ALAGAPPA UNIVERSITY

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

(A State University Established by the Government of Tamil Nadu)

KARAIKUDI – 630 003

Directorate of Distance Education

B.A. (Economics)

II - Semester

136 24

ELEMENTS OF STATISTICS

Authors:

JK Sharma, *Professor, Amity Business School, Amity University, Noida*
Units: (1-3, 11-13)

J.S. Chandan, *Professor, Medgar Evers College, City University of New York*
Units: (5, 10, 14)

Dr Deepak Chawla, *Distinguished Professor, Dean, International Management Institute (IMI), New Delhi*
Dr Neena Sondhi, *Professor, International Management Institute (IMI), New Delhi*
Units: (4, 6-9)

Vikas® is the registered trademark of Vikas® Publishing House Pvt. Ltd.

VIKAS® PUBLISHING HOUSE PVT. LTD.
E-28, Sector-8, Noida - 201301 (UP)
Phone: 0120-4078900 • Fax: 0120-4078999
Regd. Office: 7361, Ravindra Mansion, Ram Nagar, New Delhi 110 055
• Website: www.vikaspublishing.com • Email: helpline@vikaspublishing.com

Work Order No. AU/DDE/DE1-291/Preparation and Printing of Course Materials/2018 Dated 19.11.2018 Copies - 500

SYLLABI-BOOK MAPPING TABLE

Elements of Statistics

Syllabi	Mapping in Book
BLOCK I: BASIC STATISTICS AND INDEX NUMBERS **Unit - 1:** Statistics: Definition - Nature - Scope - Role and Importance of Statistics. **Unit - 2:** Index Numbers: Definition - Uses - Problems in Construction - Methods - Simple and Weighted. **Unit - 3:** Index Numbers in Economics: Laspeyer's and Paache's Index Numbers - Fishers Ideal Index Number - Marshall and Edgeworths Index Numbers.	**Unit 1:** Statistics: An Overview **(Pages 1-13);** **Unit 2:** Index Numbers **(Pages 14-28);** **Unit 3:** Index Numbers in Economics **(Pages 29-41)**
BLOCK II: CENSUS AND SAMPLING **Unit - 4:** Census and Sampling: Meaning - Features - Population and Sample. **Unit - 5:** Sampling: Meaning - Types of Sampling. **Unit - 6:** Sampling Design: Meaning - Types - Challenges. **Unit - 7:** Design of Questionnaire. **Unit - 8:** Sampling Errors.	**Unit 4:** Census and Sampling: An Introduction **(Pages 42-49);** **Unit 5:** Sampling: Meaning and Types **(Pages 50-55);** **Unit 6:** Sampling Design: Meaning, Types and Challenges **(Pages 56-72);** **Unit 7:** Design of Questionnaire **(Pages 73-93);** **Unit 8:** Sampling Errors **(Pages 94-100)**
BLOCK III: COLLECTION AND TABULATION OF DATA **Unit-9:** Collection of Data: Meaning - Types of Data: Primary and Secondary - Qualitative and Quantitative. **Unit - 10:** Tabulation of Data: Meaning - Objectives - Classification of Tabulation - Types of Tables - Presentation of Tables.	**Unit 9:** Collection of Data **(Pages 101-121);** **Unit 10:** Tabulation of Data **(Pages 122-139)**
BLOCK IV: MEASURES OF CENTRAL TENDENCY, DISPERSION AND DIAGRAMMATICS **Unit - 11:** Measures of Central Tendency: Characteristics - Median - Mode - Harmonic Mean - Geometric Mean - Simple Problems. **Unit - 12:** Measures of Dispersion-I: Features - Quartile Deviation - Mean Deviation - Standard Deviation - Its usefulness. **Unit - 13:** Measures of Dispersion-II: Range - Quartiles - Deciles - Percentiles - Characteristics - Simple Problems. **Unit - 14:** Diagrammatic and Graphic Representation - Bar Diagrams - Pie Diagrams - Histograms - Pictograms - Cartograms - Frequency Graphs - Ogives – LorenzCurve.	**Unit 11:** Measures of Central Tendency **(Pages 140-176);** **Unit 12:** Measures of Dispersion-I **(Pages 177-209);** **Unit 13:** Measures of Dispersion-II **(Pages 210-221);** **Unit 14:** Diagrammatic and Graphic Representation of Data **(Pages 222-240)**

CONTENTS

BLOCK II: CENSUS AND SAMPLING

INTRODUCTION

NOTES

Uncertainty and risk taking are inherent aspects of any business. In modern times, decisions are determined by data. In all aspects of modern-day businesses, a substantial amount and variety of data is available for inspection and analysis. Business managers and executives are increasingly required by the top management to justify the decisions taken by them based on statistical data. They need statistical decision-making systems. These systems enable them to collect, analyse and interpret data that is relevant to their decision making. Concepts pertaining to business statistics concepts and thinking enable managers and academicians to achieve the following:

- Solve a diverse range of problems
- Add value to decisions
- Reduce guesswork

Business statistics and its elements can be defined as the quantitative way to make sound decisions in the wake of uncertainty in business. Business statistics tools are used in many disciplines such as financial analysis, economics, audit, production and operations management and market research. Business statistics provides knowledge and the ability to interpret and use statistical techniques in a variety of business applications.

This book, *Elements of Statistics*, is an attempt to provide a fair idea of the concepts of statistical analysis and their applications in the business world. The book is divided into fourteen units and so designed as to provide an analytical framework for understanding all aspects of statistical analysis in the contemporary complex and dynamic business environment.

This book has been designed keeping in mind the self-instructional mode or SIM format, wherein each unit begins with an 'Introduction' to the topic and is followed by an outline of the 'Objectives'. The detailed content is then presented in a simple and structured from, interspersed with 'Check Your Progress' questions to test the student's understanding. A 'Summary' of the content, along with a list of 'Key Words' and a set of 'Self-Assessment Questions and Exercises' is provided at the end of each unit for effective recapitulation. Relevant examples/illustrations have been included for better understanding of the topics.

BLOCK - I
BASIC STATISTICS AND INDEX NUMBERS

NOTES

UNIT 1 STATISTICS: AN OVERVIEW

Structure

1.0 INTRODUCTION

Statistics refers to a mathematical science which includes methods of collecting, organizing and analysing data in a way that results in the meaningful conclusions. A better understanding of statistics is very important for anyone running a business. The unit is all about the fundamentals of statistics, its nature and types. Also, there are different definitions of statistics which help you to easily comprehend the concept of statistics. The unit goes on discussing the reasons for learning statistics, the growth and development of statistics and types of statistics. The concept and significance of statistical thinking and analysis has also been covered in the unit. In addition to this, the importance of statistics in different disciplines has also been discussed in detail in the unit.

