_r = \frac{\sum fx^r}{N} \text{ about origin}$$

$$\mu_r = \frac{1}{N} \sum f(x - \bar{x})^r \text{ about mean}$$

Relationship between Raw Moments and Central Moments:

Relation between moments about arithmetic mean and moments about an origin are given below

$$\mu_1 = \mu'_1 - \mu'_1 = 0$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2$$

$$\mu_3 = \mu'_3 - 3\mu'_2 \mu'_1 + 2(\mu'_1)^3$$

$$\mu_4 = \mu'_4 - 4\mu'_3 \mu'_1 + 6\mu'_2 (\mu'_1)^2 - 3(\mu'_1)^4$$

Example:

Calculate the first four moments from the following data:

| | | | | | | | | | |
|---|----|----|----|----|----|----|----|----|----|
| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| f | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 |

Solution:

| x | f | fx | d = x - \bar{x} x = 5.33 | fd | fd ² | fd ³ | fd ⁴ |
|---|----------------|-------------------|-------------------------------|------------------|-------------------------------|-------------------------------|--------------------------------|
| 0 | 10 | 0 | -5 | -50 | 250 | -1250 | 6250 |
| 1 | 20 | 20 | -4 | -80 | 320 | -1280 | 5120 |
| 2 | 30 | 60 | -3 | -90 | 270 | -810 | 2430 |
| 3 | 40 | 120 | -2 | -80 | 160 | -320 | 640 |
| 4 | 50 | 200 | -1 | -50 | 50 | -50 | 50 |
| 5 | 60 | 300 | 0 | 0 | 0 | 0 | 0 |
| 6 | 70 | 420 | 1 | 70 | 70 | 70 | 70 |
| 7 | 80 | 560 | 2 | 160 | 320 | 640 | 1280 |
| 8 | 90 | 720 | 3 | 270 | 810 | 2430 | 7290 |
| | N = 450 | Σfx = 2400 | Σd = -9 | Σfd = 150 | Σfd² = 2250 | Σfd³ = -570 | Σfd⁴ = 23130 |

NOTES

$$\bar{x} = \frac{\sum fx}{N} = \frac{2400}{450} = 5.33 \approx 5$$

$$\mu_1 = \frac{\sum fd}{N} = \frac{150}{450} = 0.33$$

$$\mu_2 = \frac{\sum fd^2}{N} = \frac{2250}{450} = 5$$

$$\mu_3 = \frac{\sum fd^3}{N} = \frac{-570}{450} = -1.266$$

$$\mu_4 = \frac{\sum fd^4}{N} = \frac{23130}{450} = 51.4$$

Example:

Calculate the first four moments about the arbitrary origin and then calculate the first four moments about the mean

| | | | | | | |
|---|-------|-------|-------|-------|-------|-------|
| x | 10-13 | 13-16 | 16-19 | 19-22 | 22-25 | 25-28 |
| f | 2 | 4 | 26 | 47 | 15 | 6 |

Solution:

| X | Mid values (m) | f | $d' = \frac{m-17.5}{3}$ | fd' | fd'^2 | fd'^3 | fd'^4 |
|-------|----------------|----------------|-------------------------|-----------------------------------|--------------------------------------|--------------------------------------|--------------------------------------|
| 10-13 | 11.5 | 2 | -2 | -4 | 8 | -16 | 32 |
| 13-16 | 14.5 | 4 | -1 | -4 | 4 | -4 | 4 |
| 16-19 | 17.5 | 26 | 0 | 0 | 0 | 0 | 0 |
| 19-22 | 20.5 | 47 | 1 | 47 | 47 | 47 | 47 |
| 22-25 | 23.5 | 15 | 2 | 30 | 60 | 120 | 240 |
| 25-28 | 26.5 | 6 | 3 | 18 | 54 | 162 | 486 |
| | | N = 100 | | $\sum fd' = 87$ | $\sum fd'^2 = 173$ | $\sum fd'^3 = 309$ | $\sum fd'^4 = 809$ |

*Moments, Skewness
and Kurtosis*

NOTES

$$\mu'_1 = \frac{\sum fd'}{N} \times c = \frac{87}{100} \times 3 = 2.61$$

$$\mu'_2 = \frac{\sum fd'^2}{N} \times c^2 = \frac{173}{100} \times 9 = 15.57$$

$$\mu'_3 = \frac{\sum fd'^3}{N} \times c^3 = \frac{309}{100} \times 27 = 83.43$$

$$\mu'_4 = \frac{\sum fd'^4}{N} \times c^4 = \frac{809}{100} \times 81 = 655.29$$

Moment about mean

$$\mu_1 = \mu'_1 - \mu'_1 = 2.61 - 2.61 = 0$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = 15.57 - (2.61)^2 = 8.76$$

$$\begin{aligned} \mu_3 &= \mu'_3 - 3\mu'_2 \mu'_1 + 2(\mu'_1)^3 = 83.43 - 3(2.61)(15.57) + 2(2.61)^3 \\ &= -2.91 \end{aligned}$$

$$\begin{aligned} \mu_4 &= \mu'_4 - 4\mu'_3 \mu'_1 + 6\mu'_2 (\mu'_1)^2 - 3(\mu'_1)^4 \\ &= 655.29 - 4(83.43)(2.61) + 6(15.57)(2.61)^2 - 3(2.61)^4 = 291.454 \end{aligned}$$

CHECK YOUR PROGRESS - 1

1. What are moments?
2. What is a raw moment?
3. Mention the relation between raw moments and central moments

5.3 SKEWNESS

Skewness means the symmetry or the lack of symmetry of a data. Skewness can be easily observed from the frequency curve. In frequency curve of the data and draw a reference line at the value of mode then if we find the curve on either side of the line have equal, that data is called symmetric.

Positive Skewness: Skewness is said to be positive when the tail of the curve of the frequency distribution elongates more on the right. Also, skewness is positive if mean, median and mode of the frequency distribution satisfy the following condition:

$$\text{Mean} > \text{Median} > \text{mode}$$

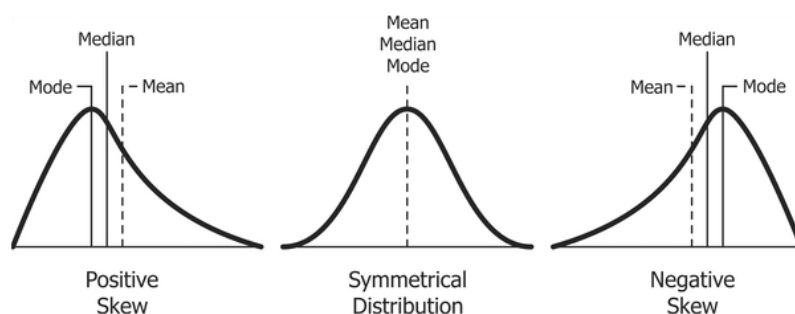
Negative Skewness: Skewness is said to be negative when the tail of the curve of the frequency distribution elongates more on the left. Also, skewness is negative if mean, median and mode of the frequency distribution satisfy the condition

$$\text{Mean} < \text{Median} < \text{Mode}$$

If the curve of the frequency distribution is symmetrical, then skewness is zero. In this case, we have the relation

Mean=Median=Mode

The figure of the symmetrical, positively skewed and negatively skewed distribution is given below:



Characteristic of a good measure of Skewness

- It should be a pure number in the sense that its value should be independent of the unit of the series and also degree of variation in the series.
- It should have zero-value, when the distribution is symmetrical.
- It should have a meaningful scale of measurement so that we could easily interpret the measured value.

Methods of ascertaining Skewness

Skewness can be studied graphically and mathematically. When we study skewness graphically, we can find out whether skewness is positive or negative or zero. This can be shown with the help of a diagram :

Mathematically skewness can be studied as :

a) Absolute Skewness

When the skewness is presented in absolute term i.e., in units, it is absolute skewness. When skewness is measured in absolute terms, then we can compare one distribution with the other if the units of measurement are same

b) Relative or Coefficient of skewness

If the value of skewness is obtained in ratios or percentages, it is called relative or coefficient of skewness.. When skewness is presented in ratios or percentages, comparison become easy. Relative measures of skewness is also called coefficient of skewness.

Mathematical measures of skewness can be calculated by :

- Karl-Pearson's Method
- Bowley's Method
- Kelly 's method

NOTES

Karl-Pearson's Method

The mean, median and mode are not equal in a skewed distribution. The Karl Pearson's measure of skewness is based upon the divergence of mean from mode in a skewed distribution. Since Mean = Mode in a symmetrical distribution, (Mean - Mode) can be taken as an absolute measure of skewness. The absolute measure of skewness for a distribution depends upon the unit of measurement.

For example, if the mean = 2.45 meters and mode = 2.14 meters, then absolute measure of skewness will be 2.45 meters - 2.14 meters = 0.31 meters. For the same distribution, if we change the unit of measurement to centimeters, the absolute measure of skewness is 245 centimeter - 214 centimeter = 31 centimeter. In order to avoid such a problem Measures Skewness and Kurtosis Karl Pearson takes a relative measure of skewness.

A relative measure, independent of the units of measurement, is defined as the Karl Pearson's Coefficient of Skewness S_k , given by

$$S_k = \frac{\text{Mean} - \text{Mode}}{S.D}$$

The sign of S_k gives the direction and its magnitude gives the extent of skewness.

If $S_k > 0$, the distribution is positively skewed, and if $S_k < 0$ it is negatively skewed.

So far we have seen that S_k is strategically dependent upon mode. If mode is not defined for a distribution we cannot find S_k . But empirical relation between mean, median and mode states that, for a moderately symmetrical distribution, we have

Mean - Mode \approx 3 (Mean - Median)

Hence Karl Pearson's coefficient of skewness is defined in terms of median as

$$S_k = \frac{3(\text{Mean} - \text{Median})}{S.D}$$

Example:

Compute the Karl Pearson's coefficient of skewness from the following data

| | | | | | | | | |
|--------------------|----|----|----|----|----|----|----|----|
| Height | 68 | 69 | 70 | 71 | 72 | 73 | 74 | 75 |
| Number of students | 10 | 18 | 30 | 42 | 35 | 28 | 16 | 8 |

NOTES

Solution:

To calculate the Mean and S.D

| Height (x) | u = x - 71 | No of students (f) | fu | fu ² |
|--------------|------------|--------------------|-----------|-----------------|
| 68 | -3 | 10 | -30 | 90 |
| 69 | -2 | 18 | -36 | 72 |
| 70 | -1 | 30 | -30 | 30 |
| 71 | 0 | 42 | 0 | 0 |
| 72 | 1 | 35 | 35 | 35 |
| 73 | 2 | 28 | 56 | 112 |
| 74 | 3 | 16 | 48 | 144 |
| 75 | 4 | 8 | 32 | 128 |
| Total | | 187 | 75 | 611 |

$$\text{Mean} = 61 + \frac{75}{187} = 61.4$$

$$\text{S.D} = \sqrt{\frac{611}{187} - \left(\frac{75}{187}\right)^2} = 1.76$$

We find that the height is a continuous variable, to find mode it is assumed that the height has been measure under the approximation that a measurement on height that is greater than 68 but less than 68.5 is consider as 68 inches, while measurement greater than or equal to 68.5 but less than 69 is taken as 69 inches. Thus the given data can be written as

| Height | No of students |
|-------------|----------------|
| 67.5 - 68.5 | 10 |
| 68.5 - 69.5 | 18 |
| 69.5 - 70.5 | 30 |
| 70.5 - 70.5 | 42 |
| 71.5 - 72.5 | 35 |
| 72.5 - 73.5 | 28 |

Moments, Skewness and Kurtosis

NOTES

| | |
|-------------|----|
| 73.5 - 74.5 | 16 |
| 74.5 - 75.5 | 8 |

The mode class is 70.5 – 71.5

$l = 70.5, \Delta_1 = 42-30 = 12, \Delta_2 = 42-35 = 7$ and $c = 1$

Therefore Mode = $70.5 + \frac{12}{12+7} \times 1 = 61.13$

Hence, the Karl Pearson’s coefficient of skewness

$$S_k = \frac{61.4 - 61.13}{1.76} = 0.153$$

Thus the distribution is positively skewed

Bowley’s Method :

This measure is based on quartiles. For a symmetrical distribution, it is seen that Q_1 and Q_3 are equidistant from median. Thus $(Q_3 - M_d) - (M_d - Q_1)$ can be taken as an absolute measure of skewness.

$$S_Q = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

Example:

Calculate the coefficient of skewness based on quartiles from the following data:

| | | | | | | | | | |
|------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| Monthly salary | 1000-1200 | 1200-1400 | 1400-1600 | 1600-1800 | 1800-2000 | 2000-2200 | 2200-2400 | 2400-2600 | 2600-2800 |
| No: of Employees | 6 | 14 | 23 | 50 | 52 | 25 | 22 | 7 | 2 |

Solution:

| Monthly salary | Frequency | Cumulative frequency |
|----------------|-----------|----------------------|
| 1000 - 1200 | 5 | 5 |
| 1200 - 1400 | 14 | 19 |
| 1400 - 1600 | 23 | 42 |
| 1600 - 1800 | 50 | 92 |
| 1800 - 2000 | 52 | 144 |
| 2000 - 2200 | 25 | 169 |

| | | |
|--------------|------------|-----|
| 2200 - 2400 | 22 | 191 |
| 2400 - 2600 | 7 | 198 |
| 2600 - 2800 | 2 | 200 |
| Total | 200 | |

Q_1 has $\frac{N}{4}$ observations or 50 observations below it . it lies in the class 1600 – 1800

$$Q_1 = l + \frac{\frac{N}{4} - c}{f} \times i = 1600 + \frac{50 - 42}{50} \times 200 = \mathbf{1632}$$

Q_2 (median) has $\frac{N}{2}$ observation or 100 observation. So it lies in the class 1800 – 2000

$$M_d = 1800 + \frac{100 - 92}{52} \times 200 = \mathbf{1830.77}$$

Q_3 has $\frac{3N}{4}$ observations or 150 observations below it . it lies in the class 2000-2200

$$Q_3 = l + \frac{\frac{3N}{4} - c}{f} \times i = 2000 + \frac{150 - 144}{25} \times 200 = \mathbf{2048}$$

$$\text{Coefficient of } S_Q = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1} = \frac{2048 + 1632 - (2 \times 1830.77)}{2048 - 1632} = \mathbf{0.044}$$

Measure of Skewness based on Moments

The measure of Skewness based on moments is denoted by β_1 and is given by

$$\beta_1 = \mu_3^2 / \mu_2^3$$

Example:

Find β_1 for the following data $\mu_1 = 0, \mu_2 = 8.76, \mu_3 = -2.91$

Solution:

$$\beta_1 = \mu_3^2 / \mu_2^3$$

$$\beta_1 = \frac{(-2.91)^2}{(8.76)^3} = \frac{8.47}{672.24} = \mathbf{0.0126}$$

NOTES

CHECK YOUR PROGRESS - 2

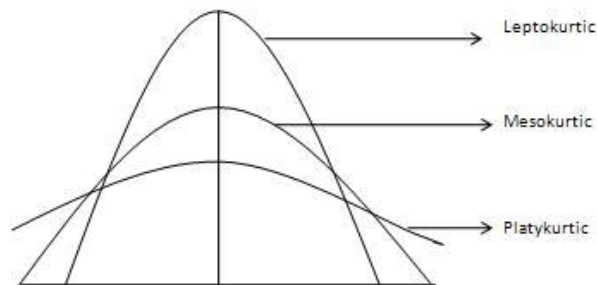
4. What are the Characteristics of a good measure of Skewness?
5. Mention the methods of ascertaining Skewness
6. State the methods of calculation of mathematical measures of Skewness.

5.4 KURTOSIS

Kurtosis is all about the tails of the distribution — not the peakedness or flatness. It is used to describe the extreme values in one versus the other tail. It is actually the measure of outliers present in the distribution.

High kurtosis in a data set is an indicator that data has heavy tails or outliers. If there is a high kurtosis, then, we need to investigate why do we have so many outliers. It indicates a lot of things, maybe wrong data entry or other things. Investigate!

Low kurtosis in a data set is an indicator that data has light tails or lack of outliers. If we get low kurtosis (too good to be true), then also we need to investigate and trim the dataset of unwanted results.



Mesokurtic: This distribution has kurtosis statistic similar to that of the normal distribution. It means that the extreme values of the distribution are similar to that of a normal distribution characteristic. This definition is used so that the standard normal distribution has a kurtosis of three.

Leptokurtic (Kurtosis > 3): Distribution is longer, tails are fatter. Peak is higher and sharper than Mesokurtic, which means that data are heavy-tailed or profusion of outliers.

Outliers stretch the horizontal axis of the histogram graph, which makes the bulk of the data appear in a narrow (“skinny”) vertical range, thereby giving the “skinniness” of a leptokurtic distribution.

Platykurtic: (Kurtosis < 3): Distribution is shorter, tails are thinner than the normal distribution. The peak is lower and broader than Mesokurtic, which means that data are light-tailed or lack of outliers.

The reason for this is because the extreme values are less than that of the normal distribution.

Measure of Kurtosis

The measure of kurtosis of a frequency distribution based moments is denoted by β_2 and is given by

$$\beta_2 = \mu_4 / \mu_2^2$$

If $\beta_2 = 3$ the distribution is said to be normal and the curve is mesokurtic.

If $\beta_2 > 3$ the distribution is said to be normal and the curve is leptokurtic.

If $\beta_2 < 3$ the distribution is said to be normal and the curve is platykurtic.

Example:

Calculate β_1 and β_2 for the following data

| | | | | | | | | | |
|---|---|----|----|----|----|----|----|----|---|
| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| f | 5 | 10 | 15 | 20 | 25 | 20 | 15 | 10 | 5 |

Solution:

$$\mu_1 = \frac{\sum fd}{N} = 0$$

$$\mu_2 = \frac{\sum fd^2}{N} = \frac{500}{125} = 4$$

$$\mu_3 = \frac{\sum fd^3}{N} = 0$$

$$\mu_4 = \frac{\sum fd^4}{N} = \frac{4700}{125} = 37.6$$

$$\beta_1 = \mu_3^2 / \mu_2^3 = 0 / 4 = 0$$

$$\beta_2 = \mu_4 / \mu_2^2 = 37.6 / 4^2 = 2.35$$

The value of β_2 is less than 3, hence the curve is platykurtic.

CHECK YOUR PROGRESS - 2

7. What is Kurtosis?
8. What is Mesokurtic?
9. Mention the measure of Kurtosis

5.5 SUMMARY

- Generally, in any frequency distribution, four moments are obtained which are known as first, second, third and fourth moments. These four moments describe the information about mean, variance, Skewness and kurtosis of a frequency distribution.

NOTES

- Raw moments can be defined as the arithmetic mean of various powers of deviations taken from origin. Central moments can be defined as the arithmetic mean of various powers of deviation taken from the mean of the distribution.
- Skewness means the symmetry or the lack of symmetry of a data. Skewness can be easily observed from the frequency curve.
- Skewness is said to be positive when the tail of the curve of the frequency distribution elongates more on the right. Skewness is said to be negative when the tail of the curve of the frequency distribution elongates more on the left.
- Kurtosis is all about the tails of the distribution — not the peakedness or flatness. It is used to describe the extreme values in one versus the other tail. It is actually the measure of outliers present in the distribution.
- **High kurtosis** in a data set is an indicator that data has heavy tails or outliers. **Low kurtosis** in a data set is an indicator that data has light tails or lack of outliers.

5.6 KEY WORDS

Moments ,Raw moments , Central moments ,Skewness , Positive Skewness , Negative Skewness , Absolute Skewness , Relative or Coefficient of Skewness , Kurtosis , High Kurtosis , Low Kurtosis , Mesokurtic , Platykurtic.

5.7 ANSWERS TO CHECK YOUR PROGRESS

1. Moments are the arithmetic means of first; second, third and so on, i.e. r^{th} power of the deviation taken from either mean or an arbitrary point of a distribution. In other words, moments are statistical measures that give certain characteristics of the distribution.
2. Raw moments can be defined as the arithmetic mean of various powers of deviations taken from origin.
3. Relation between moments about arithmetic mean and moments about an origin are given below

$$\mu_1 = \mu'_1 - \mu'_1 = 0$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2$$

$$\mu_3 = \mu'_3 - 3\mu'_2 \mu'_1 + 2(\mu'_1)^3$$

$$\mu_4 = \mu'_4 - 4\mu'_3 \mu'_1 + 6\mu'_2 (\mu'_1)^2 - 3(\mu'_1)^4$$

4. It should be a pure number in the sense that its value should be independent of the unit of the series and also degree of variation in the series.
 - It should have zero-value, when the distribution is symmetrical.
 - It should have a meaningful scale of measurement so that we

NOTES

could easily interpret the measured value.

5. Absolute Skewness and Relative or Coefficient of skewness
6. Karl-Pearson's Method, Bowley's Method, Kelly 's method
7. Kurtosis is all about the tails of the distribution — not the peakedness or flatness. It is used to describe the extreme values in one versus the other tail. It is actually the measure of outliers present in the distribution.
8. This distribution has kurtosis statistic similar to that of the normal distribution. It means that the extreme values of the distribution are similar to that of a normal distribution characteristic. This definition is used so that the standard normal distribution has a kurtosis of three.
9. The measure of kurtosis of a frequency distribution based moments is denoted by β_2 and is given by

$$\beta_2 = \mu_4 / \mu_2^2$$

- If $\beta_2 = 3$ the distribution is said to be normal and the curve is Mesokurtic.
- If $\beta_2 > 3$ the distribution is said to be normal and the curve is Leptokurtic.
- If $\beta_2 < 3$ the distribution is said to be normal and the curve is Platykurtic.

5.8 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTIONS

1. What are the measures of Skewness?
2. What is Kurtosis? What is the measure of measuring kurtosis?
3. Define moments. Distinguish between raw moment and central moment

LONG ANSWER QUESTIONS

1. Distinguish between Skewness and kurtosis and bring out their importance in describing frequency distribution
2. From the following table calculate the Karl – Pearson's coefficient of Skewness

| | | | | | | | |
|--------------|-----|-----|-----|-----|-----|-----|-----|
| Daily wages | 150 | 200 | 250 | 300 | 350 | 400 | 450 |
| No of People | 3 | 25 | 18 | 16 | 4 | 5 | 6 |

3. Using moments calculate β_1 and β_2 from the following data:

| | | | | | |
|-----------|-------|--------|---------|---------|---------|
| Size | 70-90 | 90-110 | 110-130 | 130-150 | 150-170 |
| Frequency | 8 | 11 | 18 | 9 | 4 |

NOTES

5.9 FURTHER READINGS

1. Levin, Richard I. and David S. Rubin: Statistics for Management, PrenticeHall, New Delhi.
2. Watsman terry J. and Keith Parramor: Quantitative Methods in Finance International, Thompson Business Press, London.
3. Hooda, R. P.: Statistics for Business and Economics, Macmillan, New Delhi.
4. Hein, L. W. Quantitative Approach to Managerial Decisions, Prentice Hall, NJ

UNIT 6 - CORRELATION ANALYSES

Correlation Analyses

NOTES

Structure

- 6.0 Introduction
- 6.1 Objectives
- 6.2 Correlation
- 6.3 Linear Correlation
- 6.4 Types of Correlation
- 6.5 Scatter Diagram
- 6.6 Two – Way table
- 6.7 Pearson’s Correlation Coefficient
- 6.8 Spearman’s rank Correlation Coefficient
- 6.9 Properties of Correlation Coefficient
- 6.10 Summary
- 6.11 Key Words
- 6.12 Answer to Check Your Progress
- 6.13 Questions and Exercise
- 6.14 Further Readings

6.0 INTRODUCTION

In our day to day life, we find many situations when a mutual relationship exists between two variables i.e., with change (fall or rise) in the value of one variable there may be change (fall or rise) in the value of other variable. For example, as price of a commodity increases the demand for the commodity decreases. In the increase in the levels of pressure, the volume of a gas decreases at a constant temperature. These facts indicate that there is certainly some mutual relationships that exist between the demand of a commodity and its price, and pressure and volume. Such association is studied in correlation analysis. The correlation is a statistical tool which measures the degree or intensity or extent of relationship between two variables and correlation analysis involves various methods and techniques used for studying and measuring the extent of the relationship between the two variables.

6.1 OBJECTIVES

After studying this chapter students will be able to understand

- Understand the concept of scatter Diagram
- Concept of Karl Pearson’s correlation co-efficient and the methods of computing it.
- Spearman’s Rank correlation co-efficient

Self-Instructional Material

NOTES

6.2 CORRELATION

Correlation is a statistical technique which measures and analyses the degree or extent to which two or more variables fluctuate with reference to one another. It denotes the inter-dependence amongst variables. The degrees are expressed by a coefficient which ranges between -1 to +1. The direction of change is indicated by + or - signs; the former, refers to the movement in the same direction and the later, in the opposite direction. An absence of correlation is indicated by zero. Correlation thus expresses the relationship through a relative measure of change and it has nothing to do with the units in which the variables are expressed.

6.3 LINEAR CORRELATION

If the amount of change in one variable tends to bear constant ratio to the amount of change in the other variable then the correlation is said to be linear. For example,

| | | | | | |
|---|----|-----|-----|-----|-----|
| X | 5 | 10 | 15 | 20 | 25 |
| Y | 90 | 170 | 230 | 310 | 420 |

6.4 TYPES OF CORRELATION

There are three important types of correlation. They are

1. Positive and Negative correlation
2. Simple, Partial and Multiple correlation
3. Linear and Non-Linear correlation

1. Positive and Negative correlation

Correlation is classified according to the direction of change in the two variables. In this regard, the correlation may either be positive or negative.

Positive correlation refers to the change (movement)of variables in the same direction. Both the variables are increased or decreased in the same direction, it is called positive correlation. It is otherwise called as direct correlation. For example, a positive correlation exists between ages of husband and wife, height and weight of a group of individuals, increase in rainfall and production of paddy, increase in the offer and sales.

Negative correlation refers to the change (movement) of variables in the opposite direction. In other words, an increase (decrease) in the value of one variable is followed by a decrease (increase) in the value of the other is said to be negative correlation. It is otherwise called increase correlation. For example, a negative correlation exists between price and demand, yield of crop and price.

The following expels illustrate the concept of positive correlation

and negative correlation.

Positive correlation

| | | | | | | | |
|---|----|----|----|----|----|----|----|
| X | 5 | 7 | 9 | 11 | 16 | 20 | 28 |
| y | 20 | 26 | 35 | 37 | 48 | 50 | 55 |

Negative Correlation

| | | | | | |
|---|----|----|----|----|----|
| X | 14 | 17 | 23 | 35 | 46 |
| y | 16 | 12 | 10 | 9 | 5 |

2. Simple, Partial and Multiple Correlations

Simple correlation is a measure used to determine the strength and the direction of the relationship between two variables, X and Y. A simple correlation coefficient can range from -1 to 1 . However, maximum (or minimum) values of some simple correlations cannot reach unity (i.e., 1 or -1).

When we study only two variables, the relationship is described as simple correlation; example, quantity of money and price level, demand and price, etc. But in a multiple correlation we study more than two variables simultaneously; example, the relationship of price, demand and supply of a commodity.

The study of two variables excluding some other variables is called partial correlation. For example, we study price and demand, eliminating the supply side.

3. Linear and Non-Linear Correlation

Linear correlation is a measure of the degree to which two variables vary together, or a measure of the intensity of the association between two variables.

If the ratio of change between two variables is uniform, then there will be linear correlation between them. Consider the following.

| | | | | |
|---|---|----|----|----|
| X | 6 | 12 | 18 | 24 |
| Y | 5 | 10 | 15 | 20 |

The ratio of change between the variables is same.

In a curvilinear or non linear correlation, the amount of change in one variable does not bear a constant ratio of the amount of change in the other variables. The graph of non-linear or curvilinear relationship will form a curve.

In majority of cases, we find curvilinear relationship, which is a

NOTES

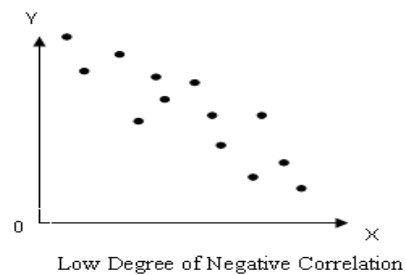
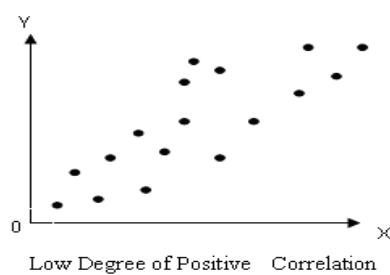
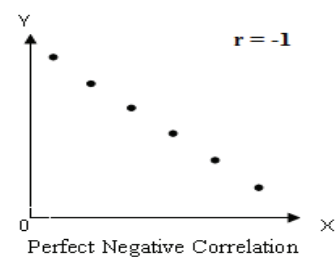
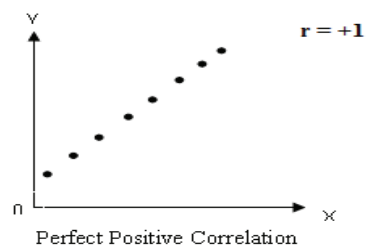
complicated one, so we generally assume that the relationship between the variables under the study is linear. In social sciences, linear correlation is rare, because the exactness is not as perfect as in natural sciences.

CHECK YOUR PROGRESS- 1

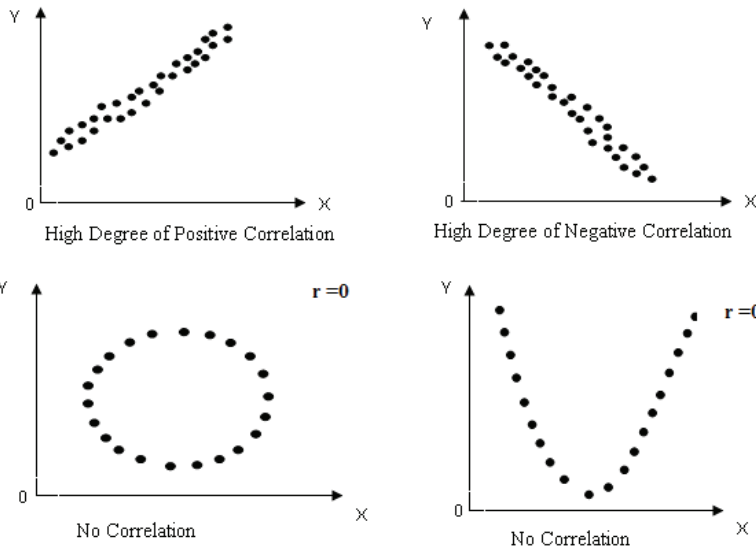
1. What is correlation?
2. Define linear correlation?
3. List out the different types of correlation?

6.5 SCATTER DIAGRAM

It is simple and attractive method of diagrammatic representation. In this method, the given data are plotted on a graph sheet in the form of dots. The x variables are plotted on the horizontal axis and y variables on the vertical axis. Now we can know the scatter or concentration of the various points. This will show the type of correlation.



NOTES



6.6 TWO-WAY TABLE

A two-way table (also called a contingency table) is a useful tool for examining relationships between categorical variables; the entries in the cells of two-way table can be frequency counts or relative frequencies (just like a one-way table).

| | Dance | Sports | TV | Total |
|-------|-------|--------|----|-------|
| Men | 2 | 10 | 8 | 20 |
| Women | 16 | 6 | 8 | 30 |
| Total | 18 | 16 | 16 | 50 |

Above a two-way table shows the favourite leisure activities for 50 adults-20 men and 30 women. Because entries in the table are frequency counts, the table is a frequency table.

6.7 PEARSON'S CO-EFFICIENT OF CORRELATION

Karl Pearson (1867-1936), the British biometrician suggested this method. It is popularly known as Pearson's co-efficient of correlation. It is mathematical method for measuring the magnitude of linear relationship between two variables.

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance.

NOTES

(a) Arithmetic mean Method

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

Example:

Find Pearson's Co-efficient of correlation from the following data

| | | | | | | | |
|--------|----|----|----|----|----|-----|-----|
| Sales | 15 | 18 | 22 | 28 | 32 | 46 | 52 |
| Profit | 52 | 66 | 78 | 87 | 96 | 125 | 141 |

Solution

Let the sales be denoted by x and the profit by y.