NOTES

1.1 OBJECTIVES

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

- Discuss the definition and nature of statistics
- Describe the different types of statistical methods
- Explain the importance and scope of statistics
- Understand the growth and development of statistics

1.2 REASONS FOR LEARNING STATISTICS

H. G. Wells' statement *statistical thinking will one day be as necessary as the ability to read and write* is valid in the context of today's competitive business environment where many organizations find themselves data rich but information poor. Thus, for decision makers, it is important to develop the ability to extract meaningful information from raw data to make better decisions. It is possible only through the careful analysis of data guided by statistical thinking.

The reason for analysis of **data** is to understand the *variation and its causes* in any phenomenon or process. As such knowledge helps in producing valuable data about a phenomenon or a process, and hence leads to better decisions. It is from this perspective that the knowledge of statistical techniques enables the decision maker to:

- Summarize and describe information (data) more precisely to understand the process at a glance.
- Capture a population's characteristics by making inferences from a sample's characteristic.
- Understand the nature of relationship between pair of variables in a process to improve its functioning.
- Make reliable forecasts of certain events of interest.

Thus, knowledge of statistical techniques or methods should enable us to gain insight into an unknown situation or to produce sophisticated analysis for numerical confirmation or a reflection of some widely held belief.

1.2.1 Growth and Development of Statistics

The views about **statistics** are numerous and have different meanings depending largely on its use: (i) for a cricket fan, statistics refers to numerical information or data relating to the runs scored by a cricketer; (ii) for an environmentalist, statistics refers to information on the quantity of pollutants

released into the atmosphere by all types of vehicles in different cities; (iii) for the census department, statistics refers to information about the birth rate and the sex ratio in different states; (iv) for a share broker, statistics refers to information on changes in share prices over a period of time; and (v) for a common person, statistics refers to increase and/or decrease in per capita income, wholesale price index, industrial production, exports, imports, crime rate and so on.

The secondary sources of such information or data are newspapers, magazines journals, reports/bulletins, radio, television and so on. In all such cases, the relevant data are collected and presented with the help of figures, charts, diagrams and pictograms. To understand and find a solution (with certain degree of precision) to problems pertaining to social, political, economic and cultural activities is, of course, unending but possible to an extent by the use of *Statistical Methods or Techniques*.

The development of fast-speed computers and use of computer software, such as Statistical Analysis System (SAS) and Statistical Product and Service Solutions (SPSS), have substantially changed the scope of application of statistical methods towards solving real-life problems. The increasing use of spreadsheet packages like Lotus 1-2-3 and Microsoft Excel have led to the incorporation of statistical features in these packages.

1.2.2 Statistical Thinking and Analysis

The objective of any organization is providing quality products or services to its customers. This objective requires statistical thinking by the management of the organization. *Statistical thinking can be defined as the thought process that focuses on ways to identify, control and reduce variations present in all phenomena or processes.* This approach helps to recognize and make interpretations of the variations in a phenomenon or a process through data analysis and, hence, enhances opportunities for improvement in the quality of products or services.

A *phenomenon or process (activity)* is a set of conditions that repeatedly come together to transform inputs into outcomes such as a business process to serve customers, length of time to complete a banking transaction, manufacturing of goods, resolution of customer complaint and so on.

The quality improvement process is comprised three factors: (i) *management philosophy*, (ii) *behavioral tools* and (iii) *statistical methods*. While management philosophy works as catalyst for laying a foundation for total quality improvement efforts, the use of behavioral tools such as brainstorming and team building, and statistical methods such as control charts and descriptive statistics, are also necessary for understanding and improving phenomena or processes.

The steps of statistical thinking necessary for understanding and improvement in the phenomena or processes are summarized in Fig.1.1.

NOTES

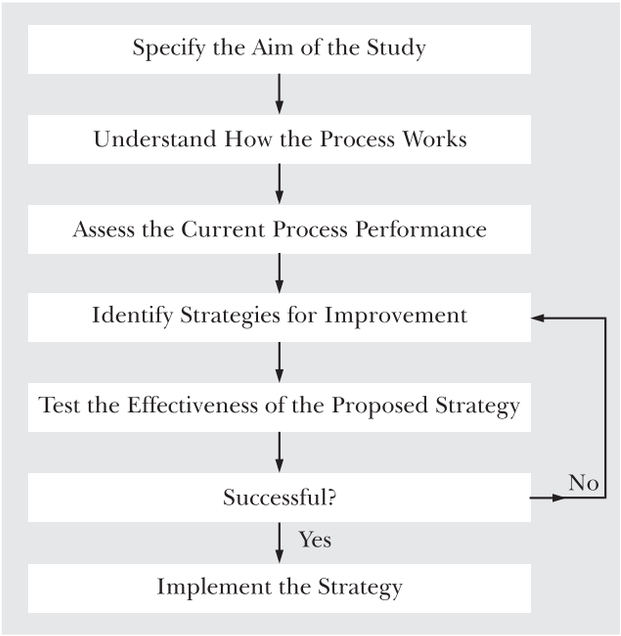


Fig. 1.1 *Flow Chart of Process Improvement*

1.3 DEFINITION AND NATURE

The word statistics refers to a collection of procedures and principles useful for gathering and analysing numerical information, called statistical data or simply data, for drawing conclusions and making decisions.

A few definitions which describe the characteristics of statistics are as follows:

- The classified facts respecting the condition of the people in a state... especially those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement.

—Webster

This definition confines the scope of statistics only to such facts and figures that are related to the conditions of the people in a state.

- By statistics we mean quantitative data affected to a marked extent by multiplicity of causes.

—Yule and Kendall

NOTES

- By statistics we mean aggregates of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated, or estimated according to reasonable standards of accuracy, collected in a systematic manner for predetermined purpose and placed in relation to each other. —Horace Secrist
- Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry.

  —Seligman
- The science of statistics is the method of judging, collecting natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates. —King
- Statistics may be called the science of counting; science of average as well as science of the measurement of social organism regarded as a whole in all its manifestations.

  — A.L.Bowley
- Statistics may be defined as a science of collection, presentation, analysis and interpretation of numerical data. — Croxton and Cowden

This definition has pointed out four stages of statistical investigation to which one more stage 'organization of data' should also to be added. Accordingly, statistics may be defined as the science of collecting, organizing, presenting, analysing and interpreting numerical data for making better decisions.

There are two branches of statistics:

(i) Mathematical statistics

(ii) Applied statistics

Mathematical statistics aims to obtain the knowledge of an experiment or theoretical nature that has direct or immediate impact on action, performance or policy decisions. Whereas, applied statistics uses statistical theory in formulating and solving problems in real life. In applied statistics, the rules necessary to solve a particular problem are not always obvious, although the guiding principles at the back of various methods are identical regardless of the field of their application.

Check Your Progress

1. How does the knowledge of statistical techniques help us?
2. Define statistical thinking.
3. Name the two branches of statistics.

1.4 TYPES OF STATISTICAL METHODS

NOTES

Statistical methods, broadly, fall into the following two categories:

(i) Descriptive statistics

(ii) Inferential statistics

Descriptive statistics includes statistical methods that are used for collecting, presenting, depicting the center, spread and shape of the data array. Hence, these methods are helpful as preliminary tools to describe the various features/characteristics of data array.

In general, methods of descriptive statistics include graphic methods and numeric measures. Bar charts, line graphs and pie charts comprise the graphic methods, whereas numeric measures include measures of central tendency, dispersion, skewness and kurtosis.

Inferential statistics includes statistical methods that are used for estimation of population characteristics on the basis of sample results and testing of statistical hypothesis.

Sample and population are two relative terms. A population or universe is the collection of elements (such as employees in a company, students in a university/college, companies, voters, households, customers, manufactured items, births and deaths, road accidents, etc.) about which we wish to make some inference. The *population element* is the individual unit or object on which the measurement is taken. A population can be *finite* or *infinite* according to the number of elements under statistical investigation. A sample is a fraction, subset or portion of that universe.

Inferential statistics can be categorized as *parametric* or *non-parametric*. The parametric statistical methods are used to draw inference about a population for which the sample is drawn on an interval or a ratio scale and population is normally distributed. Non-parametric statistical methods are used to draw inference about a population for which the sample is drawn on nominal or ordinal scale and population need not be normally distributed.

When the size of population is very large, we need to draw sample of predetermined size due to (i) lower cost; (ii) greater accuracy of results; and (iii) saving of time for data collection. The analyses of the elements of the sample reflect the characteristics of the population from which the sample is drawn.

Illustration: A manufacturer who produces electric bulbs wants to learn the average life of bulbs. For this, he selects a sample of bulbs at regular intervals of time and measures the life of each. If the sample average does not fall

within the specified range of variations, the process controls are checked and suitable actions are taken. In this example, all the bulbs being produced by the manufacturing process represents the population, the statistical variable is the life of bulb, statistic is the average life of bulbs in a given sample; parameters of interest are the average life and variation in life span among manufactured bulbs, and sampling units are the bulbs selected for the sample.

1.5 IMPORTANCE, ROLE AND SCOPE OF STATISTICS

Statistical methods are applicable in diversified fields such as economics, trade, industry, commerce, agriculture, bio-sciences, physical sciences, education, insurance, sociology, psychology and so on. Carrol D. Wright (1887), United States Commissioner of the Bureau of Labour, has explained the importance of statistics by saying:

> To a very striking degree our culture has become a statistical culture. Even a person who may never have heard of an index number is affected by those index numbers which describe the cost of living. It is impossible to understand Psychology, Sociology, Economics or a Physical Science without some general idea of the meaning of an average, of variation, of concomitance of sampling, of how to interpret charts and tables.

According to the statistician Bowley, a knowledge of statistics is like a knowledge of foreign language or of algebra, it may prove of use at any time under any circumstances.

The importance of statistics in a few important disciplines is briefly discussed below.

1.5.1 Statistics and State

For efficient governance, any State Government collects the large amount of statistics relating to prices, production, consumption, income and expenditure, investments and profits. Statistical methods such as time-series analysis, index numbers, forecasting and demand analysis are extensively practiced in formulating economic policies. Data are also collected on population dynamics in order to initiate and implement various welfare policies and programmes within a state or for whole country.

Beside statistical bureaus in all ministries and government departments in the Central and State Governments, other organizations/departments such as Central Statistical Organization (CSO), National Sample Survey Organization (NSSO) and the Registrar General of India (RGI), regularly collect data for the purpose of analysing effectiveness of various policies of government.

NOTES

1.5.2 Statistics in Economics

Various statistical methods are extensively used to derive empirical results for setting economic policies. Few examples are given below:

- Time-series analysis is used for studying the fluctuation in prices, production and consumption of commodities, savings, bank deposits, money landing to industry, etc.
- Index numbers are used for economic planning to indicate the changes over a specified period of time in (a) prices of commodities, (b) imports and exports, (c) industrial/agricultural production and (d) cost of living, and the like.
- Demand analysis is used to study the relationship between the price of a commodity and its output (supply).
- Forecasting techniques are used for curve fitting by the principle of least squares and exponential smoothing to predict inflation rate, unemployment rate or manufacturing capacity utilization.

1.5.3 Statistics in Business Management

According to Wallis and Roberts, statistics may be regarded as a body of methods for making wise decisions in the face of uncertainty. Ya-Lin-Chou gave a modified definition over this, saying that statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risks. These definitions reflect the applications of statistics in the development of general principles for dealing with uncertainty.

Statistical reports provide a summary of business activities which improves the decision makings in future. Certain activities where statistics plays an important role are discussed below:

- **Marketing:** Before a product is launched, market research team, through a pilot survey, makes use of various techniques of statistics to analyse data on population, purchasing power, habits of the consumers, competitors, pricing, etc. Such studies reveal the possible market potential for the product.

  Analysis of sales volume in relation to the purchasing power and concentration of population is helpful in establishing sales territories, routing of salesmen and advertising strategies to improve sales.
- **Production:** Statistical quality control techniques are used for improvement in the quality of the existing products and setting quality control standards for new ones. Make or buy decisions are based on statistically analysed data.
- **Finance:** The correlation analysis of profits and dividends helps organizations to predict and decide probable dividends in coming

years. Statistics tools are also used to analyse data on assets and liabilities, income and expenditure, and investment decisions under uncertainty to understand the financial results of various operations.

- **Personnel:** Statistical studies of wage rates, incentive plans, cost of living, attrition rates, employment trends, accident rates, performance appraisal, need for training and development, etc., helps Human Resource Managers in man power planning and designing welfare policies.

  The impact of various factors such as wages, grievances handling, welfare, delegation of authority, education and housing facilities, and training and development on employer–employee relationship can also be studied through statistical techniques such as multiple regression analysis and factor analysis.

1.5.4 Statistics in Physical Sciences

Statistical methods such as sampling, estimation and design of experiments are used effectively in most of physical sciences such as astronomy, engineering, geology, meteorology and certain branches of physics.

1.5.5 Statistics in Social Sciences

The following definitions reflect the importance of statistics in social sciences.

- Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations. —Bowley
- The science of statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis, enumeration or collection of estimates. —W. I. King

Few areas of applications of statistics in social sciences are as listed below:

- Regression and correlation analysis techniques are used to study those factors which bring out changes with respect to time, place and object.
- Sampling techniques are used for study pertaining to any strata of society, and drawing inferences.
- Statistical methods are also used to study mortality (death) rates, fertility (birth rates) trends, population growth and other aspects of vital statistics.

1.5.6 Statistics in Medical Sciences

For proper diagnosis of a disease, a doctor requires data relating to pulse rate, body temperature, blood pressure, heart beats and body weight.

The efficacy of a particular drug or injection meant to cure a specific disease may be verified using chi-square technique. Comparative studies for effectiveness of a particular drug/injection manufactured by different

companies can also be made by using statistical techniques such as the *t*-test and *F*-test.