Computation of coefficients of correlation

| X | X - \bar{X} | X ² | Y | Y - \bar{Y} | Y ² | XY |
|------------------|----------------------|------------------------|------------------|--------------------|--------------------------|-----------------------|
| 15 | -15.43 | 238.98 | 52 | -40.14 | 1611.22 | 619.36 |
| 18 | -12.43 | 154.50 | 66 | -26.14 | 683.30 | 324.92 |
| 22 | -8.43 | 71.06 | 78 | -14.14 | 199.94 | 119.20 |
| 28 | -2.43 | 5.90 | 87 | -5.14 | 26.42 | 12.49 |
| 32 | 1.57 | 2.46 | 96 | 3.86 | 14.90 | 6.06 |
| 46 | 15.57 | 242.42 | 125 | 32.86 | 1079.78 | 511.63 |
| 52 | 21.57 | 465.26 | 141 | 48.86 | 2387.30 | 1053.91 |
| $\sum x$ =213 | $\sum x = -$ 0.01 | $\sum x^2$ =1179.68 | $\sum y$ =645 | $\sum y =$ 0.02 | $\sum y^2 =$ 6,002.86 | $\sum xy$ =2647.57 |

NOTES

$$X = \sum x / N = 213 / 7 = 30.43$$

$$Y = \sum y / N = 645 / 7 = 92.14$$

$$\begin{aligned} \sum x^2 &= 1179.68, \sum y^2 = 6002.86, \sum xy = 2647.57 \\ r &= \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{2647.57}{\sqrt{1179.68 \times 6002.86}} = \frac{2647.57}{\sqrt{1179.68 \times 6002.86}} \\ &= \frac{2647.57}{34.35 \times 77.48} = \frac{2647.57}{2661.44} = 0.99 \end{aligned}$$

Therefore, there is a high degree positive correlation between the x and y.

6.8 SPEARMEN'S RANK CORRELATION COEFFICIENT

In statistics, Spearman's rank correlation coefficient or Spearman's rho, named after Charles Spearman and often denoted by the Greek letter ρ (rho) or as r_s is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function.

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or -1 occurs when each of the variables is a perfect monotone function of the other.

$$r_s = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

Spearman's coefficient is appropriate for both continuous and discrete ordinal variables. Both Spearman's can be formulated as special cases of a more general correlation coefficient.

Example:

Two faculty members ranked 12 candidates for scholarships. Calculate the spearman rank correlation coefficient.

| Candidate | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-------------|---|----|----|---|----|----|----|----|----|----|
| Professor A | 8 | 12 | 6 | 4 | 9 | 15 | 8 | 7 | 16 | 13 |
| Professor B | 9 | 16 | 10 | 8 | 14 | 19 | 12 | 11 | 20 | 17 |

NOTES

Solution

| R_x | R_y | $d = R_x - R_y$ | d^2 |
|-------|-------|-----------------|------------------|
| 8 | 9 | -1 | 1 |
| 12 | 16 | -4 | 16 |
| 6 | 10 | -4 | 16 |
| 4 | 8 | -4 | 16 |
| 9 | 5 | 4 | 16 |
| 15 | 10 | 5 | 25 |
| 8 | 7 | 1 | 1 |
| 7 | 11 | -4 | 16 |
| 16 | 15 | 1 | 1 |
| 13 | 18 | -5 | 25 |
| | | | $\sum d^2 = 133$ |

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6(133)}{10(100-1)} = 1 - \frac{798}{990}$$

$$= 1 - 0.8060$$

$$r = 0.194$$

CHECK YOUR PROGRESS - 2

4. What are the uses of scattered diagram?
5. Write the formula to calculate spearman's rank correlation?

6.9 PROPERTIES OF CORRELATION CO-EFFICIENT

1. Coefficient of Correlation lies between -1 and +1: The coefficient of correlation cannot take value less than -1 or more than one +1. Symbolically, $-1 \leq r \leq +1$ or $|r| < 1$.
2. Coefficients of Correlation are independent of Change of Origin: This property reveals that if we subtract any constant from all the values of X and Y, it will not affect the coefficient of correlation.
3. Coefficients of Correlation possess the property of symmetry: The degree of relationship between two variables is symmetric.
4. **Coefficient of Correlation is independent of Change of Scale:** This property reveals that if we divide or multiply all the values

NOTES

of X and Y, it will not affect the coefficient of correlation.

5. The value of the coefficient of correlation shall always lie between +1 and -1.
6. When $r = +1$, then there is perfect positive correlation between the variables.
7. When $r = -1$, then there is perfect negative correlation between the variables.
8. When $r = 0$, then there is no relationship between the variables.

The third formula given above, that is

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

It is easy to calculate, and it is not necessary to calculate the standard deviation of X and Y series separately.

6.10 SUMMARY

- The term correlation refers to the degree of relationship between two or more variables.
- Scatter diagram is a graphic device for finding correlation between two variables.
- Karl Pearson correlation coefficient $r(x,y) = r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$
- Correlation coefficient r lies between -1 and 1 . (i.e) $-1 \leq r \leq 1$
- When $r=+1$, then the correlation is perfect positive
- When $r=-1$, then the correlation is perfect negative
- When $r=0$, then there is no relationship between the variables, (i.e) the variables are uncorrelated.
- Spearman's Rank correlation deals with qualitative characteristics.

6.11 KEY WORDS

Correlation, Spearman's Rank correlation, Pearson correlation, Correlation Coefficient, Scattered Diagram

6.12 ANSWER TO CHECK YOUR PROGRESS

1. The term correlation refers to the degree of relationship between two or more variables
2. Linear correlation is a measure of the degree to which two variables vary together, or a measure of the intensity of the association between two variables
3. Positive and Negative correlation, Simple, Partial and Multiple

NOTES

correlation, Linear and Non-Linear correlation

4. Scatter diagram is a graphic device for finding correlation between two variables

$$5. r_s = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

6.13 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTIONS

1. Calculate the coefficient of correlation from the following data: $\sum X=50, \sum Y=-30, \sum X^2 =290, \sum Y^2 =300, \sum XY=-115, N=10$
2. The following data pertains to the marks in subjects *A* and *B* in a certain examination. Mean marks in *A* = 39.5, Mean marks in *B*= 47.5 standard deviation of marks in *A* =10.8 and Standard deviation of marks in *B*= 16.8. coefficient of correlation between marks in *A* and marks in *B* is 0.42. Give the estimate of marks in *B* for candidate who secured 52 marks in *A*.
3. What is scattered diagram and explain it?

LONG ANSWER QUESTIONS

1. A random sample of recent repair jobs was selected and estimated cost, actual cost were recorded. Calculate the value of Spearman's correlation

| | | | | | | | | | |
|----------------|----|----|----|----|----|----|----|----|----|
| Estimated cost | 70 | 68 | 67 | 55 | 60 | 75 | 63 | 60 | 72 |
| Actual cost | 65 | 65 | 80 | 60 | 68 | 75 | 62 | 60 | 70 |

2. Distinguish between Karl Pearson's coefficient and Spearman's correlation coefficient
3. Explain the types of correlation with examples

6.14 FURTHER READINGS

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. SahityaBhawan Publishers andDistributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing CompanyLtd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw HillPublishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., NewDelhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., NewDelhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.

UNIT 7 - REGRESSION ANALYSIS

Structure

- 7.0 Introduction
- 7.0 Objectives
- 7.0 Regression
- 7.0 Linear Regression
- 7.0 Types of Regression
 - 7.1 Regression Equation of Y on X
 - 7.2 Regression Equation of X on Y
- 7.0 Curve fitting by the Method of Least square
- 7.0 Derivations of Regression Equation
- 7.0 Properties of Correlation Coefficient
- 7.0 Summary
- 7.0 Key Words
- 7.0 Answer to Check Your Progress
- 7.0 Questions and Exercise
- 7.0 Further Readings

7.0 INTRODUCTION

Regression means stepping back or going back. It was first used by Francis Galton in 1877. He studied the relationship between the height of father and their sons. The study revealed that

- Tall fathers have tall sons and short fathers have short sons.
- The mean height of the sons of tall father is less than mean height of their fathers.
- The mean height of sons of short fathers is more than the mean height of their fathers.

The tendency to going back was called by Galton as 'Line of Regression'. This line describing the average relationship between two variables is known as the line of Regression.

In statistical modelling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modelling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning.

NOTES

7.1 OBJECTIVES

After studying this chapter students will be able to understand

- Concept of Regression and Regression coefficients
- Types of regression equations
- Regression lines both x on y and y on x

7.2 REGRESSION

Regression analysis refers to assessing the relationship between the outcome variable and one or more variables. The outcome variable is known as the dependent or response variable and the risk elements, and cofounders are known as predictors or independent variables. The dependent variable is shown by “y” and independent variables are shown by “x” in regression analysis.

7.3 LINEAR REGRESSION

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

7.4 TYPES OF REGRESSION EQUATIONS

The Regression Equation is the algebraic expression of the regression lines. It is used to predict the values of the dependent variable from the given values of independent variables. As there are two regression lines, there are two regression equations. For the two variables X and Y, there are two regression equations. They are.

- **Regression equation of X on Y.**
- **Regression equation of Y on X.**

7.4.1 Regression Equation of X on Y

The straight line equation is $X = a + by$

Here a and b are unknown constants, which determines the position. The constant a is the intercept on the other value; the constant b is the slope; the following two normal equations are derived;

$$\sum x = na + b\sum y$$

$$\sum xy = a\sum x + b\sum y^2$$

The Regression equation X on Y is used to find out the values of X for given value of Y.

7.4.2 Regression Equation of Y on X

The straight line equation is $Y = a + bx$

The following two normal equations are derived

NOTES

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

The Regression equation Y on X is used to ascertain the value of y for a given value of x.

Example:

Find out the regression equation, x on y and y on x from the following data:

| | | | | | | | |
|---|----|----|----|----|----|----|----|
| X | 15 | 20 | 25 | 30 | 35 | 40 | 45 |
| y | 8 | 14 | 20 | 26 | 32 | 38 | 44 |

Solutions

| x | y | X ² | Y ² | xy |
|----------------|----------------|-------------------|-------------------|------------------|
| 15 | 8 | 225 | 64 | 120 |
| 20 | 14 | 400 | 196 | 280 |
| 25 | 20 | 625 | 400 | 500 |
| 30 | 26 | 900 | 676 | 780 |
| 35 | 32 | 1225 | 1024 | 1120 |
| 40 | 38 | 1600 | 1444 | 1520 |
| 45 | 44 | 2025 | 1936 | 1960 |
| $\sum x = 210$ | $\sum y = 182$ | $\sum x^2 = 7000$ | $\sum y^2 = 5740$ | $\sum xy = 6300$ |

$$\sum x = 210; \sum y = 182; \sum x^2 = 7000; \sum y^2 = 5740; \sum xy = 6300$$

Regression equation x on y is $y = a + by$

Hence

$$\sum x = na + b\sum y$$

$$\sum xy = a\sum x + b\sum y^2$$

$$210 = 7a + 182b$$

(1)

$$6300 = 182a + 5740b$$

(2)

NOTES

Multiplying equation (1) by 26

$$5460 = 182a + 4732b$$

(3)

$$6300 = 182a + 5740b$$

(4)

Deducting (3) from Equation (2)

$$6300=182a+5740b$$

(4)

$$5460=182a+4732b$$

(3)

$$\begin{array}{r} (-) \quad (-) \quad (-) \\ \hline 840=0 + 1008b \end{array}$$

Therefore,

$$b=840/1008=0.83$$

Substituting the value of b in Eq.(1)

$$210=7a+(182 \times 0.83)$$

$$210=7a+151.06$$

$$7a + 151.06=210$$

$$7a=210-151.06$$

$$7a=58.94$$

$$a= 8.42$$

Hence,

$$x=a+by$$

$$x=8.42 + 0.83 y$$

Regression Eq of y on x

$$y=a+bx$$

Hence,

$$\sum y = Na + b \sum y^2$$

$$182=7a+210b$$

(1)

$$6300=210a+700b$$

(2)

Multiplying Eq.(1) by 30

$$5460=210a+6300b$$

(3)

$$6300=210a+7000b$$

(4)

Deducting Eq.(4) from Eq.(3)

$$6300=210a+7000b$$

(3)

$$\underline{5460=210a+6300b}$$

(4)

$$840=0 + 700b$$

$$700=840$$

$$b=840/700$$

$$=1.2$$

Substituting the value of b in Eq.(1)

$$182=7a+(120 \times 1.2)$$

$$182=7a+252$$

$$7a+252=182$$

$$7a=182-252$$

$$7a= -70$$

$$a= -10$$

Therefore,

$$y= -10+1.$$

7.5 CURVE FITTING BY THE METHOD OF LEAST SQUARE

1. Curve Fitting

Curve fitting is the process of introducing mathematical relationships between dependent and independent variables in the form of an equation for a given set of data.

2. Method of Least Squares

The method of least squares helps us to find the values of unknowns a and b in such a way that the following two conditions are satisfied:

- The sum of the residual (deviations) of observed values of Y and corresponding expected (estimated) values of Y will be zero. $\sum(Y-Y^{\wedge})=0$ $\sum(Y-Y^{\wedge})=0$.
- The sum of the squares of the residual (deviations) of observed values of Y and corresponding expected values (Y^{\wedge}) should be at least $\sum(Y-Y^{\wedge})^2$.

3. Fitting of a Straight Line:

A straight line can be fitted to the given data by the method of least squares. The equation of a straight line or least square line is $Y=a+bX$, where a and b are constants or unknowns.

To compute the values of these constants we need as many equations as

NOTES

the number of constants in the equation. These equations are called normal equations. In a straight line there are two constants a and b so we require two normal equations.

Normal Equation for 'a' $\sum Y = na + b\sum X$

Normal Equation for 'b' $\sum XY = a\sum X + b\sum X^2$

Example:

The given example explains how to find the equation of a straight line or a least square line by using the method of least square, which is very useful in statistics as well as in mathematics.

| | | | | | |
|---|---|---|---|---|---|
| X | 1 | 2 | 3 | 4 | 5 |
| y | 2 | 5 | 3 | 8 | 7 |

Solution

| X | Y | XY | X ² | 1.1+1.3X | Y- \hat{Y} |
|-------------|-------------|--------------|----------------|--------------|--------------------|
| 1 | 2 | 2 | 1 | 2.4 | -0.4 |
| 2 | 5 | 10 | 4 | 3.7 | 1.3 |
| 3 | 3 | 9 | 9 | 5.0 | -2 |
| 4 | 8 | 32 | 16 | 6.3 | 1.7 |
| 5 | 7 | 35 | 25 | 7.6 | -0.6 |
| $\sum X=15$ | $\sum Y=25$ | $\sum XY=88$ | $\sum X^2=55$ | Trend values | $\sum(Y- \hat{Y})$ |

The equation of least square line $Y=a + bx$

Normal equation for 'a' $\sum Y=na + b \sum X$ 25
 $= 5a + 15b$
 $\sum Y = na+b\sum X$ 25=5a+15b ---- (1)

Normal equation for 'b' $\sum XY=a\sum X+b\sum X^2$ 88=15a+55b ----(2)

Eliminate a from equation (1) and (2),multiply equation (2) by 3 and subtract from equation (2). Thus we get the values of a and b.

Here $a=1.1$ and $b=1.3X$

For the trends values, put the values of X in the above equation

7.6 DERIVATIONS OF REGRESSION EQUATION

1. When deviations are taken from Arithmetic means of X and Y

The above method of finding out the regression equation is difficult. Instead, we can use the deviations of X and Y observations from their respective averages.

(i) Regression equation of X on Y

The regression equation of Y on X can also be expressed in the following form-

$$X - \bar{X} = r \sigma_x / \sigma_y (Y - \bar{Y})$$

Here, \bar{X} is the average of X observations and \bar{Y} is the average of Y observations.

$r \sigma_x / \sigma_y$ is the regression coefficient of X on Y and is denoted by b_{xy} . b_{xy} measures the amount of change in X corresponding to a unit change in Y.

$$r \sigma_x / \sigma_y = b_{xy} = \frac{\sum xy}{\sum y^2}$$

Where $x=(X-\bar{X})$ and $y=(Y-\bar{Y})$

(ii) Regression equation of Y on X

The regression equation of Y on X can also be expressed in the following form-

$$Y - \bar{Y} = r \sigma_y / \sigma_x (X - \bar{X})$$

$r \sigma_y / \sigma_x$ is the regression coefficient of Y on X and is denoted by b_{yx} . b_{yx} measures the amount of change in Y corresponding to a unit change in X.

$$r \sigma_y / \sigma_x = b_{yx} = \frac{(\sum xy)}{\sum x^2}$$

We can calculate the coefficient of correlation which is the geometric mean of the two regression coefficients (b_{xy} & b_{yx}) i.e.

$$r = \sqrt{(b_{xy}) \times (b_{yx})}$$

2. When deviations are Taken from Assumed Mean

When instead of using actual means of X and Y observations, we use any arbitrary item (in the observation) as the mean.

We consider taking deviations of X and Y values from their respective assumed means.

The formula for calculating regression coefficient when regression is Y on X is as follows:

NOTES

$$r \frac{\sigma_y}{\sigma_x} = b_{yx} = \frac{\sum(dx)(dy) - \frac{(\sum dx)(\sum dy)}{N}}{\sum(dx)^2 - \frac{(\sum dx)^2}{N}}$$

Where $dx = (X - A_x)$ { A_x = assumed mean of X observations} and $dy = (Y - A_y)$ { A_y = assumed mean of Y observations}

The formula for calculating regression coefficient when regression is X on Y is as follows:

$$r \frac{\sigma_x}{\sigma_y} = b_{xy} = \frac{\sum(dx)(dy) - \frac{(\sum dx)(\sum dy)}{N}}{\sum(dy)^2 - \frac{(\sum dy)^2}{N}}$$

In the case of **Grouped frequency distribution**, the regression coefficients are calculated from the bivariate frequency table (or correlation table).

The formula for calculating regression coefficient (in case of grouped frequency distribution) when regression is of Y on X is as follows-

$$r \frac{\sigma_y}{\sigma_x} = b_{yx} = \frac{\sum f(dx)(dy) - \frac{(\sum f dx)(\sum f dy)}{N}}{\sum (f dx)^2 - \frac{(\sum f dx)^2}{N}} \times \frac{h_y}{h_x}$$

Where h_x = class interval of X variable and h_y = class interval of Y variable

The formula for calculating regression coefficient (in case of grouped frequency distribution) when regression is of X on Y is as follows-

$$r \frac{\sigma_x}{\sigma_y} = b_{xy} = \frac{\sum f(dx)(dy) - \frac{(\sum f dx)(\sum f dy)}{N}}{\sum (f dy)^2 - \frac{(\sum f dy)^2}{N}} \times \frac{h_x}{h_y}$$

Check your Progress - 1

1. What are regression coefficients?
2. What is the formula used to calculate assumed mean?
3. What is method of least square?

7.7 PROPERTIES OF REGRESSION EQUATION

The constant 'b' in the regression equation ($Y_e = a + bX$) is called as the **Regression Coefficient**. It determines the slope of the line, i.e. the change in the value of Y corresponding to the unit change in X and therefore, it is also called as a **"Slope Coefficient."**

1. The correlation coefficient is the geometric mean of two regression coefficients. Symbolically, it can be expressed as:

$$r = \sqrt{b_{xy} + b_{yx}}$$

2. The value of the coefficient of correlation **cannot exceed unity**

NOTES

- i.e. 1. Therefore, if one of the regression coefficients is greater than unity, the other must be less than unity.
- The sign of both the regression coefficients will be same, i.e. they will be either positive or negative. Thus, it is not possible that one regression coefficient is negative while the other is positive.
 - The coefficient of correlation will have the same sign as that of the regression coefficients, such as if the regression coefficients have a positive sign, then “r” will be positive and vice-versa.
 - The average value of the two regression coefficients will be greater than the value of the correlation. Symbolically, it can be represented as

$$\frac{b_{xy} + b_{yx}}{2} > r$$

- The regression coefficients are independent of the change of origin, but not of the scale. By origin, we mean that there will be no effect on the regression coefficients if any constant is subtracted from the value of X and Y. By scale, we mean that if the value of X and Y is either multiplied or divided by some constant, then the regression coefficients will also change.

Thus, all these properties should be kept in mind while solving for the regression coefficients.

7.8 SUMMARY

- Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data
- Regression analysis refers to assessing the relationship between the outcome variable and one or more variables
- A straight line can be fitted to the given data by the method of least squares
- The constant ‘b’ in the regression equation ($Y_e = a + bX$) is called as the **Regression Coefficient**. It determines the slope of the line, i.e. the change in the value of Y corresponding to the unit change in X and therefore, it is also called as a “**Slope Coefficient.**”

7.9 KEY WORDS

Regression, Linear regression, Types of regression coefficient, Properties of regression coefficient, straight line, Regression equation, straight line Deviations, Actual Mean,

7.10 ANSWER TO CHECK YOUR PROGRESS

- Regression analysis refers to assessing the relationship between the outcome variable and one or more variables

NOTES

2. When regression is Y on X

$$r \frac{\sigma_y}{\sigma_x} = b_{yx} = \frac{\sum(dx)(dy) - \frac{(\sum dx)(\sum dy)}{N}}{\sum(dx)^2 - \frac{(\sum dx)^2}{N}}$$

When regression is X on Y

$$r \frac{\sigma_x}{\sigma_y} = b_{xy} = \frac{\sum(dx)(dy) - \frac{(\sum dx)(\sum dy)}{N}}{\sum(dy)^2 - \frac{(\sum dy)^2}{N}}$$

3. The method of least squares helps us to find the values of unknowns a and b in such a way that the following two conditions are satisfied:
- $\sum(Y - Y^{\wedge}) = 0$ $\sum(Y - Y^{\wedge})^2 = 0$.
 - $\sum(Y - Y^{\wedge})^2 = \sum(Y - Y^{\wedge})^2$.

7.10 QUESTIONS AND EXERCISE

SHORT QUESTION ANSWER

1. What are regression coefficients?
2. Define regression and write down the two regression equations
3. Describe different types of regression
4. What are the uses of regression analysis?

LONG QUESTION AND ANSWER

1. Explain the principle of least squares
2. State the properties of regression equations
3. For 5 observations of pairs of (X, Y) of variables X and Y the following results are obtained. $\sum X=15$, $\sum Y=25$, $\sum X^2=55$, $\sum Y^2=135$, $\sum XY=83$. Find the equation of the lines of regression and estimate the values of X and Y if $Y=8$; $X=12$.
4. Using the following information you are requested to (i) obtain the linear regression of Y on X (ii) Estimate the level of defective parts delivered when inspection expenditure amounts to Rs.28,000 $\sum X=424$, $\sum Y=363$, $\sum X^2 = 21926$, $\sum Y^2 = 15123$, $\sum XY=12815$, $N=10$. Here X is the expenditure on inspection, Y is the defective parts delivered

7.11 FURTHER READINGS

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. SahityaBhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata

- McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
 5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
 5. 6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.

Regression Analysis

NOTES

Self-Instructional Material

NOTES

UNIT 8 - INDEX NUMBER

Structure

- 8.0 Introduction
- 8.1 Objectives
- 8.2 Index Numbers
 - 8.2.1 Types of Index Numbers
 - 8.2.2 Problems in construction of Index Numbers
 - 8.2.3 Methods of Constructing Index Numbers
 - 8.2.4 Quantity or Volume Index Numbers
 - 8.2.5 Test for Index Numbers
 - 8.2.6 Chain Base Index Numbers
- 8.3 Cost of living Index Numbers
 - 8.3.1 Construction of cost of living Index Numbers
 - 8.3.2 Methods to construct cost of living Index Numbers
 - 8.3.3 Uses of cost of living Index Numbers
- 8.4 Uses of Index Numbers
- 8.5 Limitations of Index Numbers
- 8.6 Summary
- 8.7 Key Words
- 8.8 Answers to Check Your Progress
- 8.9 Questions and Exercise
- 8.10 Further Readings

8.0 INTRODUCTION

Index numbers are a commonly used statistical device for measuring the combined fluctuations in group-related variables. If we wish to compare the prices of consumer items today with their prices ten years ago, we are not interested in comparing the prices of only one item, but in comparing average price levels. We may wish to compare the present agricultural production or industrial production with that at the time of independence. Here again, we have to consider all items of production and each item may have undergone a different fractional increase (or even a decrease). How do we obtain a composite measure? This composite measure is provided by index numbers, which may be defined as a device for combining the variations that have occurred to a group of related variables over a period of time, to obtain a figure that

represents the 'net' result of the change in the constitute variables. In this unit you will learn in detail about index numbers.

Index Number

8.1 OBJECTIVES

NOTES

After going through this unit, you will

- Understand about index numbers and their types
- Learn about the different methods of calculating index numbers
- Know the uses and limitations of index numbers

8.2 INDEX NUMBERS

Index numbers are meant to study changes in the effects of factors which cannot be measured directly. According to Bowley, "Index numbers are used to measure the changes in some quantity which we cannot observe directly". For example, changes in business activity in a country are not capable of direct measurement, but it is possible to study relative changes in business activity by studying the variations in the values of some such factors which affect business activity, and which are capable of direct measurement.

Index numbers may be classified in terms of the variables that they are intended to measure. In business, different groups of variables in the measurement of which index number techniques are commonly used are (i) price, (ii) quantity, (iii) value and (iv) business activity. Thus, we have an index of wholesale prices, index of consumer prices, index of industrial output, index of value of exports and index of business activity, etc. Here we shall be mainly interested in index numbers of prices showing changes with respect to time, although the methods described can be applied to other cases. In general, the present level of prices is compared with the level of prices in the past. The present period is called the current period and some period in the past is called the base period.

8.2.1 TYPE OF INDEX NUMBER

Index numbers are names after the activity they measure. Their types are as under:

Price Index: Measure changes in price over a specified period of time. It is basically the ratio of the price of a certain number of commodities at the present year as against base year.

Quantity Index : As the name suggest, these indices pertain to measuring changes in volumes of commodities like goods produced or goods consumed, etc.

Value Index: These pertain to compare changes in the monetary value of imports, exports, production or consumption of commodities.

8.2.2 PROBLEMS IN THE CONSTRUCTION OF INDEX NUMBERS

The decision regarding the following problems/aspect has to be taken before starting the actual construction of any type of index numbers.

Self-Instructional Material

NOTES

- (i) Purpose of Index numbers under construction
- (ii) Selection of base period
- (iii) Selection of items
- (iv) Selection of source data
- (v) Collection of data
- (vi) Selection of average
- (vii) System of weighting

8.2.3 METHODS OF CONSTRUCTING INDEX NUMBERS

The index number for this purpose is divided into two:

- (1) Unweighted Index number
 - Simple aggregative
 - Simple Average of price relatives
- (2) Weighted Index number
 - Weighted aggregative
 - Weighted Average of price relatives

Unweighted Index number:

There are two methods of constructing unweighted index numbers: (1) Simple Aggregative Method (2) Simple Average of Relative Method

Simple Aggregative Method

In this method, the total price of commodities in a given (current) year is divided by the total price of commodities in a base year and expressed as percentage:

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

Simple Average of Relative Method

In this method, we compute price relatives or link relatives of the given commodities and then use one of the averages such as the arithmetic mean, geometric mean, median, etc. If we use the arithmetic mean as the average, then:

$$P_{01} = \frac{1}{n} \sum \frac{P_1}{P_0} \times 100$$

The simple average of relative method is simpler and easier to apply than the simple aggregative method. The only disadvantage is that it gives equal weight to all items.

Example :

The following are the prices of four different commodities for 2017 and 2018. Compute a price index with the (1) simple aggregative method and (2) average of price relative method by using both the arithmetic mean and geometric mean, taking 2017 as the base.

NOTES

| Commodity | Cotton | Wheat | Rice | Grams |
|---------------|--------|-------|------|-------|
| Price in 2017 | 909 | 288 | 767 | 659 |
| Price in 2018 | 874 | 305 | 910 | 573 |

Solution:

| Commodity | Price in 2017 P_0 | Price in 2018 P_1 | Price relative $P = \frac{P_1}{P_0} \times 100$ | log p |
|--------------|------------------------|------------------------|--|--------------------------|
| Cotton | 909 | 874 | 69.15 | 1.9829 |
| Wheat | 288 | 305 | 105.90 | 2.0249 |
| Rice | 767 | 910 | 118.64 | 2.0742 |
| Grams | 659 | 573 | 86.95 | 1.9393 |
| Total | $\Sigma P_0 = 2623$ | $\Sigma P_1 = 2662$ | $\Sigma P = 407.64$ | $\Sigma \log P = 8.0213$ |

1. Simple Aggregative Method

$$P_{01} = \frac{\Sigma P_1}{\Sigma P_0} \times 100 = \frac{2662}{2623} \times 100 = \mathbf{101.49}$$

2. Simple Average of Price Relative Method (using the arithmetic mean)

$$P_{01} = \frac{1}{n} \Sigma \frac{P_1}{P_0} \times 100 = \frac{1}{4} (407.64) \times 100 = \mathbf{101.91}$$

3. Average of price relative method (using the geometric mean)

$$P_{01} = \text{antilog} \left(\frac{\Sigma \log P}{4} \right) = \text{antilog} \left(\frac{8.0213}{4} \right) = \mathbf{101.23}$$

Weighted Index number:

When all commodities are not of equal importance, we assign weight to each commodity relative to its importance and the index number computed from these weights is called a weighted index number.

NOTES

Weighted aggregative index number:

In order to attribute appropriate importance to each of the items used in an aggregate index number some reasonable weights must be used. There are various methods of assigning weights and consequently a large number of formulae for constructing index numbers have been devised of which some of the most important ones are:

1. Laspeyre's Index Number
2. Paasche's Index Number
3. Fisher's Ideal Index Number
4. Marshal-Edge worth Index Number

Laspeyre's Index Number:

In this index number the base year quantities are used as weights, so it also called the base year weighted index.

$$P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100$$

Paasche's Index Number:

In this index number the current (given) year quantities are used as weights, so it is also called the current year weighted index.

$$P_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100$$

Fisher's Ideal Index Number:

The geometric mean of Laspeyre's and Paasche's index numbers is known as Fisher's ideal index number. It is called ideal because it satisfies the time reversal and factor reversal test.

$$P_{01} = \sqrt{\text{Laspeyre's Index} \times \text{Paasche's Index}}$$

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100$$

Marshal-Edgeworth Index Number:

In this index number the average of the base year and current year quantities are used as weights. This index number was proposed by two English economists, Marshal and Edgeworth.

$$P_{01} = \left(\frac{\sum P_1 q_0 + \sum P_1 q_1}{\sum P_0 q_0 + \sum P_0 q_1} \right) \times 100$$

$$P_{01} = \frac{\sum P_1 (q_0 + q_1)}{\sum P_0 (q_0 + q_1)} \times 100$$

Example:

Compute the weighted aggregative price index numbers for 2011 with 2010 as the base year using (1) Laspeyre's Index Number (2) Paasche's Index Number (3) Fisher's Ideal Index Number (4) Marshal-Edgeworth Index Number.

NOTES

| Commodity | Prices | | Quantities | |
|-----------|--------|------|------------|------|
| | 2010 | 2011 | 2010 | 2011 |
| A | 10 | 12 | 20 | 22 |
| B | 8 | 8 | 16 | 18 |
| C | 5 | 6 | 10 | 11 |
| D | 4 | 4 | 7 | 8 |

Solution:

| Commodity | Prices | | Quantities | | P_1q_0 | P_0q_0 | P_1q_1 | P_0q_1 |
|-----------|---------------|---------------|---------------|---------------|--------------------------|--------------------------|--------------------------|--------------------------|
| | 2010 P_0 | 2011 P_1 | 2010 q_0 | 2011 q_1 | | | | |
| A | 10 | 12 | 20 | 22 | 240 | 200 | 264 | 220 |
| B | 8 | 8 | 16 | 18 | 128 | 128 | 144 | 144 |
| C | 5 | 6 | 10 | 11 | 60 | 50 | 66 | 55 |
| D | 4 | 4 | 7 | 8 | 28 | 28 | 32 | 32 |
| | | | | | ΣP_1q_0 = 456 | ΣP_0q_0 = 406 | ΣP_1q_1 = 506 | ΣP_0q_1 = 451 |

Laspeyre's Index Number:

$$P_{01} = \frac{\Sigma P_1q_0}{\Sigma P_0q_0} \times 100 = \frac{456}{406} \times 100 = 112.32$$

Paasche's Index Number:

$$P_{01} = \frac{\Sigma P_1q_1}{\Sigma P_0q_1} \times 100 = \frac{506}{451} \times 100 = 112.20$$

Fisher's Ideal Index Number

$$P_{01} = \sqrt{\text{Laspeyre's Index} \times \text{Paasche's Index}}$$

NOTES

$$P_{01} = \sqrt{112.32 \times 112.20} = 112.26$$

Marshal-Edgeworth Index Number

$$P_{01} = \frac{\sum P_1(q_0 + q_1)}{\sum P_0(q_0 + q_1)} \times 100 = \left(\frac{456+506}{406+451}\right) \times 100 = 112.38$$

Weighted average of price relatives:

When the specific weights are given for each commodity the weighted index number is calculated by

$$\text{Weighted Average of Price Relative index} = \frac{\sum pw}{\sum w}$$

Where w = the weight of the commodity

$$p = \text{the price relative index} = \frac{P_1}{P_0} \times 100$$

When the base year value is P_0q_0 is taken as the weight i.e. $w = P_0q_0$ then the formula is

$$\text{Weighted Average of Price Relative index} = \frac{\sum \left(\frac{P_1}{P_0} \times 100\right) \times P_0q_0}{\sum P_0q_0} = \frac{\sum P_1q_0}{\sum P_0q_0} \times 100$$

This is nothing but Laspeyre’s formula

When the weight taken as $w = P_0q_1$ then the formula is

$$\text{Weighted Average of Price Relative index} = \frac{\sum \left(\frac{P_1}{P_0} \times 100\right) \times P_0q_1}{\sum P_0q_1} = \frac{\sum P_1q_1}{\sum P_0q_1} \times 100$$

This is nothing but Paasche’s formula

Example: Compute the weighted index number for the following data

| Commodity | Price | | Weight |
|-----------|--------------|-----------|--------|
| | Current year | Base year | |
| X | 5 | 4 | 40 |
| Y | 3 | 2 | 60 |
| Z | 2 | 1 | 20 |

Solution:

| Commodity | Price | | Weight | P $= \frac{P_1}{P_0} \times 100$ | PW |
|-----------|--------------|-----------|--------|-------------------------------------|----|
| | Current year | Base year | | | |
| X | 5 | 4 | 40 | | |
| Y | 3 | 2 | 60 | | |
| Z | 2 | 1 | 20 | | |

NOTES

| | | | | | |
|---|---|---|-----|-----|-------|
| X | 5 | 4 | 40 | 125 | 5000 |
| Y | 3 | 2 | 60 | 150 | 9000 |
| Z | 2 | 1 | 20 | 200 | 4000 |
| | | | 120 | | 18000 |

$$\text{Weighted Average of Price Relative index} = \frac{\sum pw}{\sum w} = \frac{18000}{120} = \mathbf{150}$$

8.2.4 QUANTITY OR VOLUME INDEX NUMBER

Price index numbers measures and permit comparison of the price of certain goods; quantity index number, on the other hand, measures the physical volume of production, construction of employment. Though price indices are more widely used, production indices are highly significant as indicators of the level of output in the economy or in parts of it.