1.5.7 Statistics and Computers

NOTES

Computers and information technology facilities such as Excel sheet, Microsoft Word and common statistical software such as SPSS, SAS and LINDO have made data analysis readily available to any business decision-maker. Past records of operations involving payroll calculations, inventory management, railway/airline reservations and money transaction can be done with the help of a computer.

Check Your Progress

4. What are the two types of statistical methods?
5. Name the different fields where statistical methods are applied.
6. How are statistical methods used in physical science?

1.6 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The knowledge of statistical techniques or methods should enable us to gain insight into an unknown situation or to produce sophisticated analysis for numerical confirmation or a reflection of some widely held belief.
2. Statistical thinking can be defined as the thought process that focuses on ways to identify, control and reduce variations present in all phenomena or processes.
3. The two branches of statistics are: Mathematical statistics and applied statistics.
4. There are two types of statistical methods: Descriptive statistics and inferential statistics.
5. Statistical methods are applicable in diversified fields such as economics, trade, industry, commerce, agriculture, bio-sciences, physical sciences, education, insurance, sociology, psychology and so on.
6. Statistical methods such as sampling, estimation and design of experiments are used effectively in most of physical sciences such as astronomy, engineering, geology, meteorology and certain branches of physics.

1.7 SUMMARY

NOTES

- The knowledge of statistical techniques or methods should enable us to gain insight into an unknown situation or to produce sophisticated analysis for numerical confirmation or a reflection of some widely held belief.
- Statistical thinking can be defined as the thought process that focuses on ways to identify, control and reduce variations present in all phenomena or processes.
- The quality improvement process is comprised three factors: (i) *management philosophy*, (ii) *behavioral tools* and (iii) *statistical methods*.
- The word statistics refers to a collection of procedures and principles useful for gathering and analysing numerical information, called statistical data or simply data, for drawing conclusions and making decisions.
- The two branches of statistics are (i) Mathematical statistics (ii) Applied statistics
- Mathematical statistics aims to obtain the knowledge of an experiment or theoretical nature that has direct or immediate impact on action, performance or policy decisions.
- The Applied statistics uses statistical theory in formulating and solving problems in real life. In applied statistics, the rules necessary to solve a particular problem are not always obvious, although the guiding principles at the back of various methods are identical regardless of the field of their application.
- Statistical methods, broadly, fall into the following two categories: Descriptive statistics and inferential statistics.
- Descriptive statistics includes statistical methods that are used for collecting, presenting, depicting the center, spread and shape of the data array. Inferential statistics includes statistical methods that are used for estimation of population characteristics on the basis of sample results and testing of statistical hypothesis.
- Inferential statistics includes statistical methods that are used for estimation of population characteristics on the basis of sample results and testing of statistical hypothesis.
- Statistical methods are applicable in diversified fields such as economics, trade, industry, commerce, agriculture, bio-sciences, physical sciences, education, insurance, sociology, psychology and so on.

NOTES

- All ministries and government departments in the Central and State Governments, other organizations/departments such as Central Statistical Organization (CSO), National Sample Survey Organization (NSSO) and the Registrar General of India (RGI), regularly collect data for the purpose of analysing effectiveness of various policies of government.
- According to Wallis and Roberts, statistics may be regarded as a body of methods for making wise decisions in the face of uncertainty. Ya-Lin-Chou gave a modified definition over this, saying that statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risks.
- Statistical methods such as sampling, estimation and design of experiments are used effectively in most of physical sciences such as astronomy, engineering, geology, meteorology and certain branches of physics.
- According to Bowley, statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations.
- Computers and information technology facilities such as Excel sheet, Microsoft Word and common statistical software such as SPSS, SAS and LINDO have made data analysis readily available to any business decision-maker.

1.8 KEY WORDS

- **Statistics:** It refers to the practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.
- **Statistical thinking:** It refers to the process of using wide ranging and interacting data to understand processes, problems, and solutions.
- **SAS:** SAS or Statistical Analysis System is a software suite developed by SAS Institute for advanced analytics, multivariate analysis, business intelligence, data management, and predictive analytics.

1.9 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. Give the reasons of learning statistics.
2. List the three factors of quality improvement process.
3. How did Webster and Horace Secrist define statistics?

4. Write short notes on the following:
   (a) Mathematical Statistics
   (b) Applied Statistics
   (c) Descriptive Statistics
   (d) Inferential Statistics

NOTES

Long Answer Questions

1. How does statistical thinking improve business process? Explain.
2. Evaluate the growth and development of statistics.
3. What is the importance of statistics in a few important disciplines? Explain.

1.10 FURTHER READINGS

Chandan, J.S., Jagjit Singh and K.K. Khanna. 1995. *Business Statistics*, 2nd edition. New Delhi: Vikas Publishing House Pvt. Ltd.

Croxton, Frederick E., and Dudley J. Cowden. 1943. *Applied General Statistics*. New York: Prentice Hall.

Gupta, S.P. 2006. *Statistical Methods*. New Delhi: S. Chand & Company.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods*, 23rd edition. New Delhi: Vikas Publishing House Pvt.Ltd..

Levin, Richard I. and David S. Rubin. 1998. *Statistics for Management.* New Jersey: Prentice Hall.

UNIT 2 INDEX NUMBERS

NOTES

Structure

2.0 INTRODUCTION

Often an individual is interested to know, how much the prices of essential household items have increased or decreased so that necessary adjustments can be made in the monthly budget. Similarly, an organization may be concerned with the way in which prices paid for raw materials, annual income and profit, commodity prices, share prices, production volume, advertising budget, wage bills, and so on have changed over a period of time. However, while prices of a few items may have increased, others may have decreased over a given period of time. Consequently, an average measure needs to be defined to compare and describe such differences from one time period to another.

An *index number* is defined as a relative measure to compare and describe the average change in price, quantity or value of an item over a period of time in relation to its value at some fixed point in time, called the *base period*. The ratio of the current price, quantity or value to a base price, quantity or value is multiplied by 100 to express the index in terms of percentage. Since an index number is constructed as a ratio of an average change in price, quantity or value of an item over a period of time in relation to its value at some fixed point in time, therefore it has no unit of measurement and is expressed in terms of percentage as follows:

$$\text{Index number} = \frac{\text{Current period value}}{\text{Base period value}} \times 100$$

NOTES

The base period for indexes may be based at any convenient period, which is occasionally adjusted to a convenient period, and these are published at any convenient frequency. Examples of some indexes are as follows:

Daily	Stock market prices
Monthly	Unemployment figures
Yearly	Gross National Product (GNP)

For decision-making in business, it is essential to understand different published indexes and to construct one's own index. This index can be compared with a national and/or with competitor's. For example, a cement company could construct an index of its own sales and production volumes and compare it to the index of the cement industry. A graph of two indexes will provide, at a glace, a view of a company's performance within the industry as shown in Fig. 2.1.

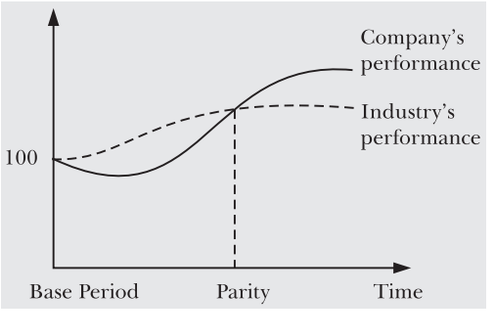


Fig. 2.1 *Graph of Two Indexes*

2.1 OBJECTIVES

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

- Understand the definition of index numbers
- Discuss the different types of price index numbers
- Describe the characteristics and uses of index numbers
- Explain the problems and methods in construction of index numbers

2.2 DEFINITION OF INDEX NUMBERS

According to Irving Fisher, Index number, almost alone in the domain of social sciences, may truly be called an exact science, if it be permissible to

NOTES

designate as science the theoretical foundations of a useful art. Definition of index numbers can be classified into the following three broad categories:

A. Measure of Change

- It is a numerical value characterizing the change in complex economic phenomena over a period of time or space. —Maslow
- An index number is a quantity which, by reference to a base period, shows by its variations the changes in the magnitude over a period of time. In general, index numbers are used to measure changes over time in magnitudes which are not capable of direct measurement.

  —John I. Raffin
- An index number is a statistical measure designed to show changes in variables or a group of related variables with respect to time, geographic location or other characteristics. —Speigel
- Index number is a single ratio (usually in percentages) which measures the combined (i.e., averaged change of several variables between two different times, places or situations.

  —A. M. Tuttle

B. Device to Measure Change

- Index numbers are devices measuring differences in the magnitude of a group of related variables. —Croxton and Cowden
- An index number is a device which shows by its variation the changes in a magnitude which is not capable of accurate measurement in itself or of direct valuation in practice. —Wheldon

C. Series Representing the Process of Change

- Index numbers are series of numbers by which changes in the magnitude of a phenomenon are measured from time to time or place to place.

  —Horace Secrist
- A series of index numbers reflects in its trend and fluctuations the movements of some quantity of which it is related. —B. L. Bowley
- An index number is a statistical measure of fluctuations in a variable arranged in the form of a series, and using a base period for making comparisons.

  —L. J. Kaplan

2.2.1 Types of Price Index Numbers

Index numbers are broadly classified into three categories: (i) price index, (ii) quantity index and (iii) value index. A brief description of each of these is as follows:

Price Index

NOTES

The price indexes are of two categories:

(i) Single price index

(ii) Composite prices index

The single price index measures the percentage change in the current price per unit of a product to its base period price. To facilitate comparisons with other years, the actual price per unit is converted into a ***price relative to*** express price per unit in each period as a percentage of price per unit in a base period. Price relatives are helpful to understand and interpret changing economic and business conditions over time. Calculations of price relatives with base year 1996 are shown in Table 2.1.

Table 2.1 *Calculation of Price Index (Base year = 2000)*

Year	*Total Wage Bill (₹ millions)*	*Ratio*	*Price Index or Percentage Relative*
(1)	*(2)*	*(3) = (2)/11.76*	*(4) = (3) × 100*
2000	11.76	11.76/11.76 = 1.000	100.0
2001	12.23	12.23/11.76 = 1.039	103.9
2002	12.84	12.84/11.76 = 1.091	109.1
2003	13.35	13.35/11.76 = 1.135	113.5
2004	13.82	13.82/11.76 = 1.175	117.5

From Table 2.1, it is observed that the price relative of 117.5 in 2004 shows an increase of 17.5 per cent in wage bill compared to the base year 2000.

A **composite price index** measures the average price change for a group of items taken as a whole from a base period to the current period. For example, the wholesale price index reflects the general price level for a group of items taken as a whole.

The ***retail price index*** indicates the general changes in the retail prices for a group of items including food, housing, clothing and so on. The consumer price index, a special type of retail price index, is the primary measure of the cost of living in a country. The consumer price index is a weighted average price index with fixed weights. The weightage applied to each item in the basket of items is derived from the urban and rural families.

Quantity Index: A quantity index indicates the relative changes in quantity levels of a group of items such as agricultural and industrial production, imports and exports, consumed (or produced) between current and base periods. The method of constructing quantity indexes is the same as that of price index except that the quantities vary from period to period.

The two most common quantity indexes are the weighted relative of aggregates and the weighted average of quantity relative.

NOTES

Value Index: A value index indicates the relative changes in total monetary value (worth) of an item, such as inventories, sales, or foreign trade, between current and base period. The value of an item is determined by multiplying its unit price by the quantity consumed (or produced). For example, comparative cost of living in terms of cost of goods and services shows whether living in a small city is cheaper than in metro cities.

Special Purpose Indexes

A few index numbers such as industrial production, agricultural production and productivity can also be constructed separately depending on the nature and degree of relationship between groups of items.

Check Your Progress

1. How has John I Raffin defined index numbers?
2. State the two categories of price index.

2.3 CHARACTERISTICS AND USES OF INDEX NUMBERS

The following are the characteristics and uses of index numbers:

2.3.1 Characteristics of Index Numbers

(i) **Specialized averages:** According to R. L. Corner, *an index number represents a special case of an average, generally weighted average, compiled from a sample of items judged to be representative of the whole.*

(ii) Index numbers can be used for comparing of two or more data sets expressed in different units of measurement. For example, consumer price index is used for comparing price for a group of items such as food, clothing, fuel, house rent and so on that are expressed in different units of measurement.

(iii) **Measure change in the value of a variable:** Index numbers represent an increase or a decrease (expressed in terms of percentage) in the value of a variable. For example, a quantity index number = 110 for cars sold in a given year when compared with that of a base year indicates that the cars sales in the given year were 10 per cent higher than in the base year (value of index number in base period is always equal to 100).

(iv) Also according to Bowley, *index numbers are used to measure the changes in some quantity which we cannot observe directly...*'. For example, cost of living cannot be measured in quantitative terms

directly because changes in it can only be studied by knowing variations in certain other related factors.

(v) Measure effect of changes with respect to time or place**:** Index numbers are helpful in comparing changes between locations, and in categories over periods of time. For example, since cost of living may be different at two different cities or town places, it can be compared over periods of time.

NOTES

2.3.2 Uses of Index Numbers

According to G. Simpson and F. Kafka, *today Index numbers are one of the most widely used statistical tools. They are used to feel the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies*. Other uses of index number are as follows:

(i) Act as economic barometers**:** Index numbers are used to measure general economic condition of a country. For example, the composite index number of prices, industrial output, foreign exchange reserves and bank deposits act as an economic barometer.

(ii) Help in policy formulation: The price index indicates changes in various segments of the economy. For example, by examining the price index of different operations of a company, the need for some remedial or preventive actions can be assessed. Similarly, by examining the population index, the need to formulate a policy for health and education can be assessed.

(iii) Reveal trends and tendencies: Since an index number describes an average change in the level of a variable between the current period and a base period, therefore it can be used to reflect typical patterns of change in the level of the variable. For example, based on the index number of industrial production, agricultural production, imports, exports, and wholesale and retail prices for the last few years, decision makers can draw the trend of variations to know how much change has taken place due to the various factors.