In constructing quantity index numbers, the problems confronting the statistician are analogous to those involved in price indices. We measure changes in quantities, and when we weigh we use prices or values as weights. Quantity indices can be obtained easily by changing p to q and q to p in the various formulae discussed above.

Thus, when Laspeyres method is used

$$Q_{01} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$$

When Paasche's formula is used

$$Q_{01} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$$

When Fisher's formula is used

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$$

These formulae represent the quantity index in which the quantities of the different commodities are weighted by their prices.

Example:

Compute the following quantity indices from the data given below (1) Laspeyre's Index Number (2) Paashe's Index Number (3) Fisher's Ideal Index Number.

| | | |
|--|------|------|
| | 2002 | 2012 |
|--|------|------|

NOTES

| Commodity | Price | Total value | Price | Total value |
|-----------|-------|-------------|-------|-------------|
| A | 10 | 200 | 12 | 360 |
| B | 12 | 480 | 15 | 900 |
| C | 15 | 450 | 17 | 680 |

Solution :

Here instead of quantity, total values are given, hence find quantities of base year and current year

$$\text{Quantity} = \frac{\text{totalvalue}}{\text{price}}$$

| Commodity | p ₀ | q ₀ | p ₁ | q ₁ | P ₀ Q ₀ | P ₀ Q ₁ | P ₁ Q ₀ | P ₁ Q ₁ |
|-----------|----------------|----------------|----------------|----------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| A | 10 | 20 | 12 | 30 | 200 | 300 | 240 | 360 |
| B | 10 | 40 | 15 | 60 | 400 | 600 | 600 | 900 |
| C | 15 | 30 | 17 | 40 | 450 | 600 | 510 | 680 |
| Total | | | | | 1050 | 1500 | 1350 | 1940 |

$$\text{Laspeyre's method } Q_{01} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100 = \frac{1500}{1050} \times 100 = \mathbf{142.86}$$

$$\text{Paasche's formula } Q_{01} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100 = \frac{1940}{1350} \times 100 = \mathbf{143.7}$$

$$\begin{aligned} \text{Fisher's formula } Q_{01} &= \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100 \\ &= \sqrt{LxP} = \sqrt{142.86 \times 143.7} \\ &= \mathbf{143.27} \end{aligned}$$

8.2.5 TEST FOR INDEX NUMBER

There are certain tests which are put to verify the consistency, or adequacy of an index number formula from different points of view. The most popular among these are the following tests:

- Order reversal test.
- Time reversal test.
- Factor reversal test.
- Unit test.

At the outset, it should be noted that it is neither possible nor necessary for an index-number formula to satisfy all the tests mentioned above. But, an ideal formula should be such that it satisfies the maximum possible tests which are relevant to the matter under study.

1. Order reversal test:

This test requires that a formula of Index number should be such that the value of the index number remains the same, even if, the order of arrangement of the items is reversed, or altered. As a matter of fact, this test is satisfied by all the formulas of index number explained in this chapter.

2. Time reversal test:

The time reversal test requires that the index number computed backwards should be the reciprocal of the index number computed forward, except for the constant of proportionality

$$P_{01} = \sqrt{\frac{\sum P_1q_0}{\sum P_0q_0} \times \frac{\sum P_1q_1}{\sum P_0q_1}}$$

$$P_{10} = \sqrt{\frac{\sum P_0q_1}{\sum P_1q_1} \times \frac{\sum P_0q_0}{\sum P_1q_0}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{\sum P_1q_0}{\sum P_0q_0} \times \frac{\sum P_1q_1}{\sum P_0q_1} \times \frac{\sum P_0q_1}{\sum P_1q_1} \times \frac{\sum P_0q_0}{\sum P_1q_0}}$$

$$P_{01} \times P_{10} = 1$$

Laspeyre's and Paasche's method do not satisfy this test but Fisher's ideal index satisfies this method. Besides both the simple and weighted geometric mean of price relatives, also, satisfy this time reversal test.

3. Factor reversal test:

This test has also been put forth by Prof. Irving Fisher, in this test the product of price index and quantity index must be equal to the value index. Thus, for the Factor Reversal test, a formula of index number should satisfy the following equation:

$$\text{Price index} \times \text{Quantity Index} = \text{Value Index}$$

$$P_{01} = \sqrt{\frac{\sum P_1q_0}{\sum P_0q_0} \times \frac{\sum P_1q_1}{\sum P_0q_1}}$$

$$Q_{01} = \sqrt{\frac{\sum q_1p_0}{\sum q_0p_0} \times \frac{\sum q_1p_1}{\sum q_0p_1}}$$

$$\therefore P_{01} \times Q_{01} = \frac{\sum p_1q_1}{\sum p_0q_0}$$

Most of the formulae of index number discussed above fail to satisfy this acid test of consistency except that of Prof. Irving Fisher.

4. Unit test

This test suggests that the formula for constructing an index should be independent of the unit of measurement in which the prices

NOTES

and quantities are quoted. Except unweighted aggregative index number all other formulas in this chapter satisfy this test.

Example:

Construct Fisher’s ideal index for the following data. Test whether it satisfies time reversal test and factor reversal test.

| Commodity | Base year | | Current year | |
|-----------|-----------|-------|--------------|-------|
| | Quantity | Price | Quantity | Price |
| A | 24 | 20 | 30 | 24 |
| B | 30 | 14 | 40 | 10 |
| C | 10 | 10 | 16 | 18 |

Solution:

| Commodity | q ₀ | p ₀ | q ₁ | p ₁ | p ₀ q ₀ | p ₀ q ₁ | p ₁ q ₀ | p ₁ q ₁ |
|-----------|----------------|----------------|----------------|----------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| A | 24 | 20 | 30 | 24 | 480 | 600 | 576 | 720 |
| B | 30 | 14 | 40 | 10 | 420 | 560 | 300 | 400 |
| C | 10 | 10 | 16 | 18 | 100 | 160 | 180 | 288 |
| | | | | | 1000 | 1320 | 1056 | 1408 |

$$\begin{aligned}
 \text{Fisher ideal index number } P_{01} &= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100 = \sqrt{\frac{1056}{1000} \times \frac{1408}{1320}} \times 100 \\
 &= \sqrt{1.056 \times 1.067} \times 100 = \sqrt{1.127} \times 100 \\
 &= 1.062 \times 100 = \mathbf{106.2}
 \end{aligned}$$

Time Reversal test:

Time Reversal test is satisfied when $P_{01} \times P_{10} = 1$

$$P_{01} = \sqrt{\frac{P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} = \sqrt{\frac{1056}{1000} \times \frac{1408}{1320}}$$

$$P_{10} = \sqrt{\frac{\sum P_{0q1} \times \sum P_{0q0}}{\sum P_{1q1} \times \sum P_{1q0}}} = \sqrt{\frac{1320}{1408} \times \frac{1000}{1056}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{1056}{1000} \times \frac{1408}{1320} \times \frac{1320}{1408} \times \frac{1000}{1056}} = \sqrt{1} = 1$$

Hence Fisher ideal index satisfy the time reversal test

Factor Reversal Test:

$$P_{01} = \sqrt{\frac{\sum P_{1q0} \times \sum P_{1q1}}{\sum P_{0q0} \times \sum P_{0q1}}} = \sqrt{\frac{1056}{1000} \times \frac{1408}{1320}}$$

$$Q_{01} = \sqrt{\frac{\sum q_{1p0} \times \sum q_{1p1}}{\sum q_{0p0} \times \sum q_{0p1}}} = \sqrt{\frac{1320}{1000} \times \frac{1408}{1056}}$$

$$\begin{aligned} \therefore P_{01} \times Q_{01} &= \sqrt{\frac{1056}{1000} \times \frac{1408}{1320} \times \frac{1320}{1000} \times \frac{1408}{1056}} = \sqrt{\left(\frac{1408}{1000}\right)^2} \\ &= \frac{1408}{1000} = \frac{\sum p_1 q_1}{\sum p_0 q_0} \end{aligned}$$

Hence Fisher ideal index number satisfy the factor reversal test

8.2.6 CHAIN BASE INDEX NUMBER

In this method, there is no fixed base period; the year immediately preceding the one for which the price index has to be calculated is assumed as the base year. Thus, for the year 1994 the base year would be 1993, for 1993 it would be 1992, for 1992 it would be 1991, and so on. In this way there is no fixed base and it keeps on changing.

The chief advantage of this method is that the price relatives of a year can be compared with the price levels of the immediately preceding year. Businesses mostly interested in comparing this time period rather than comparing rates related to the distant past will utilize this method.

$$\text{Link relative of current years} = \frac{\text{Price in the Current Year}}{\text{Price in the preceding Year}} \times 100$$

$$P_{n-1, n} = \frac{P_n}{P_{n-1}} \times 100$$

Example:

NOTES

Find the index numbers for the following data taking 2010 as the base year

| | | | | | | |
|-------|------|------|------|------|------|------|
| Years | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
| Price | 18 | 21 | 25 | 23 | 28 | 30 |

Solution:

| Year | Price | Link $\frac{P_n}{P_{n-1}} \times 100$ | Relatives Chain indices |
|------|-------|--|---|
| 2004 | 18 | $\frac{18}{18} \times 100 = 100$ | 100 |
| 2005 | 21 | $\frac{21}{18} \times 100 = 116.67$ | $\frac{100 \times 116.67}{100} = 116.7$ |
| 2006 | 25 | $\frac{25}{21} \times 100 = 119.05$ | $\frac{116.67 \times 119.05}{100} = 138.9$ |
| 2007 | 23 | $\frac{23}{25} \times 100 = 92$ | $\frac{138.9 \times 92}{100} = 127.79$ |
| 2008 | 28 | $\frac{28}{23} \times 100 = 121.74$ | $\frac{127.79 \times 121.74}{100} = 155.57$ |
| 2009 | 30 | $\frac{30}{28} \times 100 = 107.14$ | $\frac{155.57 \times 107.14}{100} = 166.68$ |

CHECK YOUR PROGRESS – 1

1. What is chain base index number
2. What is the formula for Fisher’s Ideal Index Number?
3. What is weighted index number?

8.3 COST OF LIVING INDEX NUMBER

Cost of living index numbers measure the changes in the prices paid by consumers for a special “basket” of goods and services during the current year as compared to the base year. The basket of goods and

services will contain items like (1) Food (2) Rent (3) Clothing (4) Fuel and Lighting (5) Education (6) Miscellaneous like cleaning, transport, newspapers, etc. Cost of living index numbers are also called consumer price index numbers or retail price index numbers.

8.3.1 CONSTRUCTION OF COST OF LIVING INDEX NUMBERS

The following steps are involved in the construction of Cost of living index numbers.

(1) Class of People:

The first step in the construction of the Cost of living index (CLI) is that the class of people should be defined clearly. It should be decided whether the cost of living index number is being prepared for industrial workers, or middle or lower class salaried people living in a particular area. It is therefore necessary to specify the class of people and locality where they reside.

(2) Family Budget Inquiry:

The next step in the construction of a Cost of living index number is that some families should be selected randomly. These families provide information about the cost of food, clothing, rent, miscellaneous, etc. The inquiry includes questions on family size, income, the quality and quantity of resources consumed and the money spent on them, and the weights are assigned in proportions to the expenditure on different items.

(3) Price Data:

The next step is to collect data on the retail prices of the selected commodities for the current period and the base period when these prices should be obtained from the shops situated in the locality for which the index numbers are prepared.

(4) Selection of Commodities:

The next step is the selection of the commodities to be included. We should select those commodities which are most often used by that class of people.

8.3.2 METHODS TO COMPUTE COST OF LIVING INDEX NUMBERS

There are two methods to compute cost of living index numbers: (1) Aggregate Expenditure Method (2) Family Budget Method

Aggregate Expenditure Method

In this method, the quantities of commodities consumed by the particular group in the base year are estimated and these figures or their proportions are used as weights. Then the total expenditure of each commodity for each year is calculated. The price of the current year is multiplied by the quantity or weight of the base year. These products are

Index Number

NOTES

Self-Instructional Material

NOTES

added. Similarly, for the base year the total expenditure of each commodity is calculated by multiplying the quantity consumed by its price in the base year. These products are also added. The total expenditure of the current year is divided by the total expenditure of the base year and the resulting figure is multiplied by 100 to get the required index numbers. In this method, the current period quantities are not used as weights because these quantities change from year to year.

$$P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100$$

Here,

P_1 - Represent the price of the current year,

P_0 - Represents the price of the base year and

q_0 - Represents the quantities consumed in the base year.

Family Budget Method:

In this method, the family budgets of a large number of people are carefully studied and the aggregate expenditure of the average family for various items is estimated. These values are used as weights. The current year's prices are converted into price relatives on the basis of the base year's prices, and these price relatives are multiplied by the respective values of the commodities in the base year. The total of these products is divided by the sum of the weights and the resulting figure is the required index numbers.

$$P_{01} = \frac{\sum WI}{\sum W} \times 100$$

Here, $I = \frac{\sum P_1}{\sum P_0} \times 100$ and $\sum W = P_0 q_0$

Example:

Construct the cost of living index number for 2018 on the basis of 2017 from the following data using (1) Aggregate Expenditure Method (2) Family Budget Method.

| Commodity | Quantity Consumed in 2017 (in quintal) | Prices | |
|-----------|---|--------|--------|
| | | 2017 | 2018 |
| A | 6 | 315.75 | 316.00 |
| B | 6 | 305.00 | 308.00 |

| | | | |
|---|---|---------|---------|
| C | 1 | 416.00 | 419.00 |
| D | 6 | 528.00 | 610.00 |
| E | 4 | 120.00 | 119.50 |
| F | 1 | 1020.00 | 1015.00 |

Index Number

NOTES

Solution:

The cost of living index number of 2018 by Aggregate Expenditure method:

| Commodity | Quantity Consumed in 2017 (in quintal) q_0 | Prices | | P_1q_0 | P_0q_0 |
|-----------|--|---------------|---------------|------------------------|--------------------------|
| | | 2017 P_0 | 2018 P_1 | | |
| A | 6 | 315.75 | 316.00 | 1896 | 1894.50 |
| B | 6 | 305.00 | 308.00 | 1848 | 1830.00 |
| C | 1 | 416.00 | 419.00 | 419 | 416.00 |
| D | 6 | 528.00 | 610.00 | 3660 | 3168.00 |
| E | 4 | 120.00 | 119.50 | 478 | 480.00 |
| F | 1 | 1020.00 | 1015.00 | 1015 | 1020.00 |
| | | | | $\Sigma P_1q_0 = 9316$ | $\Sigma P_0q_0 = 8808.5$ |

The cost of living index number of 2018 is

$$P_{01} = \frac{\Sigma P_1q_0}{\Sigma P_0q_0} \times 100 = \frac{9316}{8808.5} \times 100 = 105.76$$

The cost of living index number of 2018 by Family Budget Method:

| Commodity | Quantity Consumed in 2017 (in quintal) q_0 | Prices | | $W = P_0q_0$ | $I = \frac{\Sigma P_1}{\Sigma P_0} \times 100$ | Product WI |
|-----------|--|---------------|---------------|--------------|--|------------|
| | | 2017 P_0 | 2018 P_1 | | | |

Self-Instructional Material

NOTES

| | | | | | | |
|---|---|-------------|-------------|------------------------|--------|-------------------------------|
| A | 6 | 315.75 | 316.00 | 1894.5 | 100.08 | 189601.56 |
| B | 6 | 305.00 | 308.00 | 1830.0 | 100.98 | 184793.40 |
| C | 1 | 416.00 | 419.00 | 416.0 | 100.72 | 41899.52 |
| D | 6 | 528.00 | 610.00 | 3168.0 | 115.53 | 365999.04 |
| E | 4 | 120.00 | 119.50 | 480.0 | 99.58 | 47798.4 |
| F | 1 | 1020.0 0 | 1015.0 0 | 1020.0 | 99.51 | 101500.20 |
| | | | | $\Sigma W =$ 8808.5 | | ΣWI =931592.1 2 |

The cost of living index number of 2018 is

$$P_{01} = \frac{\Sigma WI}{\Sigma W} \times 100 = \frac{931592.12}{8808.5} = \mathbf{105.76}$$

8.3.3 USES OF COST OF LIVING INDEX NUMBER

- They indicate the changes in the consumer prices. Thus they help government in formulating policies regarding control of price, taxation, imports and exports of commodities, etc.
- They are used in granting allowances and other facilities to employees
- They are used for the evaluation of purchasing power of money. They are used for deflating money
- They are used for comparing changes in the cost of living of different classes of people

8.4 USES OF INDEX NUMBER

The main uses of index numbers are given below.

- Index numbers are used in the fields of commerce, meteorology, labour, industry, etc.
- Index numbers measure fluctuations during intervals of time, group differences of geographical position of degree, etc.
- They are used to compare the total variations in the prices of different commodities in which the unit of measurements differs with time and price, etc.
- They measure the purchasing power of money.
- They are helpful in forecasting future economic trends.
- They are used in studying the difference between the comparable categories of animals, people or items.
- Index numbers of industrial production are used to measure the changes in the level of industrial production in the country.

- Index numbers of import prices and export prices are used to measure the changes in the trade of a country.

Index numbers are used to measure seasonal variations and cyclical variations in a time series.

8.5 LIMITATIONS OF INDEX NUMBER

- They are simply rough indications of the relative changes.
- The choice of representative commodities may lead to fallacious conclusions as they are based on samples.
- There may be errors in the choice of base periods or weights, etc.
- Comparisons of changes in variables over long periods are not reliable.
- They may be useful for one purpose but not for another.
- They are specialized types of averages and hence are subject to all those limitations which an average suffers from.

CHECK YOUR PROGRESS - 2

4. What are the methods to compute Cost of Living Index numbers?
5. What are the popular Tests for Index number?
6. Write a few uses of index number.

8.6 SUMMARY

- Index numbers are meant to study changes in the effects of factors which cannot be measured directly. According to Bowley, “Index numbers are used to measure the changes in some quantity which we cannot observe directly”.
- . Price Index Quantity Index Value Index. Quantity Index Numbers are the types of index numbers.
- Price index numbers measures and permit comparison of the price of certain goods; quantity index number, on the other hand, measures the physical volume of production, construction of employment. Though price indices are more widely used, production indices are highly significant as indicators of the level of output in the economy or in parts of it.
- There are certain tests which are put to verify the consistency, or adequacy of an index number formula from different points of view. The most popular among these are the following tests: (1) Order reversal test. (2) Time reversal test. (3) Factor reversal test. (4) Unit test.
- In this method, there is no fixed base period; the year immediately preceding the one for which the price index has to be calculated is assumed as the base year.

NOTES

- Cost of living index numbers measure the changes in the prices paid by consumers for a special “basket” of goods and services during the current year as compared to the base year.
- There are two methods to compute cost of living index numbers: (1) Aggregate Expenditure Method (2) Family Budget Method.

8.7 KEY WORDS

Index numbers, Price Index, Quantity Index, Value Index, Laspeyre’s Index Number, Paasche’s Index Number, Fisher’s Ideal Index Number, Marshal-Edge worth Index Number, Order reversal test, Time reversal test, Factor reversal test, Unit test, Chain Base index number, Cost of living index number.

8.8 ANSWERS TO CHECK YOUR PROGRESS

1. In this method, there is no fixed base period; the year immediately preceding the one for which the price index has to be calculated is assumed as the base year.

$$2. P_{01} = \sqrt{\text{Laspeyre's Index} \times \text{Paasche's Index}}$$

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100$$

3. When all commodities are not of equal importance, we assign weight to each commodity relative to its importance and the index number computed from these weights is called a weighted index number.

4. There are two methods to compute cost of living index numbers: (1) Aggregate Expenditure Method (2) Family Budget Method

5. Order reversal test, Time reversal test, Factor reversal test, Unit test

6. Index numbers are used in the fields of commerce, meteorology, labour, industry, etc. Index numbers measure fluctuations during intervals of time, group differences of geographical position of degree, etc. They are used to compare the total variations in the prices of different commodities in which the unit of measurements differs with time and price, etc. They measure the purchasing power of money. They are helpful in forecasting future economic trends

8.9 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTIONS

1. Define index number and write the uses of index numbers
2. State the types of index numbers
3. State the methods of constructing consumer price index

LONG ANSWER QUESTIONS

1. Compute (1) Laspeyre's (2) Paasche's index number for the 2001 from the following

| Commodity | Price | | Quantity | |
|-----------|-------|------|----------|------|
| | 2002 | 2010 | 2002 | 2010 |
| W | 4 | 6 | 8 | 7 |
| X | 3 | 5 | 10 | 8 |
| Y | 2 | 4 | 14 | 12 |
| Z | 5 | 7 | 19 | 11 |

2. Calculate Fisher's ideal index method for the following data

| Commodity | 2011 | | 2012 | |
|-----------|-------|----------|-------|----------|
| | Price | Quantity | Price | Quantity |
| A | 2 | 7 | 3 | 5 |
| B | 5 | 11 | 6 | 10 |
| C | 3 | 14 | 5 | 11 |
| D | 4 | 16 | 4 | 18 |

3. Construct the consumer price index number of 2015 on the from the following data using

- (i) the average expenditure method and
(ii) the family budget method

| Commodity | Quantity Consumed in 2014 | Prices | |
|-----------|---------------------------|--------|------|
| | | 2014 | 2015 |
| A | 6 Kg | 5 | 7 |
| B | 6 Quintal | 6 | 6 |
| C | 5 Quintal | 5 | 4 |
| D | 6 Quintal | 7 | 7 |
| E | 4 Quintal | 8 | 8 |
| F | 5 Kg | 9 | 9 |

Index Number

NOTES

Self-Instructional Material

Index Number

NOTES

8.10 FURTHER READINGS

- (1) Statistics (Theory & Practice) by Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
- (2) Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
- (3) Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
- (4) Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
- (5) Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
- (6) Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.

UNIT 9 - ANALYSIS OF TIME SERIES

Analysis of Time Series

Structure

- 9.0 Introduction
- 9.1 Objectives
- 9.2 Time series Analysis
 - 9.2.1 Components of time series
 - 9.2.2 Analysis of Time Series
- 9.3 Measurement of trends
 - 9.3.1 Moving average method
 - 9.3.2 Least square method
- 9.4 Measurement of seasonal variation
 - 9.4.1 Methods of constructing seasonal indices
- 9.5 Forecasting
- 9.6 Deseasonalisation
- 9.7 Summary
- 9.8 Key Words
- 9.9 Answers to Check Your Progress
- 9.10 Questions and Exercise
- 9.11 Further Readings

NOTES

9.0 INTRODUCTION

When quantitative data are arranged in the order of their occurrence, the resulting statistical series is called a time series. The quantitative values are usually recorded over equal time interval daily, weekly, monthly, quarterly, half yearly, yearly, or any other time measure. Monthly statistics of Industrial Production in India, Annual birth-rate figures for the entire world, yield on ordinary shares, weekly wholesale price of rice, and daily records of tea sales or census data are some of the examples of time series. Each has a common characteristic of recording magnitudes that vary with passage of time. In this unit we will see about time series analysis.

9.1 OBJECTIVES

After going through this unit, you will

- Learn about time series analysis
- Know about the measurement of trends
- Understand forecasting and Deseasonalisation

Self-Instructional Material

NOTES

9.2 TIME SERIES ANALYSIS

Time series are influenced by a variety of forces. Some are continuously effective other make themselves felt at recurring time intervals, and still others are non-recurring or random in nature. Therefore, the first task is to break down the data and study each of these influences in isolation. This is known as decomposition of the time series. It enables us to understand fully the nature of the forces at work. We can then analysis their combined interactions. Such a study is known as time-series analysis.

9.2.1 COMPONENTS OF TIME SERIES:

The factors that are responsible for bringing about changes in a time series, also called the components of time series, are as follows:

- Secular Trends (or General Trends)
- Seasonal Movements
- Cyclical Movements
- Irregular Fluctuations

Secular Trends:

Secular trend is the main component of a time series which results from long term effects of socio-economic and political factors. It shows the growth or decline in a time series over a long period. It is the type of tendency which continues to persist for a very long period. Prices and export and import data, for example, reflect obviously increasing tendencies over time.

Seasonal Trends:

Seasonal trends are short term movements occurring in data due to seasonal factors. The short term is generally considered as a period in which changes occur in a time series with variations in weather or festivities. For example, it is commonly observed that the consumption of ice-cream during summer is generally high and hence an ice-cream dealer's sales would be higher in some months of the year while relatively lower during winter months. Employment, output, exports, etc., are subject to change due to variations in weather. Similarly, the sale of garments, umbrellas, greeting cards and fire-works are subject to large variations during festivals like Valentine's Day, Eid, Christmas, New Year's, etc. These types of variations in a time series are isolated only when the series is provided biannually, quarterly or monthly.

Cyclic Movements

It is a long term oscillations occurring in a time series. These oscillations are mostly observed in economics data and the periods of such oscillations are generally extended from five to twelve years or more. These oscillations are associated with the well known business

cycles. These cyclic movements can be studied provided a long series of measurements, free from irregular fluctuations, is available.

Irregular Fluctuations

It happens when a sudden changes occurring in a time series which are unlikely to be repeated. They are components of a time series which cannot be explained by trends, seasonal or cyclic movements. These variations are sometimes called residual or random components. These variations, though accidental in nature, can cause a continual change in the trends, seasonal and cyclical oscillations during the forthcoming period. Floods, fires, earthquakes, revolutions, epidemic, strikes etc., are the root causes of such irregularities.

9.2.2 ANALYSIS OF TIME SERIES

The objective of the time series analysis is to identify the magnitude and direction of trends, to estimate the effect of seasonal and cyclical variations and to estimate the size of the residual component. This implies the decomposition of a time series into its several components. Two lines of approach are usually adopted in analyzing a given time series:

- The additive model
- The multiplicative model

The additive model:

It is used when the four components of a time series are independent of one another. Independent means the magnitude and patterns of movement of the components do not affect each other. Using this assumption the magnitudes of the time series are regarded as the sum of separate influences of its four components. In additive approach, the unit of measurements remains the same for all the four components. The additive model can be written as

$$Y = T + S + C + R$$

Where Y = magnitude of a time series

T = Trend,

C = Cyclical component,

S = Seasonal component,

R = Random component.

The multiplicative model:

It is used where the forces giving rise to the four types of variations are interdependent. The magnitude of time series is the product of four components. Then the multiplicative model can be written as

$$Y = T \times S \times C \times R$$

NOTES

NOTES

The additive model is usually used when the time series is spread over a short time span or where the rate of growth or decline in the trend is small. The multiplicative model, which is used more often than the additive model, is generally used whenever the time span of the series is large or the rate of growth or decline is

$$Y - T = S + C + R \quad \text{or} \quad \frac{Y}{T} = S \times C \times R$$

Similarly, a de-trended, de-seasonalized series may be obtained as

$$Y - T - S = C + R \quad \text{or} \quad \frac{Y}{T \times S} = C \times R$$

It is not always necessary for the time series to include all four types of variations; rather, one or more of these components might be missing altogether. For example, when using annual data the seasonal component may be ignored, while in a time series of a short span having monthly or quarterly observations, the cyclical component may be ignored

9.3 MEASUREMENT OF TRENDS

- Moving average method
- Least square method

9.3.1 MOVING AVERAGE METHOD

Moving average method is a simple device of reducing fluctuations and obtaining trend values with a fair degree of accuracy. In this method the average value of a number of years (months, weeks, or days) is taken as the trend value for the middle point of the period of moving average. The process of averaging smoothes the curve and reduces the fluctuations.

The first thing to be decided in this method is the period of the moving average. What it means is to take a decision about the number of consecutive items whose average would be calculated each time. Suppose it has been decided that the period of the moving average would be 5 years (months, weeks, or days) then the arithmetic average of the first 5 items (number 1,2,3,4 and 5) would be placed against item no:3 and then the arithmetic average of item Nos:2,3,4,5 and 6 would be placed against item No: 4. This process would be repeated till the arithmetic average of the last five items has been calculated.

Odd Period of Moving Average

Calculation of three yearly moving averages includes the following steps

1. Add up the values of the first 3 years and place the yearly sum against the median year. (This sum is called moving total)
2. Leave the first year value, add up the values of the next three years and place it against its median year.
3. This process must be continued till all the values of the data are taken for calculation.

NOTES

4. Each 3-yearly moving total must be divided by 3 to get the 3-year moving averages, which is our required trend value.

The formula calculating 3 yearly moving averages is as follows

$$\frac{a+b+c}{3}, \frac{b+c+d}{3}, \frac{c+d+e}{3}$$

The formula calculating 5 yearly moving averages is as follows

$$\frac{a+b+c+d+e}{5}, \frac{b+c+d+e+f}{5}, \frac{c+d+e+f+g}{5} \dots\dots$$

Example:

Calculate the 3 yearly and 5 yearly moving averages of the data

| | | | | | | | | | | | | |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Years | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| sales | 5.2 | 4.9 | 5.5 | 4.9 | 5.2 | 5.7 | 5.4 | 5.8 | 5.9 | 6.0 | 5.2 | 4.8 |

Solution:

| Year | Sales | 3 Year Moving Total | 3 Year Moving Average (3) / 3 | 5 Year Moving Total | 5 Year Moving Average (4) / 5 |
|------|-------|---------------------|----------------------------------|---------------------|----------------------------------|
| 1 | 5.2 | --- | | -- | -- |
| 2 | 4.9 | 15.6 | 5.2 | -- | -- |
| 3 | 5.5 | 15.3 | 5.1 | 25.7 | 5.14 |
| 4 | 4.9 | 15.6 | 5.2 | 26.2 | 5.24 |
| 5 | 5.2 | 15.8 | 5.27 | 26.7 | 5.34 |
| 6 | 5.7 | 16.3 | 5.41 | 27.0 | 5.4 |
| 7 | 5.4 | 16.9 | 5.63 | 28.0 | 5.6 |
| 8 | 5.8 | 17.1 | 5.7 | 28.8 | 5.76 |

NOTES

| | | | | | |
|----|-----|------|------|------|------|
| 9 | 5.9 | 17.7 | 5.23 | 28.3 | 5.66 |
| 10 | 6.0 | 17.1 | 5.7 | 27.7 | 5.54 |
| 11 | 5.2 | 16.0 | 5.33 | --- | --- |
| 12 | 4.8 | --- | --- | --- | --- |

Even Period of Moving Average:

The period of moving average is 4,6, or 8, it is even number. The four yearly total cannot be placed against any year as median 2.5 is between the second and the third year. So the total should be placed in between the 2nd and 3rd years. We must centre the moving average in order to place the moving average against the year

Steps to find even period of moving average:

1. Add up the values of the first 4 years and place the sum against the middle of 2nd and 3rd year. (This sum is called 4 year moving total)
2. Leave the first year value and add next 4 values from the 2nd year onward and write the sum against its middle position.
3. This process must be continued till the value of the last item is taken into account.
4. Add the first two 4-years moving total and write the sum against 3rd year.
5. Leave the first 4-year moving total and add the next two 4-year moving total and place it against 4th year.
6. This process must be continued till all the 4-yearly moving totals are summed up and centered.
7. Divide the 4-years moving total by 8 to get the moving averages which are our required trend values

Example:

Find the 4 yearly moving average for determining trend values in the following time series data

| | | | | | | | |
|------------------|------|------|------|------|------|------|------|
| Year | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 |
| Profit in(000) ₹ | 12 | 14 | 16 | 15 | 13 | 14 | 18 |

Solution:

| Years | Profit | Sum of Fours | 4 years Moving Average | 4 yearly Moving Average Centered |
|-------|--------|--------------|------------------------|----------------------------------|
| 2005 | 12 | | | |
| 2006 | 14 | | | |
| | | 57 | 14.25 | $(14.25 + 14.50) / 2 = 14.38$ |
| 2007 | 16 | | | |
| | | 58 | 14.50 | $(14.50 + 14.50) / 2 = 14.50$ |
| 2008 | 15 | | | |
| | | 58 | 14.50 | $(14.50 + 15.00) / 2 = 14.75$ |
| 2009 | 13 | | | |
| | | 60 | 15.00 | |
| 2010 | 14 | | | |
| | | | | |
| 2011 | 18 | | | |

Advantages

Moving averages can be used for measuring the trend of any series. This method is applicable to linear as well as non-linear trends.