(iv) Help to measure purchasing power: Since purchasing power is related to a group of people or class rather than a particular individual or cost of a single item. Consequently, earnings of a group of people or class must be adjusted with a price index to understand an overall view of the purchasing power for the group.

For example, suppose a person earns ₹1000 per month in the year 2010. If an item costs ₹100 in that year, he could purchase 1000/100 = 10 units of the item with one month's earnings, but if in year 2011, the same person earns ₹2000 per month but the item cost is ₹250, then he could purchase 2000/ 250 = 8 units of the item. Hence, effect of monthly earning relative to the particular item is less in year 2011 than

NOTES

in 2010. The effect of price is eliminated by dividing the price of the item in both the years, For instance, in 2010, the purchasing power was 10/ 1000 = 0.01 or ₹10 that was 250/2000= 0.125 or ₹12.5 in 2011.

(v) Help in deflating various values: Adjustment of current rupee value to real terms is referred to as *deflating a value series* because prices increase over time. The price index is helpful in deflating the national income to remove the effect of inflation over a long period of time so as to understand whether there is any change in the real income of the people or not. Since retail price index compares the purchasing power of money at different points in time, it is often used to compute changes in earnings and expenditure.

2.4 PROBLEMS AND METHODS IN CONSTRUCTION

Various price indexes and their methods of construction can be classified into broad categories as shown in Fig. 2.2.

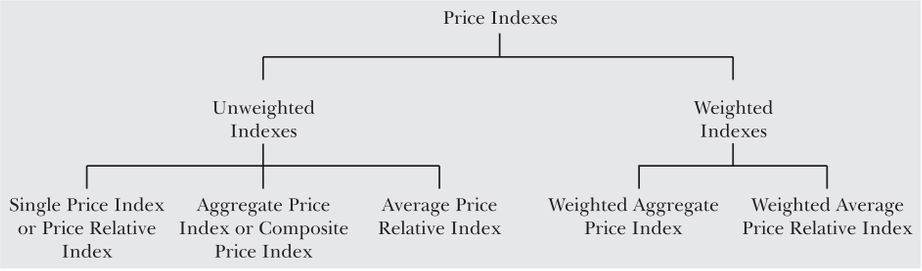


Fig. 2.2 *Classification of Index Numbers*

2.4.1 Unweighted Price Index

An *unweighted price index number measures the percentage change in price for a single item or a basket of items between any two time periods*. Unweighted implies that all the values considered in calculating the index are of equal importance.

An unweighted price index is calculated by dividing the price of an item in the given period by the price of the same item in the base period. To facilitate comparison with other years, the price of an item can be converted into a *price relative* so as to express the unit price in each year (period) in terms of percentage of the unit price in the base year.

The general formula for calculating unweighted price index or price relative index is

$$\text{Unweighted price index in period, } n = \frac{p_n}{p_0} \times 100$$

where p_n is the price per unit of an item in the nth year and p_0 is the price per unit of an item in the base year.

Example 2.1: The retail price of a commodity over a period of four years is given below:

Year	2009	2010	2011	2012
Price (₹)	24.60	25.35	26.00	26.50

(a) Find the price index based on 2009 prices.

(b) Find the percentage change in price between consecutive years (base year = 2000).

(c) Find the percentage increase between consecutive years.

Solution: (a) For the price of the commodity in the base year 2009, the price relatives for one unit of the commodity in the years 2009 to 2012 are given in Table 2.2.

(b)

Table 2.2 *Price Relatives*

Year	*Price* (₹)	*Price Relatives*	*Percentage Change*
2009	24.60	100	—
2010	25.35	$\frac{25.35}{24.60} \times 100 = 103.04$	3.04
2011	26.00	$\frac{26}{24.60} \times 100 = 105.69$	2.65
2012	26.50	$\frac{26.50}{24.60} \times 100 = 107.72$	1.92

(c) The percentage change in price relative is obtained as follows:

For year 2010: $\frac{103.04 - 100}{100} \times 100$ = 3.04 per cent

For year 2011: $\frac{105.69 - 103.04}{103.04} \times 100$ = 2.57 per cent

For year 2012: $\frac{107.72 - 105.69}{105.69} \times 100$ = 1.92 per cent

2.4.2 Aggregate Price Index

An aggregate (or composite) price index measures the average change in price for a group of related items in *the current period with respect to the base period.* For example, the change in the cost of living over a period of time is based on the price changes for a variety of commodities such as food, housing, clothing, transportation, health care and so on. If items or commodities chosen are large in number, then a sample of such items or commodities may be selected for calculating an aggregate price index.

NOTES

Method to calculate an aggregate price index irrespective of the units of measurement of price of commodities is summarized as follows:

1. Add the unit prices of a group of commodities in the year of interest.
2. Add the unit prices of a group of commodities in the base year.
3. Divide the sum obtained in Step 1 by the sum obtained in Step 2, and multiply the quotient by 100.

The formula of calculating an unweighted aggregate price index is as follows:

Aggregate price index, $$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 \quad (2.1)$$

where p_1 is the unit price of a commodity in the current period of interest and p_0 is the unit price for a commodity in the base period.

Example 2.2: The following are two sets of retail prices of a family's shopping basket. The data pertain to retail prices during 2009 and 2010.

Commodity	*Unit Price (₹)*	
	2009	*2010*
Milk (1 litre)	30	35
Eggs (1 dozen)	20	22
Butter (1 kg)	150	170
Bread (500 gm)	12	14

Calculate the unweighted aggregate price index for 2010 using 2009 as the base year.

Solution: Calculations for unweighted aggregate price index are shown in Table 2.3.

Table 2.3 *Calculation of Aggregate Price Index*

Commodity	*Unit Price (₹)*	
	2009 (p_0)	*2010* (p_1)
Milk (1 litre)	30	35
Eggs (1 dozen)	20	22
Butter (1 kg)	150	170
Bread (500 gm)	12	14
Total	212	241

The unweighted aggregate price index for expenses on food items in 2010 is given by

$$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = (241/212) \times 100 = 113.67.$$

The value P_{01} = 113.67 implies that the price of food items included in the price index has increased by 13.67 per cent over the period 2009 to 2010.

Limitations of an Unweighted Aggregate Price Index Items with large per unit price included in the price index influenced the unweighted aggregate approach of calculating a composite price index. Consequently, relatively low unit price items are dominated by the high unit price items.

Equal weights are assigned to every commodity included in the index irrespective of the relative importance of the commodity in terms of the amount purchased by a consumer. That is, no weight or importance is attached to the price change of a high-use commodity than a low-use commodity. For example, a family may purchase 30 packets of bread in a month as compare to 30 kg butter every month. A substantial price change for slow-moving items like butter and ghee can influence price index.

2.4.3 Average Price Relative Index

The average price relative index is not affected by the unit price of commodities included in the calculation of index as compare to aggregate price index. However, it also suffers from the problem of equal importance (weight) given to all commodities included in the index. Method to calculate **average price relative** index is summarized as follows:

1. Compute price relatives by dividing price of each commodity in the current year by the price of the commodity in the base year.
2. Divide the sum of the price relatives of all commodities by the number of commodities included in the calculation of the index.
3. Multiply the average value obtained in step 2 by 100 to express it in percentage.

The formula for computing the index is as follows:

$$\text{Average price relative index } P_{01} = \frac{1}{n}\sum\left(\frac{p_1}{p_0}\right)100 \qquad (2.2)$$

where n = number of commodities included in the calculation of the index.

The average value used in computing the price relatives index may be either arithmetic mean or geometric mean. If geometric mean is used for averaging the price relatives, then formula (2.2) becomes

$$\log P_{01} = \frac{1}{n}\Sigma \log\left\{\left(\frac{p_1}{p_0}\right)100\right\} = \frac{1}{n}\Sigma \log P;$$

where $P = \left(\frac{p_1}{p_0}\right)100$

Then $P_{01} = \text{antilog}\left\{\frac{1}{n}\Sigma \log p\right\}$

NOTES

Advantages and Limitations of Average Price Relative Index

Advantages

(i) Price relatives are pure numbers, therefore value of average price relative index is not affected by the unit of measurement of commodities included in the calculation of index.

(ii) Equal weights are assigned to every commodity included in the index irrespective to its price.

Limitations

(i) Since equal weights are assigned to every commodity included in the index, therefore each price relative is given equal importance. However in actual practice, it is not true.

(ii) Often arithmetic mean is used to calculate the average of price relatives, but it has a few biases. The use of geometric mean is computationally lengthy.

(iii) Price relatives index does not satisfy all criteria such as identity, time reversal and circular properties, laid down for an ideal index.

Example 2.3: From the data given below, construct the index of price relatives for the year 2002 taking 2001 as base year using (a) arithmetic mean and (b) geometric mean.

Expenses on :	*Food*	*Rent*	*Clothing*	*Education*	*Misc.*
Price (₹), 2010 :	3000	2200	1900	1600	1900
Price (₹), 2011 :	3200	2400	2100	1700	2100

Solution: Calculations of index number using arithmetic mean (A.M.) is shown in Table 2.4

Table 2.4 *Calculation of Index Using A.M.*

Expenses on	*Price in 2010(p0)*	*Price in 2011(p1)*	*Price Relatives* $\frac{p_1}{p_0} \times 100$
Food	3000	3200	106.66
Rent	2200	2400	109.09
Clothing	1900	2100	110.52
Education	1600	1800	112.50
Miscellaneous	1900	2200	115.78
			554.55

Average of price relative index $P_{01} = \frac{1}{n}\Sigma\left(\frac{p_1}{p_0}\right)100$

$= 554.55/ 5 = 110.91$

Hence, prices of commodities included in the calculation of index have increased by 10.91 per cent in the year 2011 as compared to the base year 2010.

NOTES

(b) Index number using geometric mean (G.M.) is shown in Table 2.5.

Table 2.5 *Calculations of Index Using G.M.*

Expenses on	*Price in 2010(p_0)*	*Price in 2011(p_2)*	*Price Relatives* $P = \frac{p_1}{p_0} \times 100$	*Log P*
Food	1800	2000	111.11	2.0457
Rent	1000	1200	120.00	2.0792
Clothing	700	900	128.57	2.1090
Education	400	500	125.00	2.0969
Miscellaneous	700	1000	142.86	2.1548
				10.4856

Average price relative index $P_{01} = \text{antilog}\left\{\frac{1}{n}\Sigma\log p\right\} = \text{antilog}\left\{\frac{1}{5}(10.4856)\right\}$

$= \text{antilog}\ (2.0971) = 125.00$

Check Your Progress

3. State one use of index numbers.
4. How is unweighted price index calculated?

2.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. According to John I Raffin, 'An index number is a quantity which, by reference to a base period, shows by its variations the changes in the magnitude over a period of time. In general, index numbers are used to measure changes over time in magnitudes which are not capable of direct measurement.'
2. The price indexes are of two categories:
   (i) Single price index
   (ii) Composite prices index
3. Index numbers are used to measure general economic condition of a country. For example, the composite index number of prices, industrial

NOTES

output, foreign exchange reserves and bank deposits act as an economic barometer.

4. An unweighted price index is calculated by dividing the price of an item in the given period by the price of the same item in the base period. To facilitate comparison with other years, the price of an item can be converted into a *price relative* so as to express the unit price in each year (period) in terms of percentage of the unit price in the base year.

2.6 SUMMARY

- An *index number* is defined as a relative measure to compare and describe the average change in price, quantity or value of an item over a period of time in relation to its value at some fixed point in time, called the *base period*.
- The base period for indexes may be based at any convenient period, which is occasionally adjusted to a convenient period, and these are published at any convenient frequency.
- According to Irving Fisher, Index number, almost alone in the domain of social sciences, may truly be called an exact science, if it be permissible to designate as science the theoretical foundations of a useful art.
- Index numbers are broadly classified into three categories: (i) price index, (ii) quantity index and (iii) value index.
- Index numbers represent an increase or a decrease (expressed in terms of percentage) in the value of a variable.
- Index numbers are used to measure general economic condition of a country.
- Since purchasing power is related to a group of people or class rather than a particular individual or cost of a single item. Consequently, earnings of a group of people or class must be adjusted with a price index to understand an overall view of the purchasing power for the group.
- An *unweighted price index number measures the percentage change in price for a single item or a basket of items between any two time periods*. Unweighted implies that all the values considered in calculating the index are of equal importance
- *An aggregate (or composite) price index measures the average change in price for a group of related items* in *the current period with respect to the base period.*
- The average price relative index is not affected by the unit price of commodities included in the calculation of index as compare to aggregate price index.

2.7 KEY WORDS

- **Index number:** An index number is an economic data figure reflecting price or quantity compared with a standard or base value. The base usually equals 100 and the index number is usually expressed as 100 times the ratio to the base value.
- **Unweighted index:** An unweighted index is comprised of securities with equal weight within the index. An equivalent dollar amount is invested in each of the index components.
- **Aggregate price level:** The aggregate price level is a measure of the overall level ofprices in the economy. To measure the aggregate price level, economists calculate the cost of purchasing a market basket. Aprice index is the ratio of the current cost of that market basket to the cost in a base year, multiplied by 100.

NOTES

2.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. How are index numbers constructed? What is their purpose?
2. What is an index number? Describe briefly its applications in business and industry.
3. What does an index number measure? Explain the nature and uses of index numbers.
4. What are the basic characteristics of an index number?
5. What are the main uses of an index number?

Long Answer Questions

1. Explain the significance of index numbers.
2. Explain the differences among the three principal types of indexes: price, quantity, and value.
3. Index numbers are economic barometers. Explain this statement and mention the limitations of index numbers (if any).
4. Since value of the base year is always 100, it does not make any difference which period is selected as the base on which to construct an index. Comment.
5. What is meant by the term deflating a value series? Discuss.