Disadvantages

The trend obtained by moving averages generally is neither a straight line nor a standard curve. For this reason the trend cannot be extended for forecasting future values. Trend values are not available for

NOTES

NOTES

some periods at the start and some values at the end of the time series. This method is not applicable to short time series

9.3.2 LEAST SQUARES METHOD

When the trend is linear the trend equation may be represented by $y = a + bt$ and the values of a and b for the line $y = a + bt$ which minimizes the sum of squares of the vertical deviations of the actual (observed) values from the straight line, are the solutions to the so called normal equations:

$$\Sigma y = na + b\Sigma t \dots\dots\dots (1)$$

$$\Sigma yt = a\Sigma t + b\Sigma t^2 \dots\dots\dots(2)$$

Where n is the number of paired observations

The normal equation are obtained by multiplying $y = a + bt$, by the coefficient of a and b , i.e., by 1 and t throughout and summing up.

When the Number of Years is Odd

We can use this method when we are given odd number of years. It is easy and is widely used in practice. If the number of items is odd, we can follow the following steps:

1. Denote time as the t variable and values as y
2. Middle year is assumed as the period of origin and find out deviations
3. Square the time deviations and find t^2 .
4. Multiply the given value of y by the respective deviation of t and find the total Σty .
5. Find out the values of y ; get Σy
6. The value so obtained are placed in the two quations
 - i. $\Sigma y = na + b\Sigma t$
 - ii. $\Sigma yt = a\Sigma t + b\Sigma t^2$; find out the value of a and b
7. The calculated values of a and b are substituted and the trend value of y are found for various values of t .

When the number of years is odd the calculation will be simplified by taking the mid year as origin and one year as unit and in that case

$\Sigma t = 0$ and the two normal equations take the form

$$\Sigma y = na ; \Sigma yt = b\Sigma t^2$$

$$\text{Hence } a = \frac{\Sigma y}{n} , b = \frac{\Sigma yt}{\Sigma t^2}$$

Example :

Calculate trend values by the method of least square from data given below and estimate the sales for 2003

NOTES

| | | | | | |
|--------------------------|------|------|------|------|------|
| Years: | 1996 | 1997 | 1998 | 1999 | 2000 |
| Sales of Co.A, (₹ Lakhs) | 70 | 74 | 80 | 86 | 90 |

Solution:

| Year | Sales | Deviation from 1998 | | |
|--------------|-----------------|---------------------|-----------------|----------------------------|
| | y | t | ty | t ² |
| 1996 | 70 | -2 | -140 | 4 |
| 1997 | 74 | -1 | -74 | 1 |
| 1998 | 80 | 0 | 0 | 0 |
| 1999 | 86 | 1 | 86 | 1 |
| 2000 | 90 | 2 | 180 | 4 |
| n = 5 | Σy = 400 | Σt = 0 | Σty = 52 | Σt² = 10 |

Since $\Sigma t = 0$

$$a = \frac{\Sigma y}{n} = \frac{400}{5} = 80, \quad b = \frac{\Sigma ty}{\Sigma t^2} = \frac{52}{10} = 5.2$$

Hence, $y = 80 + 5.2 \times t$

$$\text{Therefore } y_{1996} = 80 + 5.2 (-2) = 80 - 10.4 = 69.6$$

$$y_{1997} = 80 + 5.2 (-1) = 80 - 5.2 = 74.8$$

$$y_{1998} = 80 + 5.2 (0) = 80 + 0 = 80$$

$$y_{1999} = 80 + 5.2 (1) = 80 + 5.2 = 85.2$$

$$y_{2000} = 80 + 5.2 (2) = 80 + 10.4 = 90.4$$

For 2003, t will be 5. Putting $t = 5$ in the equation

$$Y_{2003} = 80 + 5.2 (5) = 80 + 26 = 106$$

Thus the estimated sales for the year 2003 is ₹106 lakhs

When the Number of Years is Even

When the number of years is even the origin is placed in the midway between the two middle years and the unit is taken to be half year instead of one year. With this change of origin and scale we have

NOTES

$$\Sigma t = 0$$

$$\text{Hence } a = \frac{\Sigma y}{n}, b = \frac{\Sigma yt}{\Sigma t^2}$$

Example:

Production of a company for 6 consecutive years is given in the following table. Calculate the trend value by using the method of least square

| | | | | | | |
|------------|------|------|------|------|------|------|
| Year | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 |
| Production | 12 | 13 | 18 | 20 | 24 | 28 |

Solution:

| Year | Sales | Deviation from 2002.5 | | | Trend values |
|--------------|-----------------|-----------------------|-------------------|------------------------------|--------------|
| | | t | ty | t ² | |
| 2000 | 12 | -2.5 | -30 | 6.25 | 11.5 |
| 2001 | 13 | -1.5 | -19.5 | 2.25 | 14.5 |
| 2002 | 18 | -0.5 | -9 | 0.25 | 17.53 |
| 2003 | 20 | 0.5 | 10 | 0.25 | 20.81 |
| 2004 | 24 | 1.5 | 36 | 2.25 | 24.09 |
| 2005 | 28 | 2.5 | 70 | 6.25 | 27.37 |
| n = 6 | Σy = 115 | Σt = 0 | Σty = 57.5 | Σt² = 17.5 | |

Since $t = 0$

$$a = \frac{\Sigma y}{n} = \frac{115}{6} = 19.17, b = \frac{\Sigma yt}{\Sigma t^2} = \frac{57.5}{17.5} = 3.28$$

Hence, $y = 19.17 + 3.28 \times t$

Therefore

$$y_{2000} = 19.17 + 3.28 (- 2.5) = 19.17 - 8.2 = 11.5$$

$$y_{2001} = 19.17 + 3.28 (- 1.5) = 19.17 - 4.92 = 14.5$$

$$y_{2002} = 19.17 + 3.28 (- 0.5) = 19.17 - 1.64 = 17.53$$

$$y_{2003} = 19.17 + 3.28 (0.5) = 19.17 + 1.64 = 20.81$$

$$y_{2004} = 19.17 + 3.28 (1.5) = 19.17 + 4.92 = 24.09$$

NOTES

$$y_{2005} = 19.17 + 3.28 (2.5) = 19.17 + 8.2 = 27.37$$

Merits

1. The method is mathematically sound.
2. The estimates a and b are unbiased.
3. The least square method gives trend values for all the years and the method is devoid of all kinds of subjectivity.
4. The algebraic sum of deviations of actual values from trend values is zero and the sum of the deviations is minimum.

Demerits

1. The least square method is highly mathematical; therefore, it is difficult for a layman to understand it.
2. The method is not flexible.
3. It has been assumed that y is only a linear function of time period n. This may not be true in any situations.

9.4 MEASUREMENT OF SEASONAL VARIATION

Seasonal variations are that rhythmic changes in the time series data that is regular and periodic variations having a period of one year duration. Some of the examples which show seasonal variations are production of cold drinks, which are high during summer months and low during winter season. Sales of sarees in a cloth store which are high during festival season and low during other periods. They have their origin in climatic or institutional factors that affect either supply or demand or both. It is important that these variations should be measured accurately. The reason for determining seasonal variations in a time series is to isolate it and to study its effect on the size of the variable in the index form which is usually referred as seasonal index.

9.4.1 METHODS OF CONSTRUCTING SEASONAL INDICES

There are four methods of constructing seasonal indices.

1. Simple averages method
2. Ratio to trend method
3. Percentage moving average method
4. Link relatives method

Simple Average Method :

The time series data for each of the 4 seasons (for quarterly data) of a particular year are expressed as percentages to the seasonal average for that year. The percentages for different seasons are averaged over the years by using simple average. The resulting percentages for each of the 4 seasons then constitute the required seasonal indices.

Steps to calculate Simple Average Method:

- (i) Arrange the data by months, quarters or years according to the data

NOTES

given.

- (ii) Find the sum of the each months, quarters or year.
- (iii) Find the average of each months, quarters or year.
- (iv) Find the average of averages, and it is called Grand Average (G)
- (v) Compute Seasonal Index for every season (i.e) months, quarters or year is given by

$$\text{Seasonal Index (S.I)} = \frac{\text{Seasonal Average}}{\text{Grandaverage}} \times 100$$

If the data is given in months

$$\text{Seasonal Index for Jan (S.I)} = \frac{\text{monthly Average (for jan)}}{\text{Grand average}} \times 100$$

$$\text{Seasonal Index for Feb (S.I)} = \frac{\text{monthly Average (for feb)}}{\text{Grandaverage}} \times 100$$

Similarly we can calculate SI for all other months

Example:

Calculate the seasonal index for the quarterly production of a computer using method of simple average

| Year | I Quarter | II Quarter | III Quarter | IV Quarter |
|------|-----------|------------|-------------|------------|
| 2011 | 355 | 451 | 525 | 500 |
| 2012 | 369 | 410 | 496 | 510 |
| 2013 | 391 | 432 | 458 | 495 |
| 2014 | 298 | 389 | 410 | 457 |
| 2015 | 300 | 390 | 431 | 459 |
| 2016 | 350 | 400 | 450 | 500 |

Solution:

| Year | I Quarter | II Quarter | III Quarter | IV Quarter |
|------|-----------|------------|-------------|------------|
| 2011 | 355 | 451 | 525 | 500 |
| 2012 | 369 | 410 | 496 | 510 |

NOTES

| | | | | |
|--------------------|--------|------|--------|--------|
| 2013 | 391 | 432 | 458 | 495 |
| 2014 | 298 | 389 | 410 | 457 |
| 2015 | 300 | 390 | 431 | 459 |
| 2016 | 350 | 400 | 450 | 500 |
| Quarterly Total | 2063 | 2472 | 2770 | 2921 |
| Quarterly Averages | 343.83 | 412 | 461.67 | 486.83 |

$$\text{Seasonal Index (S.I)} = \frac{\text{Seasonal Average}}{\text{Grand Average}} \times 100$$

$$\text{Grand average} = \frac{343.83 + 412 + 461.67 + 486.83}{4} = \frac{1704.33}{4} = 426.0825$$

$$\text{S.I for I Q} = \frac{343.83}{426.0825} \times 100 = 80.69$$

$$\text{S.I for II Q} = \frac{412}{426.0825} \times 100 = 96.69$$

$$\text{S.I for III Q} = \frac{461.67}{426.0825} \times 100 = 108.35$$

$$\text{S.I for IV Q} = \frac{486.83}{426.0825} \times 100 = 114.26$$

Advantage and Disadvantage:

- Method of simple average is easy and simple to execute.
- This method is based on the basic assumption that the data do not contain any trend and cyclic components. Since most of the economic and business time series have trends and as such this method though simple is not of much practical utility.

CHECK YOUR PROGRESS - 1

1. A time series is a set of data recorded _____
2. The terms prosperity, recession, depression and recovery are in particular attached to _____
3. What is time series?

NOTES

9.5 FORECASTING

Time series forecasting methods produce forecasts based solely on historical values and they are widely used in business situations where forecasts of a year or less are required. These methods used are particularly suited to Sales, Marketing, Finance, Production planning etc. and they have the advantage of relative simplicity. Time series forecasting is a technique for the prediction of events through a sequence of time.

The technique is used across many fields of study, from geology to economics. The techniques predict future events by analyzing the trends of the past on the assumption that future trends will hold similar to historical trends. Data is organized around relatively deterministic timestamps, and therefore, compared to random samples, may contain additional information that is tried to extract.

- Time series methods are better suited for short-term forecasts (i.e., less than a year).
- Time series forecasting relies on sufficient past data being available and that the data is of a high quality and truly representative.
- Time series methods are best suited to relatively stable situations. Where substantial fluctuations are common and underlying conditions are subject to extreme change, then time series methods may give relatively poor results.

Advantages of forecasting:

1. Helps to predict the future:
2. Learns from the past
3. Remain competitive
4. Prepare for new business

Disadvantages of forecasting:

1. Basis of forecasting
2. Reliability of past data
3. Time and cost factor

9.6 DESEASONALISATION

When the seasonal component is removed from the original data, the reduced data are free from seasonal variations and is called deseasonalised data. That is, under a multiplicative model

$$\frac{T \times S \times C \times I}{S} = T \times C \times I$$

NOTES

Deseasonalised data being free from the seasonal impact manifest only average value of data.

Seasonal adjustment can be made by dividing the original data by the seasonal index.

$$\text{Deseasonalised data} = \frac{\text{ORIGINAL DATA}}{\text{SEASONAL INDEX}} \times 100$$

where an adjustment-multiplier 100 is necessary because the seasonal indices are usually given in percentages.

In case of additive model

$$Y_t = T + S + C + I$$

$$\begin{aligned} \text{Deseasonalised data} &= \text{original data} - \frac{\text{seasonal index}}{100} \\ &= Y_t - \frac{\text{seasonal index}}{100} \end{aligned}$$

CHECK YOUR PROGRESS - 2

4. Define forecasting?
5. What are the method used for finding seasonal indices

9.7 SUMMARY

- Time series are influenced by a variety of forces. Some are continuously effective other make themselves felt at recurring time intervals, and still others are non-recurring or random in nature. Therefore, the first task is to break down the data and study each of these influences in isolation. This is known as decomposition of the time series.
- The objective of the time series analysis is to identify the magnitude and direction of trends, to estimate the effect of seasonal and cyclical variations and to estimate the size of the residual component. This implies the decomposition of a time series into its several components. Two lines of approach are usually adopted in analyzing a given time series:
 - The additive model, the multiplicative model
- Moving average method is a simple device of reducing fluctuations and obtaining trend values with a fair degree of accuracy. In this method the average value of a number of years (months, weeks, or days) is taken as the trend value for the middle point of the period of moving average. The process of averaging smoothes the curve and reduces the fluctuations.
- When the trend is linear the trend equation may be represented by

NOTES

$y = a + bt$ and the values of a and b for the line $y = a + bt$ which minimizes the sum of squares of the vertical deviations of the actual (observed) values from the straight line, are the solutions to the so called normal equations:

- Seasonal variations are that rhythmic changes in the time series data that is regular and periodic variations having a period of one year duration.
- There are four methods of constructing seasonal indices. They are Simple averages method, Ratio to trend method, Percentage moving average method, Link relatives method.
- Time series forecasting methods produce forecasts based solely on historical values and they are widely used in business situations where forecasts of a year or less are required.
- When the seasonal component is removed from the original data, the reduced data are free from seasonal variations and is called deseasonalised data.

9.8 KEY WORDS

Time series, decomposition of the time series, additive model, the multiplicative model, Moving average method , least square method , Seasonal variations , Simple averages method, Ratio to trend method, Percentage moving average method, Link relatives method, forecasting , deseasonalised.

9.9 ANSWERS TO CHECK YOUR PROGRESS

1. Periodically, at equal time intervals, at successive points of time
2. Cyclical movements
3. Time series are influenced by a variety of forces. Some are continuously effective other make themselves felt at recurring time intervals, and still others are non-recurring or random in nature.
4. Time series forecasting methods produce forecasts based solely on historical values and they are widely used in business situations where forecasts of a year or less are required
5. There are four methods of constructing seasonal indices.
 1. Simple averages method
 2. Ratio to trend method
 3. Percentage moving average method
 4. Link relatives method

9.10 QUESTION AND EXERCISE

SHORT ANSWER QUESTION

1. What is time series?
2. What are the uses of time series
3. What are basic types of variations

LONG ANSWER QUESTION

1. Explain the components of time series
2. What are the various methods of estimating the trend components
3. Explain the moving average method? How is it calculated?
4. Describe the method of finding seasonal indices
5. Calculate the trend value by using three yearly moving average of the following data

| Year | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 |
|------------|------|------|------|------|------|------|------|------|------|------|
| Production | 21 | 22 | 23 | 25 | 24 | 22 | 25 | 26 | 27 | 26 |

9.11 FURTHER READINGS

1. Spiegel, Murray R.: Theory and Practical of Statistics., London McGraw Hill Book Company.
2. Yamane, T.: Statistics: An Introductory Analysis, New York, HarperedRow Publication
3. R.P. Hooda: Statistic for Business and Economic, McMillan India Ltd.
4. G.C. Beri: Statistics for Mgt., TMH.
5. J.K. Sharma: Business Statistics, Pearson Education.
6. 6. S.P. Gupta : Statistical Methods, Sultan Chand and Sons.

UNIT 10 - SAMPLING

Structure

- 10.0 Introduction
- 10.1 Objectives
- 10.2 Basic Concept of Sampling
- 10.3 Sampling Methods
 - 10.3.1 Random Sampling
 - 10.3.2 Non Random Sampling
- 10.4 Sampling and Non Sampling Errors
- 10.5 Sampling Distribution
- 10.6 Procedure for Hypothesis
- 10.7 Null and Alternative Hypothesis
- 10.8 Errors in Hypothesis testing
- 10.9 One Tailed and Two Tailed Test
- 10.10 Summary
- 10.11 Key Words
- 10.12 Answers to Check Your Progress
- 10.13 Questions and Exercise
- 10.14 Further Readings

10.0 INTRODUCTION

In our day to day life we often examine some materials. We examine fruits and vegetables and dress materials and etc before we purchase them. This approach is applied to different fields of life. Products in factories are inspected to ensure the desired quality of the products. Medications are manufactured on a commercial scale when their effects have been tested on a sample of patients. Different fertilizers are tested on agricultural plots and different foods are tested on animals. Small dams are constructed as a sample in laboratories to study the life and other characteristics of big dams before they are actually constructed, and so on in each and every field we examine this process of inspection is very widespread and is commonly used on various occasions. However, this job is never done on a very large scale. This process is carried out on a small scale. On the basis of a small study, we make an opinion about the entire material under study.

10.1 OBJECTIVES

The student will be able to

- Understand the concept of sampling

- Understand sampling distribution of statistic;
- Understand different types of hypotheses;
- Determine type I and type II errors in hypotheses testing problems;
- Categorize one-sided and two-sided tests;
- Solve the problems of testing hypotheses concerning mean(s) and proportion(s) based on large samples.

Sampling:

A sample is defined as a smaller set of data that is chosen and/or selected from a larger population by using a predefined selection method. These elements are known as sample points, sampling units or observations. Creating a sample is an efficient method of conducting research as in most cases, it is impossible or very expensive and time consuming to research the whole population and hence researching the sample provides insights that can be applied to the whole population.

10.2 BASIC CONCEPTS OF SAMPLING

Population

The group of individuals considered under study is called as population. The word population here refers all items that have been chosen for the study. Thus in statistics, population can be number of cars manufactured in a day or week or month, number of fans manufactured in a day or week or month, number of fridge, TVs, chalk pieces, people, students, girls, boys, any manufacturing products, etc...

Finite and Infinite Population:

When the number of observations/individuals/products is countable in a group, then it is a finite population. Example: Height of all the students studying in BBA.

When the number of observations/individuals/products is uncountable in a group, then it is an infinite population. Example: number of rice in a sack, number of stones in the playground.

Sample and Sample size

A selection of a group of individuals from a population in such a way that it represents the population is called as sample and the number of individuals included in a sample is called the sample size.

Parameter and Statistic

Parameter: The statistical constants of the population like mean (m), variance (s^2) are referred as population parameters.

Statistic: Any statistical measure computed from sample is known as statistic

Sampling

NOTES

10.3 SAMPLING METHODS

The various methods of sampling can be grouped under

- 1) Probability sampling or random sampling
- 2) Non-probability sampling or non random sampling

10.3.1 RANDOM SAMPLING

Under this method, every unit of the population at any stage has equal chance (or) each unit is drawn with known probability. It helps to estimate the mean, variance etc of the population. Under probability sampling there are two procedures

1. Sampling with replacement
2. Sampling without replacement

When the successive draws are made with placing back the units selected in the preceding draws, it is known as sampling with replacement. When such replacement is not made it is known as sampling without replacement. When the population is finite sampling with replacement is adopted otherwise sampling without replacement is adopted.

Mainly there are many kinds of random sampling. Some of them are.

1. Simple Random Sampling
2. Systematic Random Sampling
3. Stratified Random Sampling
4. Cluster Sampling

1. Simple Random sampling

The basic probability sampling method is the simple random sampling. It is the simplest of all the probability sampling methods. It is used when the population is uniform. In simple random sampling the samples are selected in such a way that each and every unit in the population has an equal and independent chance of being selected as a sample. Simple random sampling may be done, with or without replacement of the samples selected. In a simple random sampling with replacement there is a chance of selecting the same sample any number of times. So, simple random sampling without replacement is followed. Thus in simple random sampling from a population of N units, the probability of drawing any unit at the first draw is $\frac{1}{N}$, the probability of drawing any unit in the second draw from among the available $(N - 1)$ units is $\frac{1}{(N-1)}$, and so on. .

Several methods have been adopted for random selection of the samples from the population. Of those, the following two methods are

generally used

- i. Lottery method
- ii. Random number table method

i) Lottery method

This is most popular method and simplest method. In this method all the items of the universe are numbered on separate slips of paper of same size, shape and color. They are folded and mixed up in a drum or a box or a container. A blindfold selection is made. Required number of slips is selected for the desired sample size. The selection of items thus depends on chance.

ii) Random number table method

As the lottery method cannot be used when the population is infinite, the alternative method is using of table of random numbers. It is difficult to number all the items on separate slips of paper of same size, shape and colour. The most practical, easy and inexpensive method of selecting a random sample can be done through "Random Number Table". The random number table has been so constructed that each of the digits 0, 1, 2... 9 will appear approximately with the same frequency and independently of each other.

There are several standard tables of random numbers. But the credit for this technique goes to

- a. Prof. LHC. Tippett (1927) random number table consists of 10,400 four-figured numbers.
- b. Fishers and Yates (1938) comprising of 15,000 digits arranged in twos.
- c. Kendall and B.B Smith (1939) consisting of 1, 00,000 numbers grouped in 25,000 sets of 4 digit random numbers,
- d. Rand corporation (1955) consisting of 2, 00,000 random numbers of 5 digits each etc.,

Merits

- There is less possibility for personal bias.
- Sampling error can be calculated.
- This method is economical as it saves time, money and labour.

Demerits

- It cannot be applied if the population is varied.
- It requires a complete list of the population but up-to-date lists are not available in many enquires.
- If the size of the sample is small, then it will not be a representative of the population.

2. Systematic sampling

The systematic sampling technique is operationally more convenient than the simple random sampling. It also ensures at the same time that each unit has equal probability of inclusion in the sample. In this method of sampling, the first unit is selected with the help of random numbers and the remaining units are selected automatically according to a predetermined pattern. This method is known as systematic sampling.

A systematic sample is formed by selecting every item from the population, where k refers to the sample interval. The sampling interval can be determined by dividing the size of the population by the size of the sample to be chosen.

That is $k = \frac{N}{n}$, where k is an integer.

k = Sampling interval,

N = Size of the population,

n = Sample size.

Procedure for selection of samples by systematic sampling method

(i) If we want to select a sample of 10 students from a class of 100 students, the sampling interval is calculated as $k = \frac{N}{n} = \frac{100}{10} = 10$

Thus sampling interval = 10 denotes that for every 10 samples one sample has to be selected.

(ii) The first sample is selected from the first 10 (sampling interval) samples through random selection procedures.

(iii) If the selected first random sample is 5, then the rest of the samples are automatically selected by incrementing the value of the sampling interval ($k = 10$) i.e., 5, 15, 25, 35, 45, 55, 65, 75, 85, 95.

Example:

Suppose we have to select 10 items out of 3,000. The procedure is to number all the 3,000 items from 1 to 3,000.

The sampling interval is calculated as $k = N/n = 3000/10 = 300$.

Thus sampling interval = 300 denotes that for every 300 samples one sample has to be selected. The first sample is selected from the first 300 (sampling interval) samples through random selection procedures. If the selected first random sample is 50, then the rest of the samples are automatically selected by incrementing the value of the sampling interval ($k=300$) i.e., 50, 350, 650, 950, 1250, 1550, 1850, 2150, 2450, 2750. Items bearing those numbers will be selected as samples from the population.

Merits

- This is simple and suitable method.

- This method distributes the sample more evenly over the entire listed population.
- The time and work is reduced much

Demerits

- Systematic samples are not random samples.
- If N is not a multiple of n , then the sampling interval (k) cannot be an integer, thus sample selection becomes difficult.

3. Stratified Sampling

When the population is varied with respect to the characteristic in which we are interested, we adopt stratified sampling.

When the varied population is divided into homogenous sub-population, the sub-populations are called strata. From each stratum a separate sample is selected using simple random sampling. This sampling method is known as stratified sampling.

We may stratify by size of farm, type of crop, soil type, etc.

The number of units to be selected may be uniform in all strata (or) may vary from stratum to stratum. There are four types of allocation of strata

1. **Equal allocation:** If the number of units to be selected is uniform in all strata it is known as equal allocation of samples.
2. **Proportional allocation:** If the number of units to be selected from a stratum is proportional to the size of the stratum, it is known as proportional allocation of samples.
3. **Neyman's allocation:** When the costs for different strata are equal, it is known as Neyman's allocation.
4. **Optimum allocation:** When the cost per unit varies from stratum to stratum, it is known as optimum allocation

Merits

- It is more representative.
- It ensures greater accuracy.
- It is easy to administrate as the universe is sub-divided.

Demerits

- To divide the population into homogeneous strata, it requires more money, time and statistical experience which is a difficult one.
- If proper stratification is not done, the sample will have an effect of bias.

4. Cluster sampling

Cluster sampling is a sampling method in which the entire population of the study is divided into externally homogeneous but

internally heterogeneous groups called clusters. Essentially, each cluster is a mini-representation of the entire population. For example if we have to conduct the an opinion poll in the city of Chennai, the city may be divided into, as 50 blocks and out of these 50 blocks 5 blocks can be picked up by random sampling and inhabitants in these five blocks can be interviewed to give their opinion on a particular issues

When using this method, it should be seen that clusters are of as small in size as possible and the number of sample units in each cluster should be more or less the same. This method is commonly used in collecting data about some common characteristics of the population.

Merits

- Cheap, quick and easy
- Large sample size
- Convenient to obtain
- Cost effective

Demerits

- Least Representative
- High sampling error
- Less efficient
- Sometimes not appropriate

10.3.2 NON-RANDOM SAMPLING

Non-probability sampling is a sampling technique in which the researcher selects samples based on the subjective judgment of the researcher rather than random selection.

In non-probability sampling, not all members of the population have a chance of participating in the study unlike probability sampling, where each member of the population has a known chance of being selected.

Non-probability sampling is most useful for exploratory studies like pilot survey (a survey that is deployed to a smaller sample compared to pre-determined sample size). Non-probability sampling is used in studies where it is not possible to draw random probability sampling due to time or cost considerations.

Non-probability sampling is a less stringent method, this sampling method depends heavily on the expertise of the researchers. Non-probability sampling is carried out by methods of observation and is widely used in qualitative research.

Types of non-probability sampling and examples

1. Convenience Sampling:

Convenience sampling is a non-probability sampling technique where samples are selected from the population only because they are conveniently available to researcher. These samples are selected only because they are easy to recruit and researcher did not consider selecting sample that represents the entire population.

Ideally, in research, it is good to test sample that represents the population. But, in some research, the population is too large to test and consider the entire population. This is one of the reasons, why researchers rely on convenience sampling, which is the most common non-probability sampling technique, because of its speed, cost-effectiveness, and ease of availability of the sample.

An example of convenience sampling would be using student volunteers known to researcher. Researcher can send the survey to students and they would act as sample in this situation.

2. Consecutive Sampling:

This non-probability sampling technique is very similar to convenience sampling, with a slight variation. Here, the researcher picks a single person or a group of sample, conducts research over a period of time, analyzes the results and then moves on to another subject or group of subject if needed.

Consecutive sampling gives the researcher a chance to work with many subjects and fine tune his/her research by collecting results that have vital insights.

3. Quota Sampling:

Hypothetically consider, a researcher wants to study the career goals of male and female employees in an organization. There are 500 employees in the organization. These 500 employees are known as population. In order to understand better about a population, researcher will need only a sample, not the entire population. Further, researcher is interested in particular strata within the population. Here is where quota sampling helps in dividing the population into strata or groups.

For studying the career goals of 500 employees, technically the sample selected should have proportionate numbers of males and females. Which means there should be 250 males and 250 females? Since, this is unlikely, the groups or strata are selected using quota sampling.

4. Judgmental or Purposive Sampling:

In judgmental sampling, the samples are selected based purely on researcher's knowledge and credibility. In other words, researchers choose only those who he feels are a right fit (with respect to attributes and representation of a population) to participate in research study.

This is not a scientific method of sampling and the downside to this sampling technique is that the results can be influenced by the preconceived notions of a researcher. Thus, there is a high amount of ambiguity involved in this research technique.

For example, this type of sampling method can be used in pilot studies.

5. Snowball Sampling:

Snowball sampling helps researchers find sample when they are difficult

to locate. Researchers use this technique when the sample size is small and not easily available. This sampling system works like the referral program. Once the researchers find suitable subjects, they are asked for assistance to seek similar subjects to form a considerably good size sample.

For example, this type of sampling can be used to conduct research involving a particular illness in patients or a rare disease. Researchers can seek help from subjects to refer other subjects suffering from the same ailment to form a subjective sample to carry out the study.

10.4 SAMPLING AND NON-SAMPLING ERRORS

Sample is a part of the total population. Sample drawn from the population depends upon chance and all the characteristics of the population may not be present in the sample drawn from the same population. The errors involved in the collection, processing and analysis of the data may be broadly classified into two categories namely,

- i. Sampling Errors
- ii. Non-Sampling Errors

i. Sampling Errors

Errors, which arise in the normal way of investigation or details on account of chance, are called sampling errors. Sampling errors are intrinsic in the method of sampling. They may arise accidentally without any bias or prejudice. Sampling Errors arise mostly due to the following reasons:

Wrong selection of the sample instead of correct sample by defective sampling technique.

- The researcher substitutes a suitable sample if the original sample is not available while research.
- In area surveys, while dealing with border lines it depends upon the investigator whether to include them in the sample or not. This is known as Faulty demarcation of sampling units.

ii. Non-Sampling Errors

The errors that arise due to human factors which always vary from one investigator to another in selecting, estimating or using measuring instruments(tape, scale)are called Non-Sampling errors. The errors may arise in the following ways:

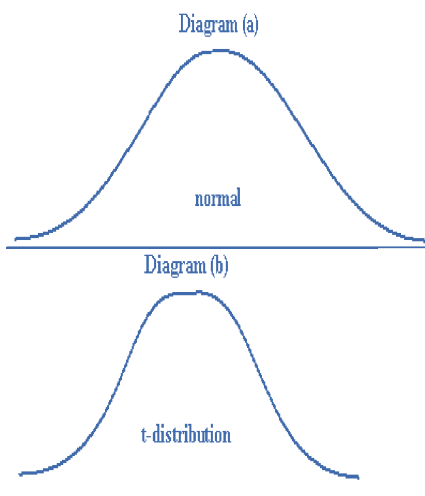
- Negligence and carelessness of the part of either researcher or respondents.
- Lack of trained and qualified investigators.
- Framing of a wrong questionnaire.
- Applying wrong statistical measure
- Incomplete investigation and sample survey.

10.5 SAMPLING DISTRIBUTION

Sampling distribution or finite-sample distribution is the probability distribution of a given statistic based on a random sample. Sampling distributions are vital in statistics because they offer a major simplification en-route to statistical implication. It can be explained that the sample distribution is the distribution that result from the collection of real data. A major attribute of a sample is that it contains a finite (countable) number of scores, the number of scores represented by the letter 'n'. The value of a statistic varies from one sample to another. Therefore, it is a random variable and its probability distribution is known as its sampling distribution.

The probability distribution of all possible values of \bar{X} calculated from all possible simple random samples are called the sampling distribution of \bar{X} . In brief, we shall call it the distribution of \bar{X} . The mean of this distribution is called the expected value of \bar{X} and is written as $E\bar{X}$ or $\mu\bar{X}$. The standard deviation (standard error) of this distribution is denoted by S.E. (\bar{X}) or $\sigma\bar{X}$ and the variance of \bar{X} is denoted by $\text{Var}(\bar{X})$ or $\sigma^2 \bar{X}$. The distribution of (\bar{X}) has some important properties:

- One important property of the distribution of \bar{X} is that it is a normal distribution when the size of the sample is large. When the sample size n is more than 30, we call it a large sample size. The shape of the population distribution does not matter. The population may be normal or non-normal, the distribution of \bar{X} is normal for $n > 30$, but this is true when the number of samples is very large. As the distribution of random variable \bar{X} is normal, \bar{X} can be transformed into a standard normal variable Z where $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$. The distribution of \bar{X} has a t-distribution when the population is normal and $n \leq 30$. Diagram (a) shows the normal distribution and diagram (b) shows the t-distribution.



- The mean of the distribution of \bar{X} is equal to the mean of the population. Thus $E(\bar{X}) = \mu$ ($\bar{X} = \mu$ (population mean)). This relation is true for small as well as large sample sizes in sampling without replacement and with replacement.
- The standard error (standard deviation) of \bar{X} is related to the standard deviation of the population σ through the relations:
$$\text{S.E.}(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$
- This is true when population is infinite, which means N is very large or the sampling is done with replacement from a finite or infinite population.

$$\text{S.E.}(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

This is true when sampling is without replacement from a finite population. The above two equations between $\sigma_{\bar{X}}$ and σ are true both for small as well as large sample sizes.

10.6 PROCEDURE FOR HYPOTHESIS

Standard Error:

The standard deviation of a sampling distribution of a statistic is often called as standard error. Standard Error

The standard deviation of a statistic is called the standard error of that statistic. If the statistic is \bar{X} the standard deviation of all possible values of \bar{X} is called the standard error of \bar{X} , which may be written as $\text{S.E.}(\bar{X})$ or $\sigma_{\bar{X}}$. Similarly, if the sample statistic is proportion \hat{p} , the standard deviation of all possible values of \hat{p} is called the standard error of \hat{p} and is denoted by $\sigma_{\hat{p}}$ or $\text{S.E.}(\hat{p})$.

10.7 NULL HYPOTHESIS AND ALTERNATIVE HYPOTHESIS

A hypothesis can be defined as a statement on the population or the values of the unknown parameters associated with the respective probability distribution. All the hypotheses should be tested for their validity using statistical concepts and a representative sample drawn from the study population. The statistical hypothesis testing plays an important role, among others, in various fields including industry, biological sciences, behavioral sciences and Economics.

A statistical test is a procedure governed by certain rules, which lead to take a decision about the null hypothesis for its rejection or otherwise on the basis of sample values. This process is called statistical hypotheses testing.

The statistical hypothesis is an assumption about the value of some unknown parameter, and the hypothesis provides some numerical

value or range of values for the parameter. In each hypotheses testing problem, we will find there are two hypotheses to choose between null hypothesis and alternative hypothesis.

Null Hypothesis:

The Null Hypothesis denoted by H_0 asserts that there is no true difference between the sample of data and the population parameter and that the difference is accidental which is caused due to the fluctuations in sampling. Thus, a null hypothesis states that there is no difference between the assumed and actual value of the parameter.

Alternative Hypothesis:

The alternative hypothesis denoted by H_1 is the other hypothesis about the population, which stands true if the null hypothesis is rejected. Thus, if we reject H_0 then the alternative hypothesis H_1 gets accepted.

For example, if we test whether the population mean has a specified value μ_0 , then the null

hypothesis would be expressed as:

$$H_0: \mu = \mu_0$$

The alternative hypothesis may be formulated suitably as anyone of the following:

1. $H_1: \mu \neq \mu_0$
2. $H_1: \mu > \mu_0$
3. $H_1: \mu < \mu_0$

The alternative hypothesis in equation (1) is known as two-sided alternative and the alternative hypothesis in equation (2) is known as one-sided (right) alternative and equation (3) is known as one-sided (left) alternative.

10.8 ERRORS IN STATISTICAL HYPOTHESES TESTING

The Hypothesis Testing is a statistical test used to determine whether the hypothesis assumed for the sample of data stands true for the entire population or not. Simply, the hypothesis is an assumption which is tested to determine the relationship between two data sets.

In hypothesis testing, two opposing hypotheses about a population are formed Viz. Null Hypothesis (H_0) and Alternative Hypothesis (H_1). The Null hypothesis is the statement which asserts that there is no difference between the sample statistic and population parameter and is the one which is tested, while the alternative hypothesis is the statement which stands true if the null hypothesis is rejected.

It must be recognized that the final decision of rejecting H_0 or not rejecting H_0 may be incorrect. The error committed by rejecting H_0 , when H_0 is really true, is called **type I error**. The error committed by not

rejecting H_0 , when H_0 is false, is called **type II error**.

Example:

A Television manufacturing company introduces a new model of television. Daily sales of the new Television, in a city, are assumed to be distributed with mean sales of ₹80,000 and standard deviation of ₹5,000 per day. The Advertising Manager of the company considers placing advertisements in TV Channels. He does this on 10 random days and tests to see whether or not sales have increased. Formulate suitable null and alternative hypotheses. What would be type I and type II errors?

Solution:

The Advertising Manager is testing whether or not sales increased more than ₹80,000. Let μ be the average amount of sales, if the advertisement does appear.

The null and alternative hypotheses can be framed based on the given information as follows:

Null hypothesis: $H_0: \mu = 80000$

i.e., The mean sales due to the advertisement is not significantly different from ₹80,000.

Alternative hypothesis: $H_1: \mu > 80000$

i.e., Increase in the mean sales due to the advertisement is significant.

(i) **If type I error** occurs, then it will be concluded as the advertisement has improved sales. But, really it is not.

(ii) **If type II error** occurs, then it will be concluded that the advertisement has not improved the sales. But, really, the advertisement has improved the sales.

The following may be the penalties due to the occurrence of these errors:

If type I error occurs, then the company may spend towards advertisement. It may increase the expenditure of the company. On the other hand, if type II error occurs, then the company will not spend towards advertisement. It may not improve the sales of the company.

Example:

In court room, a defendant is considered not guilty as long as his guilt is not proven. The prosecutor tries to prove the guilt of the defendant. Only when there is enough charging evidence the defendant is condemned. In the start of the procedure, there are two hypotheses

H_0 : "the defendant is not guilty", and

H_1 : "the defendant is guilty".

The first one is called null hypothesis, and the second one is called

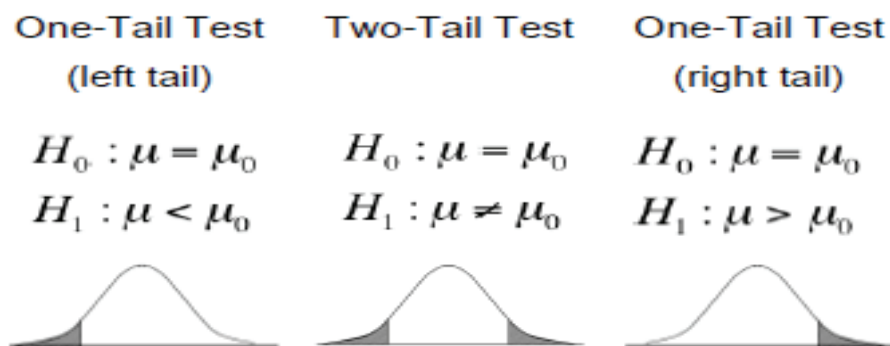
alternative (hypothesis).

10.9 ONE – TAILED AND TWO - TAILED TESTS

Two-Tailed Test is that where the hypothesis about the population parameter is rejected for the value of sample statistic failing into either tail of the distribution

When the hypothesis about the population parameter is rejected for the value of sample statistic failing into one side tail of the distribution, then it is known as one-tailed test.

If the rejection area falls on the right side, then it is called right-tailed test.) On the other hand If the rejection area falls on the left side, then it is called left-tailed test.



Example:

An insurance company is reviewing its current policy rates. When originally setting the rates they believed that the average claim amount will be maximum Rs180000. They are concerned that the true mean is actually higher than this, because they could potentially lose a lot of money. They randomly select 40 claims, and calculate a sample mean of Rs195000. Assuming that the standard deviation of claims is Rs50000 and set $\alpha = .05$, test to see if the insurance company should be concerned or not.

Solution:

Step 1: Set the null and alternative hypotheses

$$H_0 : \mu \leq 180000$$

$$H_1 : \mu > 180000$$

Step 2: Calculate the test statistic

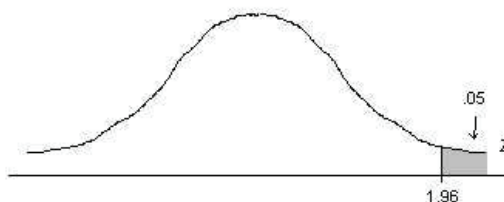
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = 1.897$$

Step 3: Set Rejection Region

Sampling

NOTES

Self-Instructional Material



Step 4: Conclude: We can see that $1.897 > 1.65$, thus our test statistic is in the rejection region. Therefore we fail to accept the null hypothesis. The insurance company should be concerned about their current policies.

Example:

A cosmetics company fills its best-selling 88 ml jars of facial cream by an automatic dispensing machine. The machine is set to dispense a mean of 8.1 ml per jar. Uncontrollable factors in the process can shift the mean away from 8.1 and cause either under fill or overfill, both of which are undesirable. In such a case the dispensing machine is stopped and recalibrated. Regardless of the mean amount dispensed, the standard deviation of the amount dispensed always has value 0.22 ml. A quality control engineer routinely selects 30 jars from the assembly line to check the amounts filled. On one occasion, the sample mean is $\bar{x} = 8.2$ ml and the sample standard deviation is $s = 0.25$ ml. Determine if there is sufficient evidence in the sample to indicate, at the 1% level of significance, that the machine should be recalibrated.

Solution:

- **Step 1.** The natural assumption is that the machine is working properly. Thus if μ denotes the mean amount of facial cream being dispensed, the hypothesis test is

$$H_0: \mu = 8.1$$

$$H_a: \mu \neq 8.1$$

$$\alpha = 0.01$$

- **Step 2.** The sample is large and the population standard deviation is known. Thus the test statistic is

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- **Step 3.** Inserting the data into the formula for the test statistic gives

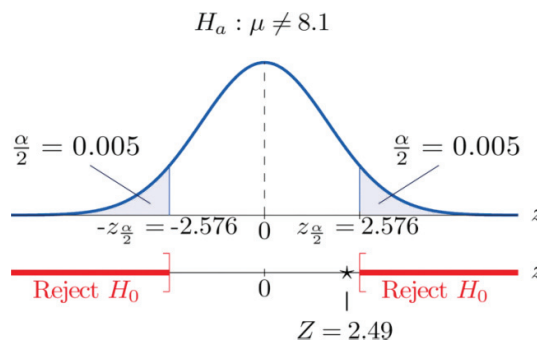
$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{8.2 - 8.1}{0.22 / \sqrt{30}} = 2.490$$

- **Step 4.** Since the symbol in H_a “ \neq ” this is a two-tailed test, so there are two critical values, $\pm z_{\alpha/2} = \pm z = 0.005$, which from the last line we read off as ± 2.576 .

The rejection region is $(-\infty, -2.576] \cup [2.576, \infty)$.

- **Step 5.** The test statistic does not fall in the rejection region. The decision is not to reject H_0 . In the context of the problem our conclusion is:

The data do not provide sufficient evidence, at the 1% level of significance, to conclude that the average amount of product dispensed is different from 8.1 ml. We conclude that the machine does not need to be recalibrated.



CHECK YOUR PROGRESS - 1

1. What is sampling?
2. When the alternative hypothesis is $H_1: \mu \neq \mu_0$, the critical region will be determined by _____
3. Rejecting H_0 when it is true called _____
4. What is null hypothesis?
5. What are type I and type II errors?

10.10 SUMMARY

- Statistic is a random variable and its probability distribution is called sampling distribution.
- Standard error is the standard deviation of the sampling distribution.
- Hypothesis is a statement on the population or the values of the parameters.
- Null hypothesis is a hypothesis which is tested for possible rejection.
- Statistical test leads to take decision on the null hypothesis.
- In each statistical hypotheses testing problem, there is one null hypothesis and one alternative hypothesis.
- Type I error is rejecting the true null hypothesis.
- Type II error is not rejecting a false null hypothesis.
- Two-Tailed Test is that where the hypothesis about the population parameter is rejected for the value of sample statistic falling into

either tail of the distribution

- When the hypothesis about the population parameter is rejected for the value of sample statistic failing into one side tail of the distribution, then it is known as one-tailed test.

10.11 KEY WORDS

Sampling, Sampling Distribution, Type I error and Type II error, One tailed and Two tailed tests, Hypothesis testing

10.12 ANSWER TO CHECK YOUR PROGRESS

1. A sample is defined as a smaller set of data that is chosen and/or selected from a larger population by using a predefined selection method
2. Both right and left tails
3. Type I error
4. A null hypothesis states that there is no difference between the assumed and actual value of the parameter
5. The error committed by rejecting H_0 , when H_0 is really true, is called **type I error**. The error committed by not rejecting H_0 , when H_0 is false, is called **type II error**.

10.13 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTIONS

1. Describe Decision Table.
2. What are type I and type II errors?
3. What do you mean by level of significance?
4. Explain one-tailed and two-tailed tests.
5. Explain critical value.

LONG ANSWR QUESTIONS

1. Describe briefly various types of sampling methods and give brief description of each
2. Discuss the null and alternative hypothesis.
3. Explain one and two tail test in detail with example
4. A set of 100 students is selected randomly from an institution. The mean height of these students is 163 *cms* and the standard deviation is 10 *cms*. Calculate the value of the test statistic under $H_0 : \mu = 167$.
5. In a random sample of 50 students from school a, 35 of them preferred junk food. In another sample of 80 from school b, 40 of them preferred junk food. Find the standard error of the difference between two sample proportions.
6. If $m = 35$, $n = 40$, $x = 10.8$, $y = 11.9$, $s_x = 3$ and $s_y = 4$, then

calculate standard error of $x - y$.

10.14 FURTHER READINGS

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. SahityaBhawan Publishers andDistributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing CompanyLtd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw HillPublishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., NewDelhi.
5. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hallof India Pvt. Ltd., New Delhi.

Sampling

NOTES

UNIT11 - TEST OF HYPOTHESIS

Structure

- 11.0 Introduction
- 11.1 Objectives
- 11.2 Hypothesis Testing on Population Mean
 - 11.2.1 Population is Known
 - 11.2.2 Population is Unknown
- 11.3 Difference Between Mean of Two Populations
 - 11.3.1 Population Variance Known
 - 11.3.2 Population Variance Unknown
- 11.4 Test of Hypothesis for Population Proportion
- 11.5 Difference Between Two Proportion
- 11.6 Summary
- 11.7 Key Words
- 11.8 Answers to Check Your Progress
- 11.9 Questions and Exercise
- 11.10 Further Readings

11.0 INTRODUCTION

Hypothesis testing was introduced by Ronald Fisher, Jerzy Neyman, Karl Pearson and Pearson's son, Egon Pearson. Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically an assumption that we make about the population parameter.

11.1 OBJECTIVES

- Understand purpose of Hypothesis testing;
- Understand the procedure for tests of hypotheses based on large samples
- Solve the problems of testing hypotheses concerning mean(s) and proportion(s) based on large samples

11.2 HYPOTHESIS TESTING ON POPULATION MEAN

Two situations may arise out of this, first one is when the population

variance is known and the second situation is if the population variance is unknown.

11.2.1 POPULATION VARIANCE KNOWN

Steps:

1. Let μ and σ^2 be respectively the mean and the variance of the population under study, where σ^2 is known. If μ_0 is an admissible value of μ , then frame the null hypothesis as $H_0: \mu = \mu_0$ and choose the suitable alternative hypothesis from

a. (i) $H_1: \mu \neq \mu_0$ (ii) $H_1: \mu > \mu_0$ (iii) $H_1: \mu < \mu_0$

2. Let (X_1, X_2, \dots, X_n) be a random sample of n observations drawn from the population, where n is large ($n \geq 30$).
3. Let the level of significance be α .
4. Consider the test statistics $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ under H_0 . Here \bar{X} represents the sample mean, The approximate sampling distribution of the test statistics under H_0 is the $N(0,1)$ distribution
5. Calculate the value of Z for the given sample (x_1, x_2, \dots, x_n) as

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

6. Find the critical value, z_e , corresponding to α and H_1 from the following table

| | | | |
|----------------------------------|------------------|---------------|---------------|
| Alternative Hypothesis (H_1) | $\mu \neq \mu_0$ | $\mu > \mu_0$ | $\mu < \mu_0$ |
| Critical Value (z_e) | $z_{\alpha/2}$ | z_α | $-z_\alpha$ |

7. Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1 .

| | | | |
|----------------------------------|---------------------------|------------------|-------------------|
| Alternative Hypothesis (H_1) | $\mu \neq \mu_0$ | $\mu > \mu_0$ | $\mu < \mu_0$ |
| Rejection Rule | $ z_0 \geq z_{\alpha/2}$ | $z_0 > z_\alpha$ | $z_0 < -z_\alpha$ |

Test of Hypothesis
NOTES

Example:

A company producing batteries finds that mean life span of the population of its batteries is 200 hours with a standard deviation of 15 hours. A sample of 100 batteries randomly chosen is found to have the mean life span of 195 hours. Test, at 5% level of significance, whether the mean life span of the batteries is significantly different from 200 hours.

Solution:

Step 1 : Let μ and σ represent respectively the mean and standard deviation of the probability distribution of the life span of the batteries. It is given that $\sigma = 15$ hours. The null and alternative hypotheses are

Null hypothesis: $H_0: \mu = 200$

i.e., the mean life span of the batteries is not significantly different from 200 hours.

Alternative hypothesis: $H_1: \mu \neq 200$

i.e., the mean life span of the batteries is significantly different from 200 hours.

It is a two-sided alternative hypothesis.

Step 2 : Data

The given sample information are

Sample size (n) = 100, Sample mean (\bar{x}) = 195 hours

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic is Z , under H_0

Under the null hypothesis H_0 , Z follows the $N(0,1)$ distribution.

Step 5 : Calculation of Test Statistic

The value of Z under H_0 is calculated from

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{195 - 200}{15/\sqrt{100}} = -3.33$$

Thus; $|z| = 3.33$

Step 6 : Critical value

Since H_1 is a two-sided alternative, the critical value at $\alpha = 0.05$ is $z_{\alpha/2} = z_{0.025} = 1.96$.

NOTES

Step 7 : Decision

Since H_1 is a two-sided alternative, elements of the critical region are determined by the rejection rule $|z_0| \geq z_e$. Thus, it is a two-tailed test. For the given sample information, the rejection rule holds *i.e.*, $|z_0| = 3.33 > z_e = 1.96$. Hence, H_0 is rejected in favour of $H_1: \mu \neq 200$. Thus, the mean life span of the batteries is significantly different from 200 hours.

11.2.2 POPULATION VARIANCE UNKNOWN

Steps:

1. Let μ and σ^2 be respectively the mean and the variance of the population under study, where σ^2 is unknown. If μ_0 is an admissible value of μ , then frame the null hypothesis as $H_0: \mu = \mu_0$ and choose the suitable alternative hypothesis from

(i) $H_1: \mu \neq \mu_0$ (ii) $H_1: \mu > \mu_0$ (iii) $H_1: \mu < \mu_0$

2. Let (X_1, X_2, \dots, X_n) be a random sample of n observations drawn from the population, where n is large ($n \geq 30$).
3. Specify the level of significance, α .
4. Consider the test statistic $Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ under H_0 , where \bar{X} and S are the sample mean and sample standard deviation respectively. It may be noted that the above test statistic is obtained from Z by substituting S for σ .

The approximate sampling distribution of the test statistic under H_0 is the $N(0,1)$ distribution.

5. Calculate the value of Z for the given sample (x_1, x_2, \dots, x_n) as $Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$. Here, \bar{X} and s are respectively the values of \bar{X} and S calculated for the given sample.
6. Find the critical value, z_e , corresponding to α and H_1 from the following table

| | | | |
|----------------------------------|------------------|---------------|---------------|
| Alternative Hypothesis (H_1) | $\mu \neq \mu_0$ | $\mu > \mu_0$ | $\mu < \mu_0$ |
| Critical Value (z_e) | $z_{\alpha/2}$ | z_α | $-z_\alpha$ |

7. Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1 .

NOTES

| | | | |
|----------------------------------|---------------------------|--------------------|---------------------|
| Alternative Hypothesis (H_1) | $\mu \neq \mu_0$ | $\mu > \mu_0$ | $\mu < \mu_0$ |
| Rejection Rule | $ z_0 \geq z_{\alpha/2}$ | $z_0 > z_{\alpha}$ | $z_0 < -z_{\alpha}$ |

Example:

A car manufacturing company desires to introduce a new model car . The company claims that the mean fuel consumption of its new model car is lower than that of the existing model of the car, which is 57 kms/litre. A sample of 100 cars of the new model car is selected randomly and their fuel consumptions are observed. It is found that the mean fuel consumption of the 100 new model car is 60 kms/litre with a standard deviation of 3 kms/litre. Test the claim of the company at 5% level of significance.

Solution:

Step 1 : Let the fuel consumption of the new model car be assumed to be distributed according to a distribution with mean and standard deviation respectively μ and σ . The null and alternative hypotheses are

Null hypothesis H_0 : $\mu = 57$

i.e., the average fuel consumption of the company's new model car is not significantly different from that of the existing model.

Alternative hypothesis H_1 : $\mu > 57$

i.e., the average fuel consumption of the company's new model car is significantly lower than that of the existing model. In other words, the number of kms by the new model car is significantly more than that of the existing model car.

Step 2 : Data:

The given sample information are

Size of the sample (n) = 100. Hence, it is a large sample.

Sample mean (\bar{x}) = 60

Sample standard deviation(s) = 3

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic under H_0 is

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Since n is large, the sampling distribution of Z under H_0 is the $N(0,1)$ distribution.

Step 5 : Calculation of Test Statistic

The value of Z for the given sample information is calculated from

$$Z_0 = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{60 - 57}{3/\sqrt{100}} = 10$$

Step 6 : Critical Value

Since H_1 is a one-sided (right) alternative hypothesis, the critical value at $\alpha = 0.05$ is $z_{\alpha} = z_{0.05} = 1.645$.

Step 7 : Decision

Since H_1 is a one-sided (right) alternative, elements of the critical region are defined by the rejection rule $z_0 > z_{\alpha} = z_{0.05}$. Thus, it is a right-tailed test. Since, for the given sample information, $z_0 = 10 > z_{\alpha} = 1.645$, H_0 is rejected.

11.3 DIFFERENCE BETWEEN MEANS OF TWO POPULATIONS

11.3.1 POPULATION VARIANCE KNOWN

Steps:

1. Let μ_x and σ_x^2 be respectively the mean and the variance of Population -1. Also, let μ_y and σ_y^2 be respectively the mean and the variance of Population -2 under study. Here σ_x^2 and σ_y^2 are known admissible values.

Frame the null hypothesis as $H_0: \mu_x = \mu_y$ and choose the suitable alternative hypothesis from

- (i) $H_1: \mu_x \neq \mu_y$ (ii) $H_1: \mu_x > \mu_y$ (iii) $H_1: \mu_x < \mu_y$
2. Let (X_1, X_2, \dots, X_m) be a random sample of m observations drawn from Population-1 and (Y_1, Y_2, \dots, Y_n) be a random sample of n observations drawn from Population-2, where m and n are large (i.e., $m \geq 30$ and $n \geq 30$). Further, these two samples are assumed to be independent.
 3. Specify the level of significance, α .

NOTES

4. Consider the test statistic $Z = \frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}}}$ under H_0 , where \bar{X} and \bar{Y} are respectively the means of the two samples described in Step-2.

The approximate sampling distribution of the test statistic $Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}}}$ under H_0 (i.e., $\mu_X = \mu_Y$) is the $N(0,1)$ distribution.

It may be noted that the test statistic, when $\sigma^2_X = \sigma^2_Y = \sigma^2$, is $Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{1}{m} + \frac{1}{n}}}$

5. Calculate the value of Z for the given samples (x_1, x_2, \dots, x_m) and (y_1, y_2, \dots, y_n) as $Z_0 = \frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}}}$.

Here, \bar{x} and \bar{y} are respectively the values of \bar{X} and \bar{Y} for the given samples.

6. Find the critical value, z_e , corresponding to α and H_1 from the following table

| | | | |
|----------------------------------|--------------------|-----------------|-----------------|
| Alternative Hypothesis (H_1) | $\mu_X \neq \mu_Y$ | $\mu_X > \mu_Y$ | $\mu_X < \mu_Y$ |
| Critical Value (z_e) | $z_{\alpha/2}$ | z_α | $-z_\alpha$ |

7. Make decision on H_0 choosing the suitable rejection rule from the following table corresponding to H_1 .

| | | | |
|----------------------------------|---------------------------|------------------|-------------------|
| Alternative Hypothesis (H_1) | $\mu_X \neq \mu_Y$ | $\mu_X > \mu_Y$ | $\mu_X < \mu_Y$ |
| Rejection Rule | $ z_0 \geq z_{\alpha/2}$ | $z_0 > z_\alpha$ | $z_0 < -z_\alpha$ |

Example:

Performance of students in a national level Olympiad exam was studied. The scores secured by randomly selected students from two districts, viz., D_1 and D_2 of a State were analyzed. The number of students randomly selected from D_1 and D_2 are respectively 1000 and 1600. Average scores secured by the students selected from D_1 and D_2 are respectively 116 and 114. Can the samples be regarded as drawn from the identical populations having common standard deviation 27 Test at 5% level of significance.

Solution:

Step 1 : Let μ_X and μ_Y be respectively the mean scores secured in the national level Olympiad examination by all the students from the districts $D1$ and $D2$ considered for the study. It is given that the populations of the scores of the students of these districts have the common standard deviation $\sigma = 2$. The null and alternative hypotheses are

Null hypothesis: $H_0: \mu_X = \mu_Y$

i.e., average scores secured by the students from the study districts are not significantly different.

Alternative hypothesis: $H_1: \mu_X \neq \mu_Y$

i.e., average scores secured by the students from the study districts are significantly different. It is a two-sided alternative.

Step 2 : Data

The given sample information are

Size of the Sample-1 (m) = 1000

Size of the Sample-2 (n) = 1600. Hence, both the samples are large.

Mean of Sample-1 (\bar{x}) = 116

Mean of Sample-2 (y) = 114

Step 3 : Level of significance

$$\alpha = 5\%$$

Step 4 : Test statistic

The test statistic under the null hypothesis H_0 is $Z = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{1}{m} + \frac{1}{n}}}$

Since both m and n are large, the sampling distribution of Z under H_0 is the $N(0, 1)$ distribution.

Step 5 : Calculation of Test Statistic

The value of Z is calculated for the given sample information from

NOTES

$$Z = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{(116 - 114)}{\sqrt{\frac{1}{1000} + \frac{1}{1600}}} = 49.628$$

Step-6 : Critical value

Since H_1 is a two-sided alternative hypothesis, the critical value at $\alpha = 0.05$ is $z_e = z_{0.025} = 1.96$.

Step-7 : Decision

Since H_1 is a two-sided alternative, elements of the critical region are defined by the rejection rule $|z_0| \geq z_e = z_{0.025}$. For the given sample information, $|z_0| = 49.628 > z_e = 1.96$. It indicates that the given sample contains sufficient evidence to reject H_0 . Thus, it may be decided that H_0 is rejected. Therefore, the average performance of the students in the districts D_1 and D_2 in the national level Olympiad examination are significantly different. Thus the given samples are not drawn from identical populations.

11.3.2 POPULATION VARIANCE UNKNOWN

Steps:

1. Let μ_x and σ_x^2 be respectively the mean and the variance of Population -1. Also, let μ_y and σ_y^2 be respectively the mean and the variance of Population -2 under study. Here σ_x^2 and σ_y^2 are known admissible values.

Frame the null hypothesis as $H_0: \mu_x = \mu_y$ and choose the suitable alternative hypothesis from

- i. (i) $H_1: \mu_x \neq \mu_y$ (ii) $H_1: \mu_x > \mu_y$ (iii) $H_1: \mu_x < \mu_y$

2. Let (X_1, X_2, \dots, X_m) be a random sample of m observations drawn from Population-1 and (Y_1, Y_2, \dots, Y_n) be a random sample of n observations drawn from Population-2, where m and n are large (i.e., $m \geq 30$ and $n \geq 30$). Further, these two samples are assumed to be independent.

3. Specify the level of significance, α .

4. Consider the test statistic $Z = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{m} + \frac{S_y^2}{n}}}$ under H_0 ,

5. (i.e., $\mu_x = \mu_y$)

i.e, the above test statistics is obtained from Z considered in the test described by substituting S^2x and S^2y respectively for σ^2_x and σ^2_y

The approximate sampling distribution of the test statistic

$$Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{S^2x}{m} + \frac{S^2y}{n}}} \text{ under } H_0 \text{ is the } N(0,1) \text{ distribution.}$$

6. Calculate the value of Z for the given samples (x_1, x_2, \dots, x_m) and

$$(y_1, y_2, \dots, y_n) \text{ as } Z_0 = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{S^2x}{m} + \frac{S^2y}{n}}}.$$

Here, \bar{x} and \bar{y} are respectively the values of \bar{X} and \bar{Y} for the given samples.

Also, s^2x and s^2y are respectively the values of S^2x and S^2y for the given samples.

7. Find the critical value, z_e , corresponding to α and H_1 from the following table

| | | | |
|----------------------------------|--------------------|-----------------|-----------------|
| Alternative Hypothesis (H_1) | $\mu_x \neq \mu_y$ | $\mu_x > \mu_y$ | $\mu_x < \mu_y$ |
| Critical Value (z_e) | $z_{\alpha/2}$ | z_α | $-z_\alpha$ |

8. Make decision on H_0 choosing the suitable rejection rule from the following table corresponding to H_1 .

| | | | |
|----------------------------------|---------------------------|------------------|-------------------|
| Alternative Hypothesis (H_1) | $\mu_x \neq \mu_y$ | $\mu_x > \mu_y$ | $\mu_x < \mu_y$ |
| Rejection Rule | $ z_0 \geq z_{\alpha/2}$ | $z_0 > z_\alpha$ | $z_0 < -z_\alpha$ |

11.4 TEST OF HYPOTHESES FOR POPULATION PROPORTION

Steps:

- Let P denote the proportion of the population possessing the qualitative characteristic (attribute) under study. If p_0 is an admissible value of P, then frame the null hypothesis as $H_0: P = p_0$ and choose the suitable alternative hypothesis from
 - $H_1: P \neq p_0$
 - $H_1: P > p_0$
 - $H_1: P < p_0$
- Let p be proportion of the sample observations possessing the attribute, where n is large, $np > 5$ and $n(1 - p) > 5$.

NOTES

- Specify the level of significance, α .
- Consider the test statistic under H_0 . Here, $\frac{p-P}{\sqrt{\frac{PQ}{n}}}$, Here, $Q = 1 - P$.

The approximate sampling distribution of the test statistic under H_0 is the $N(0, 1)$ distribution.

- Calculate the value of Z under H_0 for the given data as, $\frac{p-P}{\sqrt{\frac{p_0q_0}{n}}}$, $q_0 = 1 - p_0$.
- Choose the critical value, z_e , corresponding to α and H_1 from the following table

| | | | |
|----------------------------------|----------------|------------|-------------|
| Alternative Hypothesis (H_1) | $P \neq p_0$ | $P > p_0$ | $P < p_0$ |
| Critical Value (z_e) | $z_{\alpha/2}$ | z_α | $-z_\alpha$ |

- Make decision on H_0 choosing the suitable rejection rule from the following table corresponding to H_1 .

| | | | |
|----------------------------------|---------------------------|------------------|-------------------|
| Alternative Hypothesis (H_1) | $P \neq p_0$ | $P > p_0$ | $P < p_0$ |
| Rejection Rule | $ z_0 \geq z_{\alpha/2}$ | $z_0 > z_\alpha$ | $z_0 < -z_\alpha$ |

Example:

A survey was conducted among the students of a city to study their preference towards consumption of chocolate and ice-cream. Among 2000 randomly selected students, it is found that 1120 are chocolate and the remaining are ice-cream. Can we conclude at 1% level of significance from this information that both chocolate and ice-cream are equally preferred among the students in the city?

Solution:

Step 1 : Let P denote the proportion of students in the city who preferred to have chocolate. Then, the null and the alternative hypotheses are

Null hypothesis: $H_0 : = 0.5$

i.e., it is significant that both chocolate and ice-cream are preferred equally in the city.

Alternative hypothesis: $H_0 : \neq 0.5$

i.e., preference of chocolate and ice-cream are not

significantly equal. It is a two-sided alternative hypothesis.

Step 2 : Data

The given sample information are

Sample size (n) = 2000. Hence, it is a large sample.

No. of chocolate consumer = 1120

Sample proportion (p) = $\frac{1120}{2000} = 0.56$

Step 3 : Level of significance

$$\alpha = 1\%$$

Step 4 : Test statistic

Since n is large, $np = 1120 > 5$ and $n(1 - p) = 880 > 5$, the test statistic under the null hypothesis, is $Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}}$

Its sampling distribution under H_0 is the $N(0, 1)$ distribution.

Step 5 : Calculation of Test Statistic

The value of Z can be calculated for the sample information from

$$Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.56-0.50}{\sqrt{\frac{0.5 \times 0.5}{2000}}} = 5.3763$$

Step 6 : Critical value

Since H_1 is a two-sided alternative hypothesis, the critical value at 1% level of significance is $z_{\alpha/2} = z_{0.005} = 2.58$.

Step 7 : Decision

Since H_1 is a two-sided alternative, elements of the critical region are determined by the rejection rule $|z_0| \geq z_e$. Thus it is a two-tailed test. Since $|z_0| = 5.3763 > z_e = 2.58$, reject H_0 at 1% level of significance. Therefore, there is significant evidence to conclude that the preference of chocolate and ice-cream are different.

11.5 DIFFERENCE BETWEEN TWO PROPORTIONS

Steps:

- 1: Let P_X and P_Y denote respectively the proportions of Population-1 and Population-2 possessing the qualitative characteristic (attribute) under study. Frame the null hypothesis as H_0 :

$P_X = P_Y$ and choose the suitable alternative hypothesis from

Test of Hypothesis

NOTES

- (i) $H_1: P_X \neq P_Y$ (ii) $H_1: P_X > P_Y$ (iii) $H_1: P_X < P_Y$

2: Let P_X and P_Y denote respectively the proportions of the samples of sizes m and n drawn from Population-1 and Population-2 possessing the attribute, where m and n are large (i.e., $m \geq 30$ and $n \geq 30$). Also, $mp_x > 5$, $m(1 - p_x) > 5$, $np_y > 5$, $n(1 - p_y) > 5$. Here, these two samples are assumed to be independent.

3: Specify the level of significance, α .

4: Consider the test statistic $Z = \frac{(p_x - p_y) - (P_X - P_Y)}{\sqrt{\hat{p}\hat{q}(\frac{1}{m} + \frac{1}{n})}}$ under H_0 . Here

$\hat{p} = \frac{mp_x + np_y}{m+n}$, $\hat{q} = 1 - \hat{p}$. The approximate sampling distribution of the test statistic

under H_0 is $N(0,1)$ distribution.

5: Calculate the value of Z for the given data as $Z = \frac{(P_X - P_Y)}{\sqrt{\hat{p}\hat{q}(\frac{1}{m} + \frac{1}{n})}}$

6: Choose the critical value, z_e , corresponding to α and H_1 from the following table

| | | | |
|----------------------------------|----------------|--------------|---------------|
| Alternative Hypothesis (H_1) | $P_X \neq P_Y$ | $P_X > P_Y$ | $P_X < P_Y$ |
| Critical Value (z_e) | $z_{\alpha/2}$ | z_{α} | $-z_{\alpha}$ |

7: Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1

| | | | |
|----------------------------------|---------------------------|--------------------|---------------------|
| Alternative Hypothesis (H_1) | $P_X \neq P_Y$ | $P_X > P_Y$ | $P_X < P_Y$ |
| Rejection Rule | $ z_0 \geq z_{\alpha/2}$ | $z_0 > z_{\alpha}$ | $z_0 < -z_{\alpha}$ |

Example:

A study was conducted to investigate the interest of students in private schools. Among randomly selected 1000 students from City-1, 800 persons were found to be private school. From City-2, 1600 persons were selected randomly and among them 1200 students are from private school. Do the data indicate that the two cities are significantly different

with respect to prevalence of private school among the students? Choose the level of significance as $\alpha = 0.05$.

Solution:

Step1 : Let P_X and P_Y be respectively the proportions of private school students in City-1 and City-2. Then, the null and alternative hypotheses are

Null hypothesis: $H_0: P_X = P_Y$

i.e., there is no significant difference between the proportions of private school students in City-1 and City-2.

Alternative hypothesis: $H_1: P_X \neq P_Y$

i.e., difference between the proportions of private school students in City-1 and City-2 is significant. It is a two-sided alternative hypothesis.

Step 2 : Data

The given sample information are

| City | Sample size | Sample proportion |
|--------|-------------|----------------------------|
| City 1 | $m = 1000$ | $P_X = 800 / 1000 = 0.80$ |
| City 2 | $n = 1600$ | $P_Y = 1200 / 1600 = 0.75$ |

Here $m \geq 30$ and $n \geq 30$, $mp_x = 800 > 5$,
 $m(1 - p_x) = 200 > 5$, $np_y = 1200 > 5$, $n(1 - p_y) = 400 > 5$.

Step 3 : Level of significance

$$\alpha = 5\%$$

Step 4 : Test statistic

The test statistic under the null hypothesis is

$$Z = \frac{(p_x - p_y) - (P_X - P_Y)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m} + \frac{1}{n}\right)}} \quad \text{where } \hat{p} = \frac{mp_x + np_y}{m+n}, \hat{q} = 1 - \hat{p}$$

Step 5 : Calculation of Test Statistic

The value of Z for given sample information is calculated from

$$Z = \frac{(P_X - P_Y)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m} + \frac{1}{n}\right)}} = \frac{(0.80 - 0.75)}{\sqrt{(0.77)(0.23)\left(\frac{1}{1000} + \frac{1}{1600}\right)}} = 2.0764$$

Step 6 : Critical value

Since H_1 is a two-sided alternative hypothesis, the critical

value at 5% level of significance is $z_e = 1.96$.

Test of Hypothesis

NOTES

Step 7 : Decision

Since H_0 is a two-sided alternative, elements of the critical region are determined by the rejection rule $|z_0| > z_e$. Thus, it is a two-tailed test. For the given sample information, $z_e = 2.0764 > 1.96$. Hence, H_0 is rejected. We can conclude that the difference between the proportions of private school students in City-1 and City-2 is significant.

CHECK YOUR PROGRESS - 1

1. What is hypothesis testing?
2. A large sample theory is applicable when
3. When is standard error of the sample proportion under H_0

11.6 SUMMARY

- If the number of sample observations is greater than or equal to 30, it is called large sample.
- For hypotheses testing based on two samples, if the sizes of both the samples are greater than or equal to 30, they are called large samples.
- For testing population proportion, the sampling distribution of the test statistic is $N(0, 1)$, only when $n \geq 30$, $np > 5$ and $n(1 - p) > 5$.
- For testing equality of two population proportions, the sampling distribution of the test statistic is $N(0, 1)$ only when $m \geq 30$, $n \geq 30$, $mp > 5$, $m(1 - p) > 5$, $np > 5$ and $n(1 - p) > 5$.

11.7 KEY WORDS

Hypothesis, Population, Proportion

11.8 ANSWER TO CHECK YOUR PROGRESS

1. Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data
2. When $n \geq 30$
3. $\sqrt{\frac{PQ}{n}}$

11.9 QUESTION AND EXERCISE

SHORT ANSWER QUESTIONS

1. List the possible alternative hypotheses and the corresponding rejection rules followed in testing equality of two population means.
2. Specify the alternative hypotheses and the rejection rules prescribed for testing equality of two population proportions

LONG ANSWER QUESTIONS

1. Explain the general procedure to be followed for testing of hypotheses.
2. Explain the procedure for testing hypotheses for population mean, when the population variance is unknown.
3. How will you formulate the hypotheses for testing equality of means of two populations, when the population variances are known? Describe the method.
4. Describe the procedure for testing hypotheses concerning equality of means of two populations, assuming that the population variances are unknown.
5. Give a detailed account on testing hypotheses for population proportion.
6. Explain the procedure of testing hypotheses for equality of proportion of two populations. Interest of XII Students on Residential Schooling was investigated among randomly selected students from two regions. Among 300 students selected from Region A, 34 students expressed their interest. Among 200 students selected from Region B, 28 students expressed their interest. Does this information provide sufficient evidence to conclude at 5% level of significance that students in Region A are more interested in Residential Schooling than the students in Region B?

11.10 FURTHER READINGS

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. SahityaBhawanPublishersand Distributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills PublishingCompany Ltd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata

Test of Hypothesis

NOTES

- McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
 5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
 6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
 7. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.

UNIT 12 - CHI – SQUARE TEST

Chi –Square Test

NOTES

Structure

- 12.0 Introduction
- 12.1 Objectives
- 12.2 Characteristics of Chi –Square Test
- 12.3 Uses of Chi –Square Test
- 12.4 Steps of Chi –Square Test
- 12.5 Analysis of Variance (ANOVA)
- 12.6 Assumptions in Analysis of Variance
- 12.7. Basic steps in Analysis of Variance
 - 12.7.1 One Way ANOVA
 - 12.7.2 Two Way ANOVA
- 12.8 Summary
- 12.9 Key Words
- 12.10 Answer to Check Your Progress
- 12.11 Questions and Exercise
- 12.12 Further Readings

12.0 INTRODUCTION

A chi-squared test, also written as χ^2 test, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other qualification, 'chi-squared test' often is used as short for Pearson's chi-squared test.

The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.

Chi square test is applied in statistics to test the goodness of fit to verify the distribution of observed data with assumed theoretical distribution. Therefore, it is a measure to study the divergence of actual and expected frequencies. It has great use in statistics, specially in sampling studies, where we expect a doubted coincidence between actual and expected frequencies, and the extent to which the difference can be ignored, because of fluctuations in sampling.

12.1 OBJECTIVES

The student will be able to

- Understand the purpose for using chi-square test
- Understand the procedures for Analysis of variance
- Understand the characteristics and of chi-square test
- Solve problems to test the hypothesis whether the population has a particular variance using chi-square test

12.2 CHARACTERISTICS OF χ^2 TEST

1. Test is based on events or frequencies, whereas in theoretical

Self-Instructional Material

Chi –Square Test

NOTES

- distribution, the test is based on mean and standard deviation.
2. To draw inferences, this test is applied, specially testing the hypothesis but not useful for estimation.
 3. The test can be used between the entire set of observed and expected frequencies.
 4. For every increase in the number of degree of freedom, a new χ^2 distribution is formed.
 5. It is a general purpose test and as such is highly useful in research.

12.3 USES OF χ^2 TEST

χ^2 Test of goodness of fit

Through the test we can find out the deviations between the observed values and expected values. Here we are not concerned with the parameters but concerned with the form of distribution. Karl Pearson has developed a method to test the difference between the theoretical value (hypothesis) and the observed value. A Greek letter χ^2 is used to describe the magnitude of difference between the fact and theory.

The χ^2 may be defined as,

$$\chi^2 = \frac{O-E^2}{E}$$

O = Observed Frequencies

E = Expected Frequencies

12.4 STEPS OF χ^2 TEST

1. A hypothesis is established along with the significance level.
2. Compute deviation between observed value and expected value (O-E).
3. Square the deviations calculated (O-E)².
4. Divide the (O-E)² by its expected frequency.
5. Add all the values obtained in step 4.
6. Find the value of χ^2 table at certain level of significance, usually 5% level.

If the calculated value of χ^2 is greater than the table value of χ^2 , at certain level of significance, we reject the hypothesis. If the computed value of χ^2 is less than the table value, at a certain degree of level of significance, it is said to be non-significant. This implies that the discrepancy between the observed frequencies may be due to fluctuations in the simple sampling.

Example:

In a certain sample of 2000 families, 1400 families are consumers of tea. Out of 1800 Hindu families, 1236 families consume tea. Use χ^2 test and state whether there is any significant difference between consumption of tea among Hindu and non-Hindu families.

| | Hindu | Non – Hindu | Total |
|---------------------|-------|-------------|-------|
| Consuming Tea | 1236 | 164 | 1400 |
| Non – Consuming Tea | 564 | 36 | 600 |
| Total | 1800 | 200 | 2000 |

Solution

On tabulation of the information in a 2x2 contingency table, we get:

Observed Frequencies

| | Hindu | Non Hindu | Total |
|---------------------|-------|-----------|-------|
| Consuming Tea | 1236 | 164 | 1400 |
| Non – Consuming Tea | 564 | 36 | 600 |
| Total | 1800 | 200 | 2000 |

Expected Frequencies

| | Hindu | Non Hindu | Total |
|---------------------|-------|-----------|-------|
| Consuming Tea | 1260 | 140 | 1400 |
| Non – Consuming Tea | 540 | 60 | 600 |
| Total | 1800 | 200 | 2000 |

Calculation of χ^2

| O | E | O – E | (O-E) ² | (O-E) ² / E |
|------|------|-------|--------------------|------------------------|
| 1236 | 1260 | -24 | 576 | 0.457 |
| 564 | 540 | 24 | 576 | 1.068 |
| 164 | 140 | 24-24 | 576 | 4.114 |
| 36 | 60 | | 576 | 9.600 |
| | | | | $\sum(O-E)^2/E=15.239$ |

d.f is 1, Table value of $r_{2, 0.05}$ for 1 d.f = 3.841.

For a contingency table, 2x2 table, the degree of freedom is

$$V = (c-1)(r-1) = (2-1)(2-1) = 1.$$

The calculated value of χ^2 15.239 is higher than the table value i.e., 3.841; therefore the null hypothesis is rejected.

Hence, the two communities differ significantly as far as consumption of a tea is concerned.

12.5 ANALYSIS OF VARIANCE (ANOVA)

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher. The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation.

Chi –Square Test

NOTES

In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means.

12.6 ASSUMPTIONS IN ANALYSIS OF VARIANCE

1. Each of the samples is strained from a normal distribution.
2. The variances for the population from which samples have been drained are equal.
3. The variation of each value around its own grand mean should be independent for each value.

12.7 BASIC STEPS IN ANALYSIS OF VARIANCE

Determine

1. One estimate of the population variance from the variance among the sample means.
2. Determine a second estimate of the population variance from the variance within the sample.
3. Compare these two estimates if they are approximately equal in value, accept the null hypothesis.

12.7.1 One-Way Anova

In statistics, one-way analysis of variance (abbreviated one-way ANOVA) is a technique that can be used to compare means of two or more samples (using the F distribution). This technique can be used only for numerical response data, the "Y", usually one variable, and numerical or (usually) categorical input data, the "X", always one variable, hence "one-way".

The ANOVA tests the null hypothesis that samples in all groups are drawn from populations with the same mean values. To do this, two estimates are made of the population variance. These estimates rely on various assumptions.

The ANOVA produces an F-statistic, the ratio of the variance calculated among the means to the variance within the samples. When there are only two means to compare, the t-test and the F-test are equivalent; the relation between ANOVA and t is given by $F = t^2$. An extension of one-way ANOVA is two-way analysis of variance that examines the influence of two different categorical independent variables on one dependent variable.

Example:

In order to determine whether there is significant difference in the durability of 3 makes of computers, samples of size 5 are selected from each make and the frequency of repair during the 1st year of purchase is observed. The results are as follows:

| Makes | | |
|-------|----|-----|
| I | II | III |
| 4 | 7 | 6 |
| 6 | 9 | 4 |

| | | |
|---|----|---|
| 8 | 11 | 6 |
| 9 | 12 | 3 |
| 7 | 5 | 2 |

In view of the above data, what conclusion can you draw?

Solution:

Null Hypothesis H_0 = there is no significant difference in the durability of 3 makes of computers.

| Computer I | | Computer II | | Computer III | |
|---------------|------------------|---------------|------------------|---------------|------------------|
| X_1 | X_1^2 | X_2 | X_2^2 | X_3 | X_3^2 |
| 4 | 16 | 7 | 49 | 6 | 36 |
| 6 | 36 | 9 | 81 | 4 | 16 |
| 8 | 64 | 11 | 121 | 6 | 36 |
| 9 | 81 | 12 | 144 | 3 | 9 |
| 7 | 49 | 5 | 25 | 2 | 4 |
| $\sum X_1=34$ | $\sum X_1^2=246$ | $\sum X_2=44$ | $\sum X_2^2=420$ | $\sum X_3=21$ | $\sum X_3^2=101$ |

Step – 1

$$\begin{aligned} \text{Sum of all items (T)} &= \sum X_1 + \sum X_2 + \sum X_3 \\ &= 34 + 44 + 21 \\ &= 99 \end{aligned}$$

Step – 2

$$\text{Correction factor (C.F)} = \frac{T^2}{N} = \frac{(99)^2}{15} = 653.4$$

Step – 3

$$\begin{aligned} \text{TSS} &= \text{Sum of Squares of all the items} - \text{C.F} \\ &= \sum X_1^2 + \sum X_2^2 + \sum X_3^2 - \frac{T^2}{N} \\ &= 246 + 420 + 101 - 653.4 = 113.6 \end{aligned}$$

Step – 4

$$\begin{aligned} \text{SSC} &= \text{Sum of Squares between samples} - \text{C.F} \\ &= \frac{(\sum X_1)^2}{n} + \frac{(\sum X_2)^2}{n} + \frac{(\sum X_3)^2}{n} - \text{C.F} \\ &= \frac{(34)^2}{5} + \frac{(44)^2}{5} + \frac{(21)^2}{5} - 653.4 \\ &= 231.2 + 387.2 + 88.5 - 653.4 = 53.5 \end{aligned}$$

Step – 5

$$\begin{aligned} \text{MSC} &= \frac{\text{Sum of squares between samples}}{\text{d.f}} \\ &= \frac{53.5}{2} \end{aligned}$$

Chi –Square Test

NOTES

Chi –Square Test

NOTES

$$= 26.75$$

Step – 6

SSE = Total sum of squares – Sum of Squares between samples

$$= 113.6 - 53.5$$

$$= 60.1$$

Step – 7

MSE = $\frac{\text{Sum of squares within samples}}{\text{d.f}}$

$$= \frac{60.1}{12}$$

$$= 5.00$$

ANOVA TABLE

| Source of variations | Sum of squares | Degrees of freedom | Mean Squares | F - ratio |
|----------------------|----------------|--------------------|---|-----------------------------------|
| Between samples | SSC = 53.5 | 3-1=2 | MSC = $\frac{SSC}{\text{d.f}}$ = 26.75 | $F_c = \frac{MSC}{MSE}$ = 5.35 |
| Within samples | SSE = 60.1 | 15-3=12 | MSE = $\frac{SSE}{\text{d.f}}$ = 5.00 | |

Tabulated value of F for $V_1=2$ and $V_2=12$ at 5% level of significance is 3.88. $F_{\text{Tab}}=3.88$. Calculated value of F is $F_c = 5.35$. Since $F_c > F_{\text{Tab}}$. We reject the null hypothesis H_0 . There is significant difference in the durability of 3 makes of computers.

12.7.2 TWO-WAY ANOVA

The two-way ANOVA compares the mean differences between groups that have been split on two independent variables (called factors). The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable.

This is a term which stems from agricultural research in which several variables or treatments are applied to different blocks of land for repetition or replication of the experimental effects. The advantages of a completely randomized experimental design are as follows.

- Easy to lay out.
- Allows flexibility.
- Simple statistical analysis.
- The lot of information due to missing data is smaller than with any other design.

But this design is usually suited (i) only for small number of treatments and (ii) for homogeneous experimental material.

Example:

There varieties A, B, C of crop are tested in a randomized block design with four replications. The plot yields in pounds are as follows

| Varieties | Yields | | | |
|-----------|--------|---|----|---|
| | 1 | 2 | 3 | 4 |
| A | 6 | 5 | 8 | 9 |
| B | 8 | 4 | 6 | 9 |
| C | 7 | 6 | 10 | 6 |

Solution

Null hypothesis H_0 : There is no significant difference between varieties (rows) and between yields,(blocks).

| Varieties | Yields | | | | |
|-----------|--------|----|----|----|-------|
| | 1 | 2 | 3 | 4 | Total |
| A | 6 | 4 | 6 | 6 | 24 |
| B | 7 | 5 | 8 | 9 | 28 |
| C | 8 | 6 | 10 | 9 | 32 |
| Total | 21 | 15 | 24 | 24 | 84 |

Step -1

Grand total (T) = 84

Step - 2

Correction factor (C.F) = $\frac{T^2}{N} = \frac{(84)^2}{12} = 588$

Step - 3

SSC

$$\begin{aligned}
 &= \text{Sum of squares between blocks (columns)} \\
 &= \frac{(21)^2}{3} + \frac{(15)^2}{3} + \frac{(24)^2}{3} + \frac{(24)^2}{3} - C.F \\
 &= 606 - 588 \\
 &= 18
 \end{aligned}$$

Step - 4

SSR

$$\begin{aligned}
 &= \text{Sum of squares between varieties (Rows)} \\
 &= \frac{(24)^2}{4} + \frac{(28)^2}{4} + \frac{(32)^2}{4} - C.F \\
 &= 596 - 588 \\
 &= 8
 \end{aligned}$$

Step - 5

TSS

$$\begin{aligned}
 &= \text{Total sum of squares} - C.F \\
 &= [(6)^2+(7)^2+(8)^2+(4)^2+(6)^2+(5)^2+(8)^2+(6)^2+(10)^2+(6)^2+(9)^2+(9)^2] - 588 \\
 &= 624 - 588 \\
 &= 36
 \end{aligned}$$

Step - 6

SSE

$$\begin{aligned}
 &= \text{Residual sum of squares} \\
 &= TSS-(SSC+SSR) \\
 &= 36 - (18+8) = 10
 \end{aligned}$$

Chi –Square Test

NOTES

Chi –Square Test

NOTES

Step- 7

$$\begin{aligned} \text{d.f} &= v_3 &&= (c-1)(r-1) \\ &&&= (3)(2) \\ &&&= 6 \end{aligned}$$

ANOVA TABLE

| Source of variation | Sum of squares | Degree of freedom | Mean Squares | F-ratio |
|--------------------------|----------------|-------------------|---|----------------------------------|
| Between Blocks (Columns) | SSC = 18 | c-1 4-1= 3 | MSC= $\frac{SSC}{\text{d.f}}$ = 6 | $F_c = \frac{MSC}{MSE}$ = 3.6 |
| Between Varieties (Rows) | SSR=8 | r-1 3-1=2 | MSR= $\frac{SSR}{\text{d.f}}$ = 4 | $F_R = \frac{MSR}{MSE}$ = 2.4 |
| Residual | SSE=10 | (r-1)(c-1) = 6 | MSE= $\frac{SSE}{\text{d.f}}$ =1.667 | |

- (i) The tabulated value of F for (3,6) d.f at 5 % level of significance is 4.76. $F_{\text{tab}}=4.76$. since $F_c < F_{\text{tab}}$, we accept the null hypothesis H_0 . That is there is no significant difference between yields.
- (ii) The tabulated value of F for (2,6) d.f at 5 % level of significance is 5.14. $F_{\text{tab}}=5.14$. since $F_R < F_{\text{tab}}$, we accept the null hypothesis H_0 . That is there is no significant difference between varieties.

CHECK YOUR PROGRESS - 1

1. By which other name is the Chi-Square goodness of fit test known?
2. What type of data do you need in Chi-Square test?
3. What symbol is used to represent Chi-Square?
4. What is Analysis of Variance?
5. What is the main purpose of Two way ANOVA test?
6. The variation of each value around its own grand mean should be _____ for each value

12.8 SUMMARY

- The uses of distribution are testing the specified variance of a normal population, testing goodness of fit and testing independence of attributes
- Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures used to analyze

the differences among group means in a sample

- One-way analysis of variance (abbreviated one-way ANOVA) is a technique that can be used to compare means of two or more samples
- The two-way ANOVA compares the mean differences between groups that have been split on two independent variables

12.9 KEY WORDS

Chi-square, Analysis of Variance, One way method, Two way method

12.10 ANSWER TO CHECK YOUR PROGRESS

1. One sample chi square
 2. Categorical
 3. χ^2
 4. ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the *t*-test beyond two means
 5. The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable
 6. Independent
-

12.11 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTION:

1. Define chi square test
2. What are the condition for the validity of chi square test
3. Five applications of chi square test
4. What is analysis of variance?
5. What are the assumptions of ANOVA

LONG ANSWER QUESTION:

1. Explain the steps of chi-square test
 2. Write down steps for testing the significance of goodness of fit
 3. Write the model ANOVA table for one way classification
 4. Compare one way and two way ANOVA
-

12.12 FURTHER READINGS

1. Spiegel, Murray R.: Theory and Practical of Statistics., London
2. McGraw Hill Book Company.
3. Yamane, T.: Statistics: An Introductory Analysis, New York, HarperedRow Publication
4. R.P. Hooda: Statistic for Economic and Management McMillan IndiaLtd.
5. G.C. Beri: Statistics for Mgt., TMA
6. J.K. Sharma: Business Statistics, Pearson Education

UNIT13 - PROBABILITY

Structure

- 13.0 Introduction
- 13.1 Objectives
- 13.2 Importance Terms
- 13.3 Types of Probability
- 13.4 Basic relationship of Probability
- 13.5 Addition Theorem of Probability
- 13.6 Multiplication Theorem of Probability
- 13.7. Condition Probability
 - 13.7.1 Combined Use Of Addition And Multiplication Theorem
- 13.8 Baye's Theorem and its application
- 13.9 Summary
- 13.10 Key Words
- 13.11 Answer to Check Your Progress
- 13.12 Questions and Exercise
- 13.13 Further Readings

13.0 INTRODUCTION

In our day to day life the “probability” or “chance” is very commonly used term. Sometimes, we use to say “Probably it may rain tomorrow”, “Probably Mr. X may come for taking his class today”, “Probably you are right”. All these terms, possibility and probability convey the same meaning. But in statistics probability has certain special connotation unlike in Layman's view.

The theory of probability has been developed in 17th century. It has got its origin from games, tossing coins, throwing a dice, drawing a card from a pack. In 1954 Antoine Gornband had taken an initiation and an interest for this area.

After him many authors in statistics had tried to remodel the idea given by the former. The “probability” has become one of the basic tools of statistics. Sometimes statistical analysis becomes paralyzed without the theorem of probability. “Probability of a given event is defined as the expected frequency of occurrence of the event among events of a like sort.” (Garrett)

The probability theory provides a means of getting an idea of the likelihood of occurrence of different events resulting from a random experiment in terms of quantitative measures ranging between zero and one. The probability is zero for an impossible event and one for an event

which is certain to occur.

13.1 OBJECTIVES

The students will be able to understand

- The important terms in probability
- Concept of conditional probability, addition theorem and multiplication theorem.
- Baye's theorem and its applications

13.2 IMPORTANT TERMS

1. **Probability or Chance:** Probability or chance is a common term used in day-to-day life. For example, we generally say, 'it may rain today'. This statement has a certain uncertainty. Probability is quantitative measure of the chance of occurrence of a particular event.
2. **Experiment:** An experiment is an operation which can produce well-defined outcomes.
3. **Random Experiment:** If all the possible outcomes of an experiment are known but the exact output cannot be predicted in advance, that experiment is called a random experiment.
Examples: Tossing of a fair coin: When we toss a coin, the outcome will be either Head (H) or Tail (T)
4. **Trial :** Any particular performance of a random experiment is called trial
Example: Tossing 4 coins, rolling a die, picking ball from a bag containing 10 balls of which 4 is red and 6 is blue.
5. **Event :** Any subset of a Sample Space is an event. Events are generally denoted by capital letters A, B , C, D etc.
Examples:
 - i. When a coin is tossed, outcome of getting head or tail is an event**Types of Events:**
 - **Simple Events:** In the case of simple events, we take the probability of occurrence of single events.
Examples: Probability of getting a Head (H) when a coin is tossed
 - **Compound Events:** In the case of compound events, we take the probability of joint occurrence of two or more events
Examples: When two coins are tossed, probability of getting a Head (H) in the first toss and getting a Tail (T) in the second toss..

Probability

NOTES

Self-Instructional

6. **Sample Space** :Sample Space is the set of all possible outcomes of an experiment. It is denoted by S.

Examples : When a coin is tossed, $S = \{H, T\}$ where H = Head and T = Tail

7. **Mutually Exclusive Events**: Two or more than two events are said to be mutually exclusive if the occurrence of one of the events excludes the occurrence of the other

Example :When a coin is tossed, we get either Head or Tail. Head and Tail cannot come simultaneously. Hence occurrence of Head and Tail are mutually exclusive events.

8. **Equally Likely Events**: Events are said to be equally likely if there is no preference for a particular event over the other.

Examples: When a coin is tossed, Head (H) or Tail is equally likely to occur.

9. **Independent Events**: Events can be said to be independent if the occurrence or non-occurrence of one event does not influence the occurrence or non-occurrence of the other.

Example:

- i. When a coin is tossed twice, the event of getting Tail(T) in the first toss and the event of getting Tail(T) in the second toss are independent events. This is because the occurrence of getting Tail(T) in any toss does not influence the occurrence of getting Tail(T) in the other toss.

10. **Exhaustive Events**: Exhaustive Event is the total number of all possible outcomes of an experiment.

Examples: When a coin is tossed, we get either Head or Tail. Hence there are 2 exhaustive events.

11. **Favorable Events**: The outcomes which make necessary the happening of an event in a trial are called favorable events.

Examples:if two dice are thrown, the number of favorable events of getting a sum 5 is four, i.e., (1, 4), (2, 3), (3, 2) and (4, 1).

13.3 TYPES OF PROBABILITY

1. Classical Approach (Priori Probability):

According to this approach, the probability is the ratio of favorable events to the total no. of equally likely events. In tossing a coin the probability of the coin coming down is 1, of the head coming up is $\frac{1}{2}$ and of the tail coming up is $\frac{1}{2}$.

The probability of one event as 'P' (success) and of the other event as 'q' (failure) as there is no third event.

Probability NOTES

$$p = \frac{\text{Number of favourable cases}}{\text{Total number of equally likely cases}}$$

If an event can occur in 'a' ways and fail to occur in 'b' ways and these are equally to occur, then the probability of the event occurring, $a/a+b$ is denoted by p. Such probabilities are known as unitary or theoretical or mathematical probability. p is the probability of the event happening and q is the probability of its not happening.

$$p = \frac{a}{a+b} \text{ and } q = \frac{b}{a+b}$$

$$\text{Hence } p+q = \frac{a+b}{a+b}$$

$$\text{Therefore } p+q = 1$$

Probabilities can be expressed either as ratio, fraction or percentage, such as $\frac{1}{2}$ or 0.5 or 50%. **Example:** Tossing of a coin.

Limitations:

- This definition is confined to the problems of games of chance only and can not explain the problem other than the games of chance.
- This method can not be applied, when the outcomes of a random experiment are not equally likely.
- The classical definition is applicable only when the events are mutually exclusive.

2. Relative Frequency Theory of Probability:

In this approach, the probability of happening of an event is determined on the basis of past experience or on the basis of relative frequency of success in the past.

Example: If a machine produces 100 articles in the past, 2 articles were found to be defective, and then the probability of the defective articles is $2/100$ or 2%.

The relative frequency obtained on the basis of past experience can be shown to come to very close to the classical probability.

Limitations:

- The experimental conditions may not remain essentially homogeneous and identical in a large number of repetitions of the experiment.
- The relative frequency m/n , may not attain a unique value no matter how large.

- Probability $p(A)$ defined can never be obtained in practice. We can only attempt at a close estimate of $p(A)$ by making N sufficiently large.

3. Subjective Approach :

The subjective approach is also known as subjective theory of probability. The probability of an event is considered as a measure of one's confidence in the occurrence of that particular event

This theory is commonly used in business decision making. The decision reflects the personality of the decision maker. Persons may arrive at different probability assignment because of differences in value at experience etc. The personality of the decision maker is reflected in a final decision. The decision under this theory is taken on the basis of the available data plus the effects of other factors many of which may be subjective in nature.

Example: A student would top in B. Com Exam this year.

A subjective would assign a weight between zero and one to this event according to his belief for its possible occurrence.

4. Axiomatic Approach:

The probability calculations are based on the axioms. The axiomatic probability includes the concept of both classical and empirical definitions of probability.

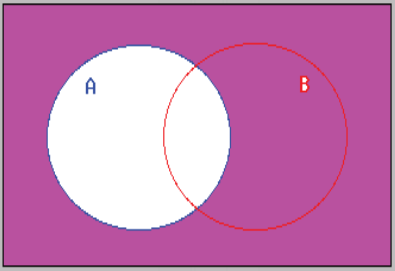
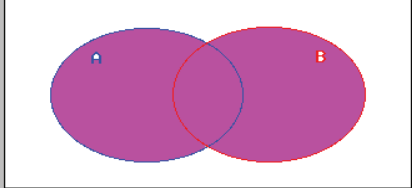
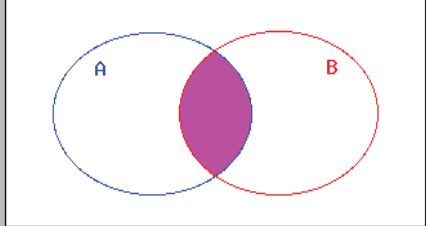
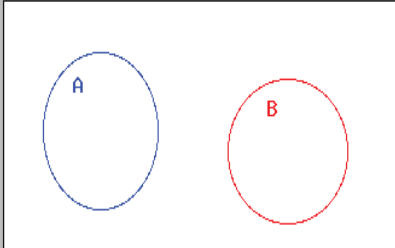
The approach assumes finite sample spaces and is based on the following three axioms:

- i) The probability of an event ranges from 0 to 1. If the event cannot take place its probability shall be '0' and if it is bound to occur its probability is '1'.
- ii) The probability of the entire sample space is 1, i.e. $p(S)=1$.
- iii) If A and B are mutually exclusive events then the probability of occurrence of either A or B denoted by $p(A \cup B) = p(A) + p(B)$
- iv) If A and B are happening together events then the probability of occurrence of probability of A intersection B denoted by $p(A \cap B) = p(A) \cdot p(B)$

13.4 BASIC RELATIONSHIPS OF PROBABILITY

There are some basic probability relationships that can be used to compute the probability of an event without knowledge of all the sample point probabilities.

**Probability
NOTES**

| | |
|---|--|
|  | <p>Complement of an Event: The complement of any event A is the event (not A), i.e., the event that A does not occur. The event A and its complement (not A) are mutually exclusive and exhaustive. It is denoted A', A^c or \bar{A}</p> |
|  | <p>Union of Two Events: the union of events A and B is the event containing all sample points that are in A or B or both. It is denoted by $A \cup B$</p> |
|  | <p>Intersection of Two Events: The intersection of events A and B is the set of all sample points that are in both A and B. It is denoted by $A \cap B$</p> |
|  | <p>○ Mutually Exclusive Events: two sets are mutually exclusive (also called disjoint) if they do not have any elements in common; they need not together comprise the universal set.</p> |

13.5 ADDITION THEOREM OF PROBABILITY

The probability of an event in a random experiment as well as axiomatic approach formulated by Russian Mathematician A.N. Kolmogorov and observed that probability as a function of outcomes of an experiment. By now you know that the probability $P(A)$ of an event A associated with a discrete sample space is the sum of the probabilities assigned to the sample points in A as discussed in axiomatic approach of probability. Here we will learn Addition Theorem of Probability to find probability of occurrence for simultaneous trials under two conditions when events are mutually exclusive and when they are not mutually exclusive.

1. Addition Theorem For Mutually Exclusive Events

Statement: If A and B are two mutually exclusive events, then the probability of occurrence of either A or B is the sum of the individual probabilities of A and B. Symbolically

*Self-Instructional
Material*

Probability
NOTES

$$P(A \cup B) = P(A \text{ or } B) = P(A) + P(B)$$

Proof : Let N be the total number exhaustive and equally likely cases of an experiment. Let m_1 and m_2 be the number of cases favourable to the happening of events A and B respectively. Then

$$P(A) = \frac{n(A)}{n(S)} = \frac{m_1}{N}$$

and

$$P(B) = \frac{n(B)}{n(S)} = \frac{m_2}{N}$$

Since the events A and B are mutually exclusive, the total number of events favorable to either A or B i.e. $n(A \cup B) = m_1 + m_2$, then

$$P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{n(A \cup B)}{N} = \frac{m_1 + m_2}{N} = \frac{m_1}{N} + \frac{m_2}{N} = P(A) + P(B)$$

Example 1: A card is drawn at random from a pack of 52 cards. Find the probability that the drawn card is either a club or an ace of diamond.

Solution : Let A : Event of drawing a card of club and

B: Event of drawing an ace of diamond

$$P(A) = \frac{13}{52}$$

The probability of drawing a card of club

$$P(B) = \frac{1}{52}$$

The probability of drawing an ace of diamond

Since the events are mutually exclusive, the probability of the drawn card being a club or an ace of diamond is:

$$P(A \cup B) = P(A) + P(B) = \frac{13}{52} + \frac{1}{52} = \frac{14}{52} = \frac{7}{26}$$

2. Addition Theorem For Non-Mutually Exclusive Events

The addition theorem discussed above is not applicable when the events are not mutually exclusive. For example, if one card is drawn at random from a pack of 52 cards then in order to find the probability of either a spade or a king card, it cannot be calculated by simply adding the probabilities of spade and king card because the events are not mutually exclusive as there is one card which is a spade as well as a king. Thus, the events are not mutually exclusive; therefore, the addition theorem is modified as:

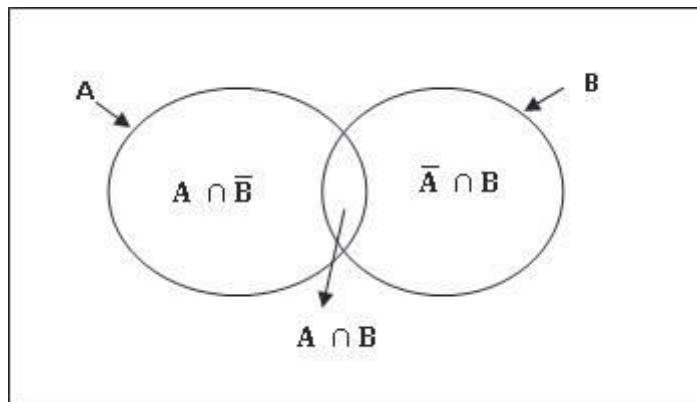
Statement: If A and B are not mutually exclusive events, the probability of the occurrence of either A or B or both is equal to the probability that event A occurs, plus the probability that event B occurs minus the probability of occurrence of the events common to both A and B. In other words the probability of occurrence of at least one of them is given by

$$P(A \cup B) = P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$$

Proof: Let us suppose that a random experiment results in a sample space S with N sample points (exhaustive number of cases). Then by definition

$$P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{n(A \cup B)}{N}$$

Where $n(A \cup B)$ is the number of occurrences (sample points) favorable to the event (AUB)



Addition theorem for non-mutually exclusive events

From the above diagram, we get:

$$\begin{aligned} P(A \cup B) &= \frac{[n(A) - n(A \cap B)] + n(A \cap B) + [n(B) - n(A \cap B)]}{N} \\ &= \frac{n(A) + n(B) - n(A \cap B)}{N} \\ &= \frac{n(A)}{N} + \frac{n(B)}{N} - \frac{n(A \cap B)}{N} \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Example 2

A card is drawn at random from a pack of 52 cards. Find the probability that the drawn card is either a spade or a king.

Solution: Let A: Event of drawing a card of spade and

B: Event of drawing a king card

The probability of drawing a card of spade

$$P(A) = \frac{13}{52}$$

The probability of drawing a king card

$$P(B) = \frac{4}{52}$$

Because one of the kings is a spade card also therefore, these events are not mutually exclusive. The probability of drawing a king of spade is

$$P(A \cap B) = \frac{1}{52}$$

So, the probability of the drawing a spade or king card is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

13.6 MULTIPLICATION THEOREM OF PROBABILITY

In the addition theorem of probability for mutually exclusive events as well as for those events which are not mutually exclusive. In many situations we want to find the probability of simultaneous occurrence of two or more events. Sometimes the information is available that an event A has occurred and one is required to find the probability of occurrence of another event B utilizing the information about event A. Such a probability is known as conditional probability. Here we shall discuss the important concept of conditional probability of an event which will be helpful in understanding the concept of multiplication theorem of probability as well as independence of events.

1. Multiplication Theorem for Independent Events

Statement: This theorem states that if two events A and B are independent then the probability that both of them will occur is equal to the product of their individual probabilities.

$$P(A \cap B) = P(A) \cdot P(B)$$

Proof

If an event A can happen in n_1 ways out of which a_1 are favorable and the event B can happen in n_2 ways out of which a_2 are favorable, we can combine each favorable event in the first with each favorable event in the second case. Thus, the total number of favorable cases is $a_1 \times a_2$. Similarly, the total number of possible cases is $n_1 \times n_2$. Then by definition the probability of happening of both the independent events is

$$P(A \cap B) = P(A \text{ and } B) = \frac{a_1 \times a_2}{n_1 \times n_2} = \frac{a_1}{n_1} \times \frac{a_2}{n_2} = P(A) \times P(B)$$

$$\text{as } P(A) = \frac{a_1}{n_1} \text{ \& } P(B) = \frac{a_2}{n_2}$$

Similarly we can extend the theorem to three events

$$P(A \cap B \cap C) = P(A) \cdot P(B/A) \cdot P(C/A \cdot B)$$

Example 1. From a pack of 52 cards, two cards are drawn at random one after the other with replacement. What is the probability that both cards are kings?

Solution:

The probability of drawing a king $P(A) = \frac{4}{52}$

The probability of drawing again the king after replacement $P(B) = \frac{4}{52}$

Since the two events are independent, the probability of drawing two kings is:

$$P(A \text{ and } B) = P(A \cap B) = P(A) \times P(B) = \frac{4}{52} \times \frac{4}{52} = \frac{1}{169}$$

2. Multiplication Theorem of Probability for Dependent Events

Statement: The probability of simultaneous happening of two events A and B is given by:

$$P(A \cap B) = P(A) \cdot P(B|A); P(A) \neq 0$$

$$P(B \cap A) = P(B) \cdot P(A|B); P(B) \neq 0$$

Where $P(B|A)$ is the conditional probability of happening of B under the condition that A has happened and $P(A|B)$ is the conditional probability of happening of A under the condition that B has happened.

Proof:

Let A and B be the events associated with the sample space S of a random experiment with exhaustive number of outcomes (sample points) N, i.e., $n(S) = N$. Then by definition

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)}$$

For the conditional event $A|B$ (i.e., the happening of A under the condition that B has happened), the favorable outcomes (sample points) must be out of the sample points of B. In other words, for the event $A|B$, the sample space is B and hence

$$P(A|B) = \frac{n(A \cap B)}{n(B)}$$

Similarly, we have

Probability

NOTES

$$P(B|A) = \frac{n(B \cap A)}{n(A)}$$

On multiplying and dividing equation (1) by $n(A)$, we get

$$\begin{aligned} P(A \cap B) &= \frac{n(A)}{n(S)} \times \frac{n(A \cap B)}{n(A)} \\ &= P(A) \cdot P(B|A) \end{aligned}$$

Also

$$\begin{aligned} P(A \cap B) &= \frac{n(B)}{n(S)} \times \frac{n(A \cap B)}{n(B)} \\ &= P(B) \cdot P(A|B) \end{aligned}$$

Example

A bag contains 5 white and 8 red balls. Two successive drawings of 3 balls are made such that (a) the balls are replaced before the second drawing, and (b) the balls are not replaced before the second draw. Find the probability that the first drawing will give 3 white and the second 3 red balls in each case.

Solution:

(a) When balls are replaced.

Total balls in the bag = $8 + 5 = 13$

3 balls can be drawn out of total of 13 balls in ${}^{13}C_3$ ways.

3 white balls can be drawn out of 5 white balls in 5C_3 ways.

$$\text{Probability of 3 white balls} = \frac{P(3W)}{{}^{13}C_3} = \frac{{}^5C_3}{286} = \frac{10}{286}$$

Since the balls are replaced after the first draw so again there are 13 balls in the bag 3 red balls can be drawn out of 8 red balls in 8C_3 ways.

$$\text{Probability of 3 red balls} = \frac{P(3R)}{{}^{13}C_3} = \frac{{}^8C_3}{286} = \frac{56}{286}$$

Since the events are independent, the required probability is:

$$P(3W \text{ and } 3R) = P(3W) \times P(3R) = \frac{{}^5C_3}{{}^{13}C_3} \times \frac{{}^8C_3}{{}^{13}C_3} = \frac{10}{286} \times \frac{56}{286} = \frac{140}{20,449}$$

(b) When the balls are not replaced before second draw

Total balls in the bag = $8 + 5 = 13$

3 balls can be drawn out of 13 balls in ${}^{13}C_3$ ways.

3 white balls can be drawn out of 5 white balls in 5C_3 ways.

The probability of drawing 3 white balls =
$$P(3W) = \frac{{}^5C_3}{{}^{13}C_3}$$

After the first draw, balls left are 10, 3 balls can be drawn out of 10 balls in ${}^{10}C_3$ ways.

3 red balls can be drawn out of 8 balls in 8C_3 ways. Probability of drawing 3 red balls =
$$\frac{{}^8C_3}{{}^{10}C_3}$$
.

Since both the events are dependent, the required probability is:

$$P(3W \text{ and } 3R) = P(3W) \times P(3R|3W) = \frac{{}^5C_3}{{}^{13}C_3} \times \frac{{}^8C_3}{{}^{10}C_3} = \frac{5}{143} \times \frac{7}{15} = \frac{7}{429}$$

13.7 CONDITIONAL PROBABILITY

When the occurrence of an event A and are required to find out the probability of the occurrence of another event B. Two events A and B are said to be dependent when event A can occur only when event B is known to have occurred (or vice versa). The probability attached to such an event is called the conditional probability and is denoted by P (A|B) or in other words, probability of A given that B has occurred. For example, if we want to find the probability of an ace of spade if we know that card drawn from a pack of cards is black. Let us consider another problem relating to dairy plant. There are two lots of full cream packets A and B, each containing some defective packets. A coin is tossed and if it turns up with its head upside lot A is selected and if it turns with tail up, lot B is selected. In this problem we are interested to know the probability of the event that a milk packet selected from the lot obtained in this manner is defective.

Definition: If two events A and B are dependent, then the conditional probability of B given that event A has occurred is defined as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \text{ if } P(A) \geq 0$$

Let us consider the experiment of throwing of a die once. The sample space of this experiment is {1, 2, 3, 4, 5, and 6}.

Let E_1 : an even number and E_2 : multiple of 3 p.

Then E_1 : {2, 4, 6} and E_2 : {3, 6}.

Hence, $P(E_1) = 3/6 = 1/2$ and $P(E_2) = 2/6 = 1/3$

In order to find the probability of occurrence of E_2 when it is given that E_1 has occurred We know that in a single throw of die 2 or 4 or 6 has come up. Out of these only 6 is favorable to E_2 . So the probability of occurrence of E_2 when it is given that E_1 has occurred is equal to 1/3. This probability of E_2 when E_1 has occurred is written as

Probability
NOTES

$P(E_2|E_1)$. Here we find that $P(E_2|E_1) = P(E_2)$.

Let us consider the event

E_3 : a number greater than 3 then $E_3: \{4,5,6\}$ and $P(E_3) = 3/6 = 1/2$
Out of 2,4 and 6, two numbers namely 4 and 6 are favorable to E_3 .

Therefore, $P(E_3|E_1) = 2/3$.

The events of the type E_1 and E_2 are called independent events as the occurrence or non-occurrence of E_1 does not affect the probability of occurrence or non-occurrence of E_2 . The events E_1 and E_3 are not independent.

13.7.1 Combined Use Of Addition And Multiplication Theorem

In probability both addition and multiplication theorems are used simultaneously. The following examples illustrate the combined use of addition and multiplication theorems.

Example

A bag contains 5 white and 4 black balls. A ball is drawn from this bag and is replaced and then second draw of a ball is made. What is the probability that two balls are of different colors.

Solution: There are two possibilities

- i) First ball is white and the second ball drawn is black.
- ii) First ball is black and the second ball drawn is white.

Since the events are independent, so by using multiplication theorem we have

- i) Probability of drawing First ball white and the second ball black = $\frac{5}{9} \times \frac{4}{9} = \frac{20}{81}$

- ii) Probability of drawing First ball black and the second ball white = $\frac{4}{9} \times \frac{5}{9} = \frac{20}{81}$

Since these probabilities are mutually exclusive, by using addition theorem

$$\text{Probability that two balls are of different colors} = \frac{20}{81} + \frac{20}{81} = \frac{40}{81}$$

CHECK YOUR PROGRESS

- 1. What is sample space?

2. What is an event?
3. Write the formula for addition probability theorem
4. Mention the types of probability
5. How Baye's theorem is calculated

13.8 BAYES' THEOREM AND ITS APPLICATIONS

There are many situations where the ultimate outcome of an experiment depends on what happens in various intermediate stages. This issue is resolved by the Bayes'

There is a very big difference between $P(A | B)$ and $P(B | A)$

Suppose that a new test is developed to identify people who are liable to suffer from some genetic disease in later life. Of course, no test is perfect; there will be some carriers of the defective gene who test negative, and some non-carriers who test positive. So, for example, let A be the event 'the patient is a carrier', and B the event 'the test result is positive'.

The scientists who develop the test are concerned with the probabilities that the test result is wrong, that is, with $P(B | A')$ and $P(B' | A)$. However, a patient who has taken the test has different concerns.

If I tested positive, what is the chance that I have the disease?

If I tested negative, how sure can I be that I am not a carrier? In other words, $P(A | B)$ and $P(A' | B')$.

These conditional probabilities are related by Bayes' Theorem:

Let A and B be events with non-zero probability. Then

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

The proof is not hard. We have

$$P(A | B) \cdot P(B) = P(A \cap B) = P(B | A) \cdot P(A),$$

using the definition of conditional probability twice. (Note that we need both A and B to have non-zero probability here.) Now divide this equation by P(B) to get the result.

If $P(A) \neq 0$, and $P(B) \neq 0$, then

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B | A) \cdot P(A) + P(B | A') \cdot P(A')} \cdot$$

Bayes' Theorem is often stated in this form.

Example

Consider the clinical test described at the start of this section. Suppose that 1 in 1000 of the population is a carrier of the disease. Suppose also

Probability
NOTES

that the probability that a carrier tests negative is 1%, while the probability that a non carrier tests positive is 5%. (A test achieving these values would be regarded as very successful.) Let A be the event ‘the patient is a carrier’, and B the event ‘the test result is positive’. We are given that $P(A) = 0.001$ (so that $P(A') = 0.999$), and that $P(B | A) = 0.99$, $P(B | A') = 0.05$.

(a) A patient has just had a positive test result. What is the probability that the patient is a carrier? The answer is

$$\begin{aligned}
 P(A | B) &= \frac{P(B | A) \cdot P(A)}{P(B | A) \cdot P(A) + P(B | A') \cdot P(A')} \\
 &= \frac{0.99 \times 0.001}{(0.99 \times 0.001) + (0.05 \times 0.999)} \\
 &= \frac{0.00099}{0.05094} = 0.0194.
 \end{aligned}$$

(b) A patient has just had a negative test result. What is the probability that the patient is a carrier? The answer is

$$\begin{aligned}
 P(A | B') &= \frac{P(B' | A)P(A)}{P(B' | A)P(A) + P(B' | A')P(A')} \\
 &= \frac{0.01 \times 0.001}{(0.01 \times 0.001) + (0.95 \times 0.999)} \\
 &= \frac{0.00001}{0.94095} = 0.00001.
 \end{aligned}$$

13.9 SUMMARY

- Bayes’ Theorem is often stated in the form. If $P(A) \neq 0,1$ and $P(B) \neq 0$, then

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B | A) \cdot P(A) + P(B | A') \cdot P(A')}$$
- **Conditional Probability:** Two events A and B are said to be dependent when event A can occur only when event B is known to have occurred (or vice versa).
- **Multiplication Probability :**The probability of simultaneous occurrence of two or more events
- **Addition Probability:** If A and B are not mutually exclusive events, the probability of the occurrence of either A or B or both is equal to the probability that event A occurs, plus the probability that event B occurs minus the probability of occurrence of the events common to both A and B
- **Types Of Probability:** Axiomatic Approach, Classical Approach ,Relative Frequency Theory of Probability, Subjective Approach

13.10 KEY WORDS

Probability, Sample, Events, Variables, Addition theorem, Multiplication theorem, Axiomatic approach, Classical approach, Relative frequency theory, Subjective approach, Baye’s theorem

13.11 ANSWER TO CHECK YOUR PROGRESS

1. **Sample Space** :Sample Space is the set of all possible outcomes of an experiment. It is denoted by S
2. **Event** :Any subset of a Sample Space is an event. Events are generally denoted by capital letters A, B , C, D etc.
3. Addition Theorem For Mutually Exclusive Events
 $P(A \cup B) = P(A \text{ or } B) = P(A) + P(B)$
Addition Theorem For Non-Mutually Exclusive Events
 $P(A \cup B) = P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$
4. **Types Of Probability:** Axiomatic Approach, Classical Approach ,Relative Frequency Theory of Probability, Subjective Approach
5. Bayes' Theorem is often stated in the form. If $P(A) \neq 0, 1$ and $P(B) \neq 0$, then

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B | A) \cdot P(A) + P(B | A') \cdot P(A')}$$

13.12 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTION:

1. Define probability
2. What are sample space
3. Define random variable
4. State the Baye's theorem
5. Explain mutually exclusive event

LONG ANSWER QUESTIONS:

1. Define probability and bring out the importance of probability
2. Distinguish between independent and dependents events
3. Explain briefly Baye's theorem
4. If 20% of the bottles produced by machine are defective, determine the probability that out of 4 bottles (i) 0, (ii) 1, (iii) at most 2 bottles will be defective

13.13 FURTHER READINGS

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.

UNIT14 - PROBABILITY DISTRIBUTION

Structure

- 14.0 Introduction
- 14.1 Objectives
- 14.2 Random Variable
- 14.3 Types of Random Variable
- 14.4 Binomial Distribution
- 14.5 Poisson Distribution
- 14.6 Normal Distribution
- 14.7 Summary
- 14.8 Key Words
- 14.9 Answer to Check your progress
- 14.10 Questions and Exercise
- 14.11 Further Reading

14.0 INTRODUCTION

A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence. It describes the range of possible values that a random variable can attain and the probability that the value of the random variable is with any subset of that range. For example 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)

14.1 OBJECTIVES

The students will be able to understand

- The random variable and its types in probability distribution
- Concept of Binomial Distribution, Poisson Distribution and Normal Distribution
- Concept of Mean and Standard deviation of Binomial and Poisson Distribution

A random variable is defined as a real number X connected with the outcome of a random experiment E . For example, if E consists of three tosses of a coin, we may consider random variable X which denotes the number of heads (0, 1, 2 or 3)

| | | | | | | | | |
|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|
| Outcome : | HHH | HTH | THH | TTH | HTT | THT | TTH | TTT |
| Value of X : | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 0 |

Thus, to every outcome there corresponds a real number $X(w)$. Since the

NOTES

points of the sample space corresponds to outcomes, this means that a real number, which we denote by $X(w)$, is defined for each $w \in S$ and let us denote them by w_1, w_2, \dots, w_8 i.e. $X(w_1) = 3, X(w_2) = 2, \dots, X(w_8) = 0$. Thus, we define a random variable as a real valued function whose domain is the sample space associated with a random experiment and range is the real line. Generally it is denoted by capital letters X, Y, Z, \dots etc.

14.3 TYPES OF RANDOM VARIABLE

1. Discrete random variable

If a random variable X assumes only a finite or countable set of values, it is called a discrete random variable. In other words, a real valued function defined on a discrete sample space is called a discrete random variable. In case of discrete random variable we usually talk of values at a point. Generally it represents counted data. For example, number of defective milk packet in a milk plant, number of students in a class etc.

2. Continuous random variable

A random variable is said to be continuous if it can assume infinite and uncountable set of values. A continuous random variable is in which different values cannot be put in one to one correspondence with a set of positive integers. For example, weight of baby elephant take any possible value in the interval of 160 kg to 260 kg, say 189 kg or 189.4356 kg; likewise, marks scored by the students in a class etc. In case of continuous random variable we usually take the values in a particular interval. Continuous random variables represent measured data.

Probability Distribution of a Random Variable

The concept of probability distribution is equivalent to the frequency distribution. It depicts how total probability of one is distributed among various values which a random variable can take.

Mean and Variance of a Random variable

Let X denotes the random variable which assumes values x_1, x_2, \dots, x_n with corresponding probabilities p_1, p_2, \dots, p_n . Then the probability distribution be as follow:

| | | | | |
|---------|-------|-------|--------|-------|
| $X:$ | x_1 | x_2 | | x_n |
| $P(X):$ | p_1 | p_2 | | p_n |

Then

$$\sum_{i=1}^n p_i = p_1 + p_2 + \dots + p_n = 1$$

The mean (μ) of the above probability distribution is defined as:

$$\mu = \frac{p_1 x_1 + p_2 x_2 + \dots + p_n x_n}{p_1 + p_2 + \dots + p_n} = \frac{\sum p_i x_i}{\sum p_i} = \sum p_i x_i$$

NOTES

The variance (σ^2) is defined as:

$$\begin{aligned} \sigma^2 &= \sum (x_i - \mu)^2 p_i = \sum (x_i^2 + \mu^2 - 2x_i \mu) p_i = \sum x_i^2 p_i + \mu^2 \sum p_i - 2\mu \sum x_i p_i \\ &= \sum x_i^2 p_i + \mu^2(1) - 2\mu(\mu) = \sum x_i^2 p_i - \mu^2 = \sum x_i^2 p_i - \left(\sum p_i x_i\right)^2 \end{aligned}$$

Mean of a random variable X is also known as expected value and is denoted by E(X)

$$E(X) = \mu = p_1 x_1 + p_2 x_2 + \dots + p_n x_n = \sum p_i x_i$$

$$\text{Variance } (\sigma^2) = E(X^2) - (E(X))^2$$

Example

A die is tossed twice. Getting a number greater than 4 is considered a success. Find the variance of the probability distribution of the number of success.

Solution:

Here p, probability of a number greater than 4 = 2/6 = 1/3 and q,

probability of a number not greater than 4 = 1 - 1/3 = 2/3

$$P(X = 0) = q \times q = \frac{2}{3} \times \frac{2}{3} = \frac{4}{9}$$

$$P(X = 1) = p \times q + q \times p = \frac{1}{3} \times \frac{2}{3} + \frac{2}{3} \times \frac{1}{3} = \frac{4}{9}$$

$$P(X = 2) = p \times p = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$$

Thus, we have:

| x_i | p_i | $p_i x_i$ | x_i^2 | $p_i x_i^2$ |
|-------|-------|-----------|---------|-------------|
| 0 | 4/9 | 0 | 0 | 0 |
| 1 | 4/9 | 4/9 | 1 | 4/9 |
| 2 | 1/9 | 2/9 | 4 | 4/9 |
| Total | | 6/9 | | 8/9 |

Hence, the

$$\text{mean } \mu = \sum p_i x_i = \frac{6}{9} = \frac{2}{3}$$

The variance

$$\sigma^2 = \sum p_i x_i^2 - \mu^2 = \frac{8}{9} - \left(\frac{6}{9}\right)^2 = \frac{8}{9} - \frac{36}{81} = \frac{72-36}{81} = \frac{36}{81} = \frac{4}{9}$$

14.4 BINOMIAL DISTRIBUTION

Binomial distribution is a discrete probability distribution. This distribution was discovered by a Swiss Mathematician James Bernoulli (1654-1705). A Bernoullian trial is an experiment having only two possible outcomes i.e. success or failure. In other words the result of the trial are dichotomous e.g. in tossing of a coin either head or tail, the sex

NOTES

of a calf can be either male or female, a manufactured milk product or an engineering equipment or spare part will be either defective or non defective etc. This distribution can be used under the following conditions:

- a) The random experiment is performed repeatedly a finite and fixed number of times i.e. n , the number of trials is finite and fixed.
- b) The outcome of a trial results in the dichotomous classification of events i.e. each trial must result in two mutually exclusive outcomes, success or failure.
- c) Probability of success (or failure) remains same in each trial i.e. in each trail the probability of success, denoted by p remains constant. $q=1-p$, is then termed as the probability of failure (non-occurrence).
- d) Trials are independent i.e. the outcome of any trial does not affect the outcomes of the subsequent trials.

Theorem:

If X denotes the number of successes in n trials satisfying the above conditions, then X is a random variable which can take values $0,1,2,\dots,n$ i.e. no success, one success, two successes,, or all the n successes. The general expression for the probability of r successes is given by: $P(r) = P(X = r) = {}^n C_r p^r q^{n-r}$ for $r=0,1,2,\dots,n$

Proof :

By the theorem of compound probability, the probability that r trials are success and the remaining $(n-r)$ are failures in a sequence of n trials in a specified order say S,F,S,F,S,\dots,S is given by

$$\begin{aligned}
 P(S \cap F \cap S \cap F \cap \dots \cap S) &= P(S)P(F)P(S)P(F)P(F) \dots P(S) \\
 &= p \cdot q \cdot p \cdot q \dots p \\
 &= (p \times p \times p \dots r \text{ times}) \times (q \times q \times q \dots (n-r) \text{ times}) = p^r q^{(n-r)}
 \end{aligned}$$

But we are interested in any r trials being successes and since r trials can be chosen out of n trials in ${}^n C_r$ (mutually exclusive) ways. Therefore, by the theorem of total probability, the chance $P(r)$ of r successes in a series of n independent trials is given by

$$P(r) = {}^n C_r p^r q^{n-r} \quad 0 \leq r \leq n$$

r can take only positive integer values.

Thus, the chance variate i.e. the number of successes, can take the values $0,1,2,\dots,r,\dots,n$ with corresponding probabilities

$$q^n, {}^n C_1 p q^{n-1}, \dots, {}^n C_r p^r q^{n-r}, \dots, p^n$$

- The probability distribution of the number of successes so obtained is called the binomial probability distribution for the obvious reason that the probabilities are the various terms of the binomial expansion of $(q+p)^n$.
- The sum of probabilities

NOTES

$$\sum_{r=0}^n p(r) = p(0) + p(1) + p(2) + \dots + p(r)$$

$$= q^n + {}^nC_1 p q^{n-1} + \dots + {}^nC_r p^r q^{n-r} + \dots + p^n = (q + p)^n$$

$$= 1$$

- The expression for P (X = r) is known as probability mass function of the Binomial distribution with parameter **n** and **p**. The random variable **X** following this probability law is called binomial variate with parameter n and p denoted as **X~B(n,p)**. Hence binomial distribution can be completely determined if n and p are known .

Example :

It is known that 40 % patients affected by tuberculosis die every year. 6 patients are admitted to a hospital suffering from tuberculosis. What is the probability that

- (i) Three patients will die.
- (ii) at least patients will die
- (iii) all patients will be cured
- (iv) no patients will be saved.

Solution

we have **p = 0.4** , **q = 1- 0.40 = 0.6** and **n=6**

In binomial distribution we have

$$P(r) = {}^nC_r \cdot p^r \cdot q^{n-r}$$

- (i) Prob. [Three patients will die]

$$P[r = 3] = P(3) = {}^6C_3 \cdot (0.4)^3 (0.6)^3$$

$$P(3) = \frac{6!}{3!3!} (0.4)^3 (0.6)^3 = 20(0.4)^3(0.6)^3 = 0.2765$$

- (ii) Prob. (at least five patients will die)

$$P(5) + P(6) = {}^6C_5 (0.4)^5 (0.6)^1 + {}^6C_6 (0.4)^6(0.6)^0$$

$$= 6 (0.4)^5 (0.6)^1 + (0.4)^6$$

$$= 0.0369 + 0.0041 = 0.0410$$

- (iii) Prob. (all patients will be cured) = 1 - P (no patients will die)

$$1 - P(0) = 1 - {}^6C_0 (0.4)^0(0.6)^6$$

$$= 1 - (0.6)^6$$

$$= 1 - 0.0467 = 0.9533$$

- (iv) Prob. (no patients will be saved) = P (all patients will die)

$$= P(6)$$

$$= {}^6C_6 (0.4)^6 (0.6)^0$$

$$= (0.4)^6 = 0.0041$$

Example of Binomial distribution

- The number of heads/tails in a sequence of coin flips
- Vote counts for two different candidates in an election

- The number of male/female employees in a company
- The number of accounts that are in compliance or not in compliance with an accounting procedure
- The number of successful sales calls
- The number of defective products in a production run

Properties of Binomial Distribution

i) **Mean of binomial distribution is np .**

Proof: First raw moment

$$\begin{aligned}\mu'_1 &= E(r) = \sum_{r=0}^n r \cdot n C_r p^r q^{n-r} = \sum_{r=0}^n r \frac{n!}{r!(n-r)!} p p^{r-1} q^{n-r} \\ &= \sum_{r=0}^n \frac{r \cdot n(n-1)!}{r(r-1)!(n-1-(r-1))!} p p^{r-1} q^{(n-1)-(r-1)} = np \sum_{r=0}^n n-1 C_{r-1} p^{r-1} q^{(n-1)-(r-1)} \\ &= \sum_{r=0}^n np n-1 C_{r-1} p^{r-1} q^{(n-1)-(r-1)} = \\ &= np \sum_{r=0}^n n-1 C_{r-1} p^{r-1} q^{(n-1)-(r-1)} = np(q+p)^{n-1} = np\end{aligned}$$

ii) **Variance of binomial distribution is npq**

Proof: Second raw moment

$$\begin{aligned}\mu'_2 &= E(r^2) = \sum_{r=0}^n r^2 n C_r p^r q^{n-r} = \sum_{r=0}^n \{r + r(r-1)\} n C_r p^r q^{n-r} \\ &= \sum_{r=0}^n r n C_r p^r q^{n-r} + \sum_{r=0}^n r(r-1) n C_r p^r q^{n-r} = np + n(n-1)p^2 = np + n^2p^2 - np^2\end{aligned}$$

Variance

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = np + n^2p^2 - np^2 - n^2p^2 = np(1-p) = npq$$

For the binomial distribution if mean and variance are known, we can arrive at the frequency distribution and variance is less than mean.

iii) The third and fourth central moment μ_3 and μ_4 can be obtained on the same lines.

$$\begin{aligned}\mu_3 &= npq(q-p) \\ \mu_4 &= npq[1 + 3(n-2)pq]\end{aligned}$$

iv) Pearson's constants β_1 & β_2 as well as γ_1 and γ_2 are given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{[npq(1-2p)]^2}{(npq)^3} = \frac{[(1-2p)]^2}{npq} \quad \gamma_1 = \sqrt{\beta_1} = \frac{(1-2p)}{\sqrt{npq}}$$

NOTES

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{npq[1+3(n-2)pq]}{(npq)^2} = 3 + \frac{1-6pq}{npq}, \quad \gamma_2 = \beta_2 - 3 = \frac{1-6pq}{npq}$$

γ_1 shows that the binomial distribution is positively skewed if $q > p$ or $p < 1/2$ and it is negatively skewed if $q < p$ or $p > 1/2$ and it is symmetrical if $p = q = 1/2$. The binomial distribution is leptokurtic if $pq < 1/6$ and platykurtic if $pq > 1/6$.

v) Mode of binomial distribution is determined by the value $(n+1)p$. If this value is an integer equal to k then the distribution is bi-modal, the two modal values being $X=k$ and $X=k-1$. When this value is not an integer then the distribution has unique mode at $X=k_1$, the integral part of $(n+1)p$.

vi) Additive property: If X_1 is $B(n_1, p)$ and X_2 is $B(n_2, p)$ and they are independent then their sum $X_1 + X_2$ is also a binomial variate $B(n_1 + n_2, p)$.

Example:

If the mean and variance of a Binomial Distribution are respectively 9 and 6, find the distribution.

Solution: Mean of Binomial Distribution is np and variance is npq

$$\therefore np = 9 \text{ and } npq = 6$$

$$\text{Now } \frac{npq}{np} = \frac{6}{9} \Rightarrow q = \frac{2}{3}$$

$$\therefore p = 1 - q = 1 - \frac{2}{3} = \frac{1}{3}$$

$$\therefore np = 9 \Rightarrow n \cdot \frac{1}{3} = 9 \Rightarrow n = 3 \times 9 = 27$$

Hence, the Binomial Distribution is $\left(\frac{2}{3} + \frac{1}{3}\right)^{27}$

$$\text{i.e. } {}^{27}C_r \left(\frac{1}{3}\right)^r \left(\frac{2}{3}\right)^{27-r}$$

14.5 POISSON DISTRIBUTION

Poisson distribution is a limiting case of Binomial distribution under the following conditions:

- n , the no. of trials is indefinitely large i.e., $n \rightarrow \infty$
- p , the constant probability of success for each trial is indefinitely small i.e. $p \rightarrow 0$
- $np = m$ (say) is finite. Thus, $p = m/n$, $q = 1 - m/n$ where m is a positive real number.

Under, the above three conditions the probability mass function of

NOTES

binomial distribution tends to the probability mass function of the Poisson distribution whose definition and derivation given below:

A random variable X is said to follow a Poisson distribution if it assumes only non-negative values and its probability mass function is given by

$$P(r, m) = P(x = m) = \frac{e^{-m} m^r}{r!}; \quad r = 0, 1, 2, \dots, \infty$$

where m is known as the parameter of the distribution.

e = 2.7183 (the base of the natural logarithm)

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots + \infty$$

$$e^{-x} = 1 - \frac{x}{1!} + \frac{x^2}{2!} - \frac{x^3}{3!} + \dots + (-1)^n \frac{x^n}{n!} + \dots + \infty$$

Proof: As $n \rightarrow \infty$ and $np = m$
 $p = m/n$ and $q = 1 - m/n$

Probability function of binomial distribution is

$$\begin{aligned} P(r) &= {}^n C_r p^r q^{n-r} = \frac{n!}{r!(n-r)!} p^r q^{n-r} \\ &= \frac{n(n-1)(n-2)\dots[n-(r-1)]}{r!} \left(\frac{m}{n}\right)^r \left(1 - \frac{m}{n}\right)^{n-r} \\ &= \frac{m^r}{r!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{r-1}{n}\right) \times \left(1 - \frac{m}{n}\right)^n \left(1 - \frac{m}{n}\right)^{-r} \end{aligned}$$

Taking limit as $n \rightarrow \infty$

$$= \frac{m^r}{r!} (1 - 0)(1 - 0) \dots (1 - 0) \times \lim_{n \rightarrow \infty} \left(1 - \frac{m}{n}\right)^n \lim_{n \rightarrow \infty} \left(1 - \frac{m}{n}\right)^{-r}$$

We know that

$$\lim_{n \rightarrow \infty} \left(1 - \frac{m}{n}\right)^n = e^{-m}$$

$$\lim_{n \rightarrow \infty} \left(1 - \frac{m}{n}\right)^a = 1 \quad \text{a is not a function of n}$$

$$\text{Thus, } P(r) = \frac{m^r}{r!} \frac{e^{-m} \cdot 1}{1} = \frac{e^{-m} m^r}{r!}; \quad r = 0, 1, 2, \dots, \infty$$

Putting $r = 0, 1, 2, \dots$ in above equation, we obtain the probabilities of $r =$

$0, 1, 2, \dots$ successes respectively we get $e^{-m}, \frac{e^{-m} m^1}{1!}, \frac{e^{-m} m^2}{2!}, \dots$

Total probability is 1:

$$\sum_{r=0}^{\infty} P(r) = \sum_{r=0}^{\infty} P(x = r) = \sum_{r=0}^{\infty} \frac{e^{-m} m^r}{r!} = \sum_{r=0}^{\infty} P(x = r) = e^{-m} + m e^{-m} + \frac{e^{-m} m^2}{2!} + \dots$$

NOTES

$$= e^{-m} \left(1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \dots \right) = e^{-m} \sum_{r=0}^{\infty} \frac{m^r}{r!} = e^{-m} \cdot e^m = 1$$

If we know m, all the probabilities of the Poisson distribution can be obtained, therefore m is the only parameter of the Poisson distribution. The application of this distribution in solving problems is illustrated through following examples.

Example

A manufacturer of screws knows that 5% of his product is defective. If he sells his product in a carton of 100 items and guarantees that not more than 10 items will be defective. What is the probability that the carton will fail to meet the guaranteed quality?

Solution:

In this example $p = 0.05$, $n = 100$. Therefore, $m = n \cdot p = 100(0.05) = 5$

Prob. [That the carton will fail to meet the guaranteed quality] = 1 - Prob. [The carton will meet the guaranteed quality] = Prob. [Not more than 10 items will be defective] = $1 - P[r \leq 10]$
 = $1 - [P(0) + P(1) + P(2) + P(3) + \dots + P(10)]$

$$P(r) = \frac{e^{-m} m^r}{r!}$$

In case of Poisson distribution

Therefore, we have

$$P(r > 10) = 1 - P(r \leq 10) = 1 - \left(\frac{e^{-5} 5^0}{0!} + \frac{e^{-5} 5^1}{1!} + \frac{e^{-5} 5^2}{2!} + \dots + \frac{e^{-5} 5^{10}}{10!} \right)$$

$$= 1 - e^{-5} \left[1 + 5 + \frac{5^2}{2!} + \frac{5^3}{3!} + \dots + \frac{5^{10}}{10!} \right] = 1 - 0.9865 = 0.0135$$

Examples of Poisson Distribution

- The hourly number of customers arriving at a bank
- The daily number of accidents on a particular stretch of highway
- The hourly number of accesses to a particular web server
- The daily number of emergency calls in Dallas
- The number of typos in a book
- The monthly number of employees who had an absence in a large company
- Monthly demands for a particular product

Properties of Poisson Distribution

i) Mean of the Poisson distribution is m

$$\begin{aligned}\mu'_1 = \text{Mean} &= \sum_{r=0}^{\infty} r \frac{e^{-m} m^r}{r!} = \sum_{r=0}^{\infty} r \frac{e^{-m} m^r}{(r-1)!} = m \sum_{r=0}^{\infty} \frac{e^{-m} m^{r-1}}{(r-1)!} \\ &= m e^{-m} \left[1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \dots \right] = m e^{-m} e^m = m\end{aligned}$$

ii) Variance of the Poisson distribution is

$$\text{Variance} = \sum_{r=0}^{\infty} r^2 p(r) - \left(\sum_{r=0}^{\infty} r p(r) \right)^2 = \sum_{r=0}^{\infty} r^2 p(r) - (m)^2$$

where,

$$\begin{aligned}\mu'_2 &= \sum_{r=0}^{\infty} r^2 \frac{e^{-m} m^r}{r!} = \sum_{r=0}^{\infty} [r + r(r-1)] \frac{e^{-m} m^r}{r!} = \sum_{r=0}^{\infty} r \frac{e^{-m} m^r}{r!} + \sum_{r=0}^{\infty} r(r-1) \frac{e^{-m} m^r}{r!} \\ &= m + e^{-m} m^2 \sum_{r=0}^{\infty} \frac{m^{r-2}}{(r-2)!} = m + e^{-m} m^2 \left[1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \dots \right] \\ &= m + e^{-m} m^2 e^m = m + m^2\end{aligned}$$

$$\text{Variance} = \mu'_2 - (\mu'_1)^2 = m + m^2 - (m)^2 = m$$

Hence, for Poisson distribution with parameter m mean is equal to variance.

iii) Third and fourth central moments μ_3 and μ_4

$$\mu_3 = m, \quad \mu_4 = 3m^2 + m$$

iv) Pearson's constants β_1 & β_2 as well as γ_1 and γ_2 are given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(m)^2}{(m)^3} = \frac{1}{m}, \quad \gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{m}}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3m^2+m}{(m)^2} = 3 + \frac{1}{m}, \quad \gamma_2 = \beta_1 - 3 = \frac{1}{m}$$

It may be noted that the first three central moments of the Poisson distribution are identical and are equal to the value of parameter itself namely 'm'. Hence Poisson distribution is always a positively skewed distribution as $m > 0$ as well as leptokurtic. As the value of m increases γ_1 decreases and the thus Skewness is reduced for increasing values of m. As $m \rightarrow \infty$, γ_1 and γ_2 tend to zero. So we conclude that as $m \rightarrow \infty$, the curve of the Poisson distribution tends to be symmetrical curve for large values of m.

v) Mode of Poisson distribution is determined by the value m. If m is an integer then the distribution is bi-modal, the two modal values being $X=m$ and $X=m-1$. When m is not an integer then the distribution has unique modal value being integral part of m.

vi) Additive property: If X_1 and X_2 are two independent Poisson variate with parameters m_1 and m_2 then their sum $X_1 + X_2$ is also a Poisson variate with parameter $m_1 + m_2$.

Example

The mean of the Poisson distribution is 2.25. Find the other constants of the distribution.

NOTES

Solution:

We have $m = 2.25$

$$\sigma = \sqrt{m} = \sqrt{2.25} = 1.5$$

$$\mu_1 = 0$$

$$\mu_2 = m = 2.25$$

$$\mu_3 = m = 2.25$$

$$\mu_4 = m + 3m^2 = 2.25 + 3(2.25)^2 = 2.25 + 15.1875 = 17.4375$$

$$\beta_1 = \frac{1}{m} = \frac{1}{2.25} = 0.444$$

$$\beta_2 = 3 + \frac{1}{m} = 3 + 0.444 = 3.444$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{m}} = \frac{1}{1.5} = 0.67$$

$$\gamma_2 = \beta_2 - 3 = 3 + \frac{1}{m} - 3 = \frac{1}{2.25} = 0.444$$

This curve is positively Skewed and Leptokurtic.

CHECK YOUR PROGRESS - 1

1. List the types of random variable
2. What are the properties of Binominal distribution?
3. List out few example of Poisson distribution

14.6 NORMAL DISTRIBUTION

Normal distribution is one of the important distribution in continuous probability distribution. Normal distribution is probably the most important and widely used theoretical distribution. Normal distribution unlike the Binomial and Poisson is a continuous probability distribution. It has been observed that a vast number of variables arising in studies of agricultural and dairying, social, psychological and economic phenomena tend to follow normal distribution. The normal distribution was first discovered by French Mathematician Abraham De-Moivre in 1733, who obtained this continuous distribution as a limiting case of the Binomial distribution. But it was later rediscovered and applied by Laplace and Karl Gauss. It is also known as Gaussian distribution after the name of Karl Friedrich Gauss.

A continuous random variable X is said to have a normal distribution with parameters μ (mean) and σ (standard deviation), if its density function is given by the probability law

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

where π and e are given by $\pi = 22/7$ and $e=2.7183$ (base of natural logarithms).

Properties of normal distribution:

1) A random variable X with mean μ and variance σ^2 following the normal law given above is represented as $X \sim N(\mu, \sigma^2)$.

2) If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$, Z is defined as a standard normal variate with $E(Z)=0$ and $\text{Var}(Z)=1$ and we write $Z \sim N(0, 1)$

3) The p.d.f. of a standard normal variate Z is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}, -\infty < z < \infty, \text{ where } Z = \frac{X - \mu}{\sigma}$$

4) Normal distribution is a limiting form of the binomial distribution when

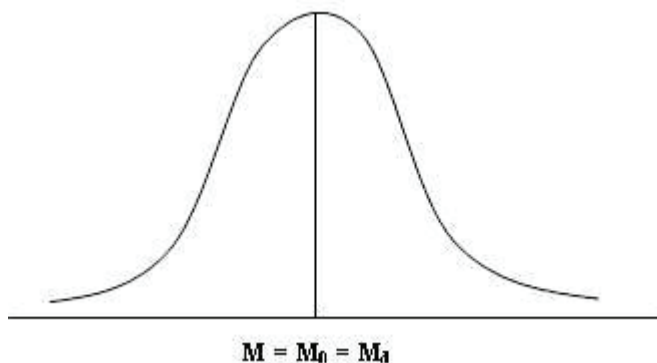
- n , the number of trials is indefinite large, i.e. $n \rightarrow \infty$ and
- neither p nor q is very small.

5) Normal distribution is a limiting form of Poisson distribution when its mean m is large and n is also large.

Characteristics of Normal Distribution

It has the following properties:

- The graph of $f(x)$ is bell shaped unimodal and symmetric curve as shown in the Fig. 12.1. The top of the bell is directly above the mean (μ).



Normal probability curve

- The curve is symmetrical about the line $X = \mu$, ($Z = 0$) i.e., it has the same shape on either side of the line $X = \mu$, (or $Z = 0$). This is because the equation of the curve $\phi(z)$ remains unchanged if we change z to $-z$.

NOTES

3. Since the distribution is symmetrical, mean, median and mode coincide. Thus, Mean = Median = Mode = μ
4. Since Mean = Median = Mode = μ , the ordinate at $X = \mu$, ($Z = 0$) divides the whole area into two equal parts. Further, since total area under normal probability curve is 1, the area to the right of the ordinate as well as to the left of the ordinate at $X = \mu$ (or $Z = 0$) is 0.5
5. Also, by virtue of symmetry the quartiles are equidistant from median (μ), i.e,

$$Q_3 - M_d = M_d - Q_1$$

6. Since the distribution is symmetrical, all moments of odd order about the mean are zero. Thus $\mu_{2n+1} = 0$; ($n = 0,1,2,\dots$) i.e. $\mu_1 = \mu_3 = \mu_5 = \dots = 0$.
7. The moments (about mean) of even order are given by

$$\mu_{2n} = 1.3.5 \dots (2n - 1)\sigma^{2n}, (n = 1,2,3 \dots)$$

Putting $n=1$ and 2 we get

$$\mu_2 = \sigma^2 \text{ and } \mu_4 = 3\sigma^4$$

$$\therefore \beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$$

and

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\sigma^4}{\sigma^4} = 3$$

8. Since the distribution is symmetrical, the moment coefficient of skewness based on moments is given by $\beta_1 = 0 \Rightarrow \gamma_1 = 0$
9. The coefficient of kurtosis is given by $\beta_2 = 3 \Rightarrow \gamma_2 = 0$
10. No portion of the curve lies below the x-axis, since $f(x)$ being the probability can never be negative.
11. Theoretically, the range of the distribution is from $-\infty < \text{to } < \infty$. But practically, range = 6σ
12. As x increases numerically [i.e. on either side of $X = \mu$], the value of $f(x)$ decreases rapidly, the maximum probability occurring at $X = \mu$ and is given by

$$[f(x)]_{\max} = \frac{1}{\sqrt{2\pi}\sigma}$$

Thus maximum value of $f(x)$ is inversely proportional to the standard deviation. For large values of σ , $f(x)$ increases, i.e., the curve has a normal peak.

13. Distribution is unimodal with the only mode occurring at $X = \mu$.
14. X-axis is an asymptote to the curve i.e., for numerically large value of X (on either side of the line ($X = \mu$)), the curve becomes parallel to the X-axis and is supposed to meet it at infinity.
15. A linear combination of independent normal variates is also a normal variate. If X_1, X_2, \dots, X_n are independent normal variates with mean $\mu_1, \mu_2, \dots, \mu_n$ and standard deviations $\sigma_1, \sigma_2, \dots$,

NOTES

σ_n respectively then their linear combination

$$a_1X_1 + a_2X_2 + \dots + a_nX_n$$

Where a_1, a_2, \dots, a_n are constants, is also a normal variate with Mean = $a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n$ and Variance = $a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots + a_n^2\sigma_n^2$. In particular, if we take $a_1 = a_2 = \dots = a_n = 1$ then we get $X_1 + X_2 + \dots + X_n$ is a normal variate with mean $\mu_1 + \mu_2 + \dots + \mu_n$ and variance $\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2$. Thus, the sum of independent normal variates is also a normal variate with mean equal to sum of their means and standard deviation equal to square root of sum of the squares of their standard deviations. This is known as the Re-productive or Additive Property of the Normal distribution.

16. Mean Deviation (M.D.) about mean or median or mode is given by

$$M.D. = \sqrt{\frac{2}{\pi}} \cdot \sigma \cong \frac{4}{5} \sigma$$

17. Quartiles are given (in terms of μ and σ) by

$$Q_1 = \mu - 0.6745\sigma \text{ and } Q_3 = \mu + 0.6745\sigma$$

18. Quartile deviation (Q.D.) is given by

$$Q.D. = \frac{Q_3 - Q_1}{2} = 0.6745\sigma \cong \frac{2}{3} \sigma \text{ Also}$$

$$Q.D. = \frac{2}{3} \sigma = \frac{4}{6} \sigma = \frac{5}{6} \times \frac{4}{5} \sigma = \frac{5}{6} M.D.$$

$$\therefore Q.D. = \frac{5}{6} M.D.$$

19. We have (approximately):

$$Q.D. : M.D. : S.D. :: \frac{2}{3} \sigma : \frac{4}{5} \sigma : \sigma :: \frac{2}{3} : \frac{4}{5} : 1 \Rightarrow Q.D. : M.D. : S.D. :: 10 : 12 : 15$$

From property 18 we also have $4S.D. = 5M.D. = 6Q.D.$

20. Points of inflexion of the normal curve are at $X = \mu \pm \sigma$ i.e. they are equidistant from mean at a distance of σ and are given by :

$$X = \mu \pm \sigma, \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2}$$

21. Area property: One of the most fundamental property of the normal probability curve is the area property. If $X \sim N(\mu, \sigma^2)$, then the probability that random value of X will lie between $X = \mu$ and $X = x_1$ is given

$$P(\mu < X < x_1) = \int_{\mu}^{x_1} f(x) \cdot dx = \frac{1}{\sqrt{2\pi} \sigma} \int_{\mu}^{x_1} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$\text{put } z = \frac{x-\mu}{\sigma} \Rightarrow x = \mu + \sigma z ; \therefore \text{ at } x = \mu, z = 0; \text{ and at } x = x_1, z = \frac{x_1 - \mu}{\sigma} = z_1$$

NOTES

$$\therefore P(0 < x < x_1) = P(0 < z < z_1)$$

$$= \frac{1}{\sqrt{2\pi}} \int_0^{z_1} e^{-\frac{1}{2}z^2} dz = \int_0^{z_1} \phi(z) dz$$

Where $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$, is the probability function of standard normal variate. The definite integral $\int_0^{z_1} \phi(z) dz$, is known as **Normal Probability integral** and gives the area under standard normal curve between the ordinate $z=0$ and $z = z_1$. These areas have been provided in the form of table for different values of z_1 at the intervals of 0.01 which are available in any standard text books of statistics.

Particular Cases:

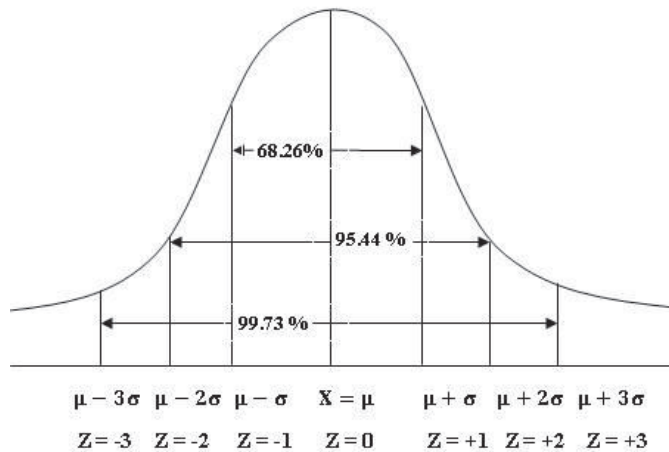
1. In particular, the probability that a random variable X lies in the interval $(\mu - \sigma, \mu + \sigma)$ is given by

$$P(\mu - \sigma < X < \mu + \sigma) = \int_{\mu - \sigma}^{\mu + \sigma} f(x) dx.$$

$$P(-1 < Z < 1) = \int_{-1}^1 \phi(z) dz = 2 \int_0^1 \phi(z) dz = 2(0.3413) = 0.6826$$

The area under the normal probability curve between the ordinates at $X = \mu - \sigma$ and $X = \mu + \sigma$ is 0.6826. In other words, the range $X = \mu - \sigma$ covers 68.26% of the observations (as shown in Fig). This is known as 1σ limit of normal distribution

This is known as 1σ limit of normal distribution



1σ, 2σ and 3σ under Normal Probability Curve

2. The probability that random variable X lies in the interval $(\mu - 2\sigma, \mu + 2\sigma)$ is given by

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = \int_{\mu - 2\sigma}^{\mu + 2\sigma} f(x) dz \Rightarrow P(-2 < Z < 2) = \int_{-2}^2 \phi(z) dz$$

$$= 2 \int_0^2 \phi(z) dz = 2(0.47725) = 0.95445$$

The area under the normal probability curve between the ordinates at

NOTES

$X = \mu - 2\sigma$ and $X = \mu + 2\sigma$ is 0.95445. In other words, the range $X = \mu + 2\sigma$ covers 95.445% of the observations (as shown in Fig.). This is known as 2σ limits of normal distribution and is considered as warning limit in case of statistical quality control which implies that it is a warning to the manufacturer that the manufacturing process is going out of control.

3. The probability that random variable X lies in the interval $(\mu - 3\sigma, \mu + 3\sigma)$ is given by

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = P(-3 < Z < 3)$$

$$\int_{-3}^3 \phi(z) dz = 2 \int_0^3 \phi(z) dz = 2(0.49865) = 0.9973$$

The area under the normal probability curve between the ordinates at $X = \mu - 3\sigma$ and $X = \mu + 3\sigma$ is 0.9973. In other words, the range $X = \mu + 3\sigma$ covers 99.73% of the observations (as shown in Fig.). This is known as 3σ limits of normal distribution and it implies the manufacturing process is out of control in case of statistical quality control.

Thus, the probability that a normal variate X lies outside the range $\mu - 3\sigma$ is given as

$$P(|X - \mu| > 3\sigma) = P(|z| > 3) = 1 - P(-3 < z < 3) = 1 - 0.9973 = 0.0027$$

Thus, in all probability, we should expect a normal variate to lie within the range $\mu - 3\sigma$ though theoretically may range from $-\infty$ to ∞ .

Example

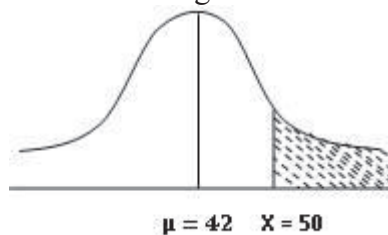
An Intelligence test was administered to 1000 students. The average score of students was 42 with standard deviation of 24. Find

- Number of students exceeding a score of 50
- Number of students scoring between 30 & 58
- Value of score exceeded by top 100 students.

Solution:

In this problem $\mu = 42$ and $\sigma = 24$ and let X denote the score obtained

- Number of students exceeding score 50



As shown in figure we want to find $P(X > 50)$ i.e. probability of shaded portion

$$\text{At } X=50, \quad Z = \frac{50-42}{24} = \frac{8}{24} = 0.334$$

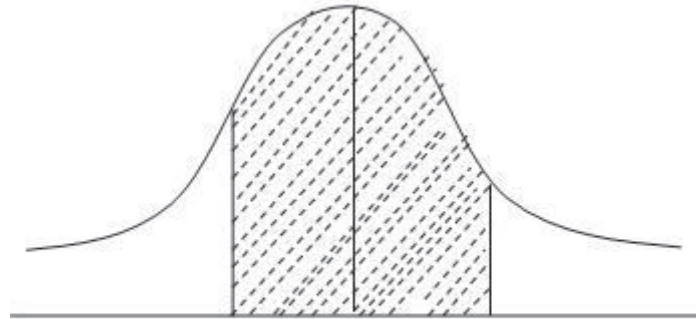
$$P(X > 50) = P(Z > 0.334) = 0.5 - P(0 \leq Z \leq 0.334) = 0.5 - 0.1308 = 0.3692$$

$$\text{No of students} = 1000 * 0.3692 = 369.2 \sim 369 \text{ students}$$

- Number of students scoring between 30 and 58

NOTES

As shown in figure we want to find $P(30 < X < 58)$ i.e. probability of shaded portion



$$\begin{matrix} X_1 = 30 & \mu = 42 & X_2 = 58 \\ Z_1 = -0.5 & Z = 0 & Z_2 = 0.6667 \end{matrix}$$

At $X_1 = 30$ $Z_1 = \frac{30 - 42}{24} = -0.5$

$P(Z_1 > -0.5) = P(0 \leq Z_1 \leq 0.5) = 0.1915$

At $X_2 = 58$ $Z_2 = \frac{58 - 42}{24} = 0.6667$

$P(Z_2 < 0.6667) = P(0 \leq Z_2 \leq 0.6667) = 0.2476$

$P(30 < X < 58) = P(-0.5 \leq Z \leq 0.6667) = 0.1915 + 0.2476 = 0.4391$

No of students = $1000 * .4391 = 439.1 \sim 439$ students

(c) Value of score exceeded by top 100 students.

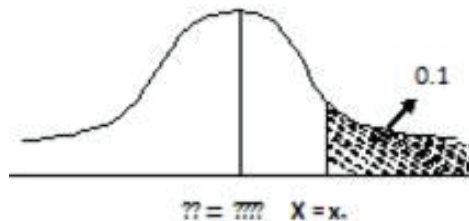
Let x_1 be the value of score exceeded by top 100 students, the probability of top 100 students = $100/N = 100/1000 = 0.1$ such that $P(X > x_1) = 0.1$

At $X = x_1$, $Z = \frac{x_1 - 42}{24} = Z_1$.

From Fig the $P(X > x_1)$ shown as shaded region

$P(X > x_1) = P(Z > Z_1) = 0.1 \Rightarrow P(0 \leq Z \leq Z_1) = 0.4 \Rightarrow \frac{x_1 - 42}{24} = 1.286$

$x_1 = 72.86 \sim 73$



Conclusion

- (a) 369 students scored more than 50.
- (b) 439 students scored between 30 & 58.

(c) Minimum score of top 100 students is 73.

14.7 SUMMARY

- Conditions for the binomial probability distribution are
 - i. The trials are independent
 - ii. The number of trials is finite
 - iii. Each trial has only two possible outcomes called success and failure.
 - iv. The probability of success in each trial is a constant.
- The parameters of the binomial distributions are n and p
- The mean of the binomial distribution is np and variance are npq
- Poisson distribution as limiting form of binomial distribution when n is large, p is small and np is finite.
- The Poisson probability distribution is $x = 0, 1, 2, 3, \dots$. Where $\lambda = np$
- The mean and variance of the Poisson distribution is λ .
- The λ is the only parameter of Poisson distribution.
- Poisson distribution can never be symmetrical.
- It is a distribution for rare events.
- Normal distribution is the limiting form of binomial distribution when n is large and neither p nor q is small

14.8 KEY WORDS

Random Variables, Binomial Distribution, Poisson Distribution, Normal Distribution

14.9 ANSWER TO CHECK YOUR PROGRESS

1. Discrete random variable, continuous random variable,
2. Mean = np , SD = \sqrt{npq} , variance = npq
3. Examples
 - The hourly number of customers arriving at a bank
 - The daily number of accidents on a particular stretch of highway
 - The hourly number of accesses to a particular web server

Probability Distribution

NOTES

Self-Instructional Material

NOTES

- The daily number of emergency calls in 108
- The number of types in a book
- The monthly number of employees who had an absence in a large company

14.10 QUESTIONS AND EXERCIS

SHORT ANSWER QUESTIONS:

1. Define Binomial distribution
2. Mention the properties of Normal distribution
3. What are the main characteristics of Poisson distribution
4. Determine the binomial distribution for which the mean is 4 and variance 3. Also find $P(X = 15)$

LONG ANSWER QUESTIONS

1. What is meant by probability distribution of a discrete random variable?
2. Define Binomial distribution? what are the main characteristics of binomial distribution
3. Write the main characteristics of normal distribution
4. Fit the Poisson distribution to the following

| | | | | | | |
|---|-----|----|----|----|---|---|
| X | 0 | 1 | 2 | 3 | 4 | 5 |
| f | 120 | 82 | 52 | 22 | 4 | 0 |

14.11 FURTHER READINGS

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. SahityaBhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., NewDelhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
7. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.

DISTANCE EDUCATION
B.B.A. DEGREE EXAMINATION, DECEMBER 2019
BUSINESS STATISTICS

Time: Three hours

Maximum: 75 marks

PART A – (10 X 2 = 20 marks)

Answer ALL questions.

1. Define statistics
2. What are the different variables used in Statistics? Give examples.
3. Find the median and mode for the weights (kgs) of 15 persons given as 58, 75, 60, 55, 61, 57, 55, 45, 70, 52, 55, 54, 60, 50, 46.
4. What is a random number? How it is useful in sampling?
5. Define standard error and mention its importance.
6. Define Null Hypothesis and Alternative Hypothesis with examples
7. Mention any four uses of Chi-square distribution in test of hypothesis.
8. Distinguish between one-way and two-way analysis of variance.
9. Define a Poisson distribution and mention its mean and variance
10. Mention the properties of a discrete probability distribution.

PART B – (5 X 5 = 25 marks)

Answer ALL questions.

11. a. Discuss the standard error of proportion.
OR
b. Describe the applications of chi square test.
12. a. Explain forecasting methods used in time series analysis.
OR
b. Briefly explain correlation coefficient between two variables.
13. a. Distinguish between correlation and regression.
OR

b. List the causes for seasonal fluctuations in a time series.

14. a. Describe chain base index number in detail

OR

b. State Bayes' theorem and mention a situation in commerce for its application.

15. a. State the conditions under which a binomial distribution becomes a normal distribution

OR

b. Compute mean, median and GM for

| Class : | 0-9 | 10-19 | 20-29 | 30-39 | 40-49 | 50-59 | 60-69 | 70-79 |
|---------|-----|-------|-------|-------|-------|-------|-------|-------|
| f | 5 | 8 | 14 | 16 | 35 | 28 | 16 | 8 |

PART C – (3 X 10 = 30 marks)

Answer ANY THREE questions

16. What is data? Explain its types in detail

17. Explain Stratified sampling technique and discuss how it is better than simple random sampling in a particular situation

18. What are the assumptions made by the regression model in estimating the parameters and in significance testing?

19. How large sample is useful in estimation and testing?

20. If X follows a normal distribution with mean 100cm and variance 25 cm, find the probabilities for (i) $X \leq 88$ (ii) $X \geq 92$, and (iii) $76 \leq X \leq 83$.