2.9 FURTHER READINGS

NOTES

Chandan, J.S., Jagjit Singh and K.K. Khanna. 1995. *Business Statistics*, 2nd edition. New Delhi: Vikas Publishing House Pvt. Ltd.

Croxton, Frederick E., and Dudley J. Cowden. 1943. *Applied General Statistics*. New York: Prentice Hall.

Gupta, S.P. 2006. *Statistical Methods*. New Delhi: S. Chand & Company.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods*, 23rd edition. New Delhi: Vikas Publishing House Pvt.Ltd..

Levin, Richard I. and David S. Rubin. 1998. *Statistics for Management.* New Jersey: Prentice Hall.

UNIT 3 INDEX NUMBERS IN ECONOMICS

NOTES

Structure

3.0 INTRODUCTION

A price-weighted index is an index in which the member companies are weighted in proportion to their price per share, rather than by number of shares outstanding, market capitalization or other factors.

A price-weighted average is a simple mathematical average of several stock prices, and is often used to construct a price-weighted index.

3.1 OBJECTIVES

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

- Describe the commonly used price indexes
- Understand the calculation of Laspeyre's weighting method
- Describe Marshall-Edgeworth method
- Explain Walsch and Kelly's method

3.2 WEIGHTED PRICE INDEXES

While computing weighted price index certain weights are assigned to all commodities in accordance of their relative importance. The weights are of

NOTES

two types: *quantity weights* and *value weights*. There are two commonly used price indexes:

(i) Weighted aggregate price index

(ii) Weighted average of price relative index

Under these, we will focus mainly on Laspeyer, Paasche, Fisher, Marshall and Edgeworth.

Weighted Aggregate Price Index

In a weighted aggregate price index, each item in the basket of items included for calculation of the index is assigned a weight according to its importance. Often, the quantity of usage is considered as the measure of importance. Such weightage improves the accuracy of the general price level estimate.

Several methods (or approaches) to determine weights (value) to be assigned to each item in the basket are as follows:

- Laspeyre's method
- Marshall-Edgeworth's method
- Paasche's method
- Walsch's method
- Dorbish and Bowley's method
- Kelly's method
- Fisher's ideal method

3.2.1 Laspeyre's Weighting Method

Laspeyre's price index, named after the statistician Laspeyre's, treats quantities of usage as constant at base period level and is used for weighting price of each item both in base period and current period. The formula for calculating Laspeyre's price index is given by

$$\text{Laspeyre's price index} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

where $p1$ are prices in the current period; $p0$ are prices in the base period and $q0$ are quantities consumed/sold in the base period.

Advantages and Disadvantages

Advantages: This method uses quantities consumed only in the base period and therefore there is no need not keep record of quantities consumed in each period. Moreover, having used quantities consumed/sold only in the base period, the index of one period may be compared with another period.

Disadvantages: Since in this index fixed quantities consumed in the base period determine weights, it does not adjust changes in consumption patterns. Moreover, consumption of commodities/sold decreases with relatively large increase in price and vice versa.

NOTES

Example 3.1: Compute the cost of living index number using Laspeyre's method from the following information:

Commodity	*Unit Consumption in Base Period*	*Price in Base Period*	*Price in Current Period*
Wheat	200	1.0	1.2
Rice	50	3.0	3.5
Pulses	50	4.0	5.0
Ghee	20	20.0	30.0
Sugar	40	2.5	5.0
Oil	50	10.0	15.0
Fuel	60	2.0	2.5
Clothing	40	15.0	18.0

Solution: Calculation of cost of living index by Laspeyre's method is shown in Table 3.1.

Table 3.1 *Laspeyre's Method*

Commodity	*Base Period Quantity*	*Base Period Price*	*Current Price*		
	($q0$)	($p0$)	($p1$)	$p1q0$	$p0q0$
Wheat	200	1.0	1.2	240	200
Rice	50	3.0	3.5	175	150
Pulses	50	4.0	5.0	250	200
Ghee	20	20.0	30.0	600	400
Sugar	40	2.5	5.0	200	100
Oil	50	10.0	15.0	750	500
Fuel	60	2.0	2.5	150	120
Clothing	40	15.0	18.0	720	600
Total	510			3085	2270

$$\text{Cost of living index} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{3085}{2270} \times 100 = 135.9$$

3.2.2 Paasche's Weighting Method

Paasche's price index treats quantities of usage at current period as constant and is used for weighting price of each item both in base period and current period. The formula for calculating Paasche's index is given by

$$\text{Paasche price index} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}$$

where $p1$ are prices in current period; $p0$ are prices in base period and $q1$ are quantities in current period.

Advantages and Disadvantages

Advantages: In this index, the effects of changes in price and consumption patterns of quantities during the current period are combined. Thus, it provides

NOTES

a better estimate of changes in the economy than Laspeyre's method. If the price or quantity of all items changes in the same ratio, then Laspeyre's and Paasche's indexes will be same.

Disadvantages: This method requires data on quantities consumed of several commodities in each period. Getting the data on the quantities for each period is either expensive or time consuming. Moreover, indexes of different periods by Paasche's method may not be compared because index number for the previous period requires computation to reflect the effect of the new quantity weights.

Example 3.2: For the following data, calculate the price index number of 2011 with 2012 as the base year, using: (a) Laspeyre's method and (b) Paasche's method.

	2008		*2009*	
Commodity	*Price*	*Quantity*	*Price*	*Quantity*
A	20	18	40	16
B	50	10	60	15
C	40	15	50	15
D	20	20	20	25

Solution: Information necessary for both Laspeyre's and Paasche's methods are shown in Table 3.2.

Table 3.2 *Calculation of Laspeyre's and Paasche's Indexes*

Commodity	*Base Period, 2008*		*Current year, 2009*					
	Price ($p0$)	*Quantity* ($q0$)	*Price* ($p1$)	*Quantity* ($q1$)	$p1\,q0$	$p0\,q0$	$p1\,q1$	$p0\,q1$
A	20	8	40	6	320	160	240	120
B	50	10	60	5	600	500	300	250
C	40	15	50	15	750	600	750	600
D	20	20	20	25	400	400	500	500
					2070	1660	1790	1470

$$\text{Laspeyre's price index} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{2070}{1660} \times 100 = 124.7$$

$$\text{Paasche's price index} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100 = \frac{1790}{1470} \times 100 = 121.77$$

The Paasche's price index shows an increase of 21.77 per cent in the price while Laspeyre's index shows an increase of 24.7 per cent. Hence, it may be concluded that Paasche's index shows a trend towards less expensive commodities.

3.2.3 Dorbish and Bowley's Method

Dorbish and Bowley's method is the simple *arithmetic mean* of the Laspeyre's and Paasche's indexes and takes into account the influence of quantity weights

NOTES

of both base period and current period. The formula for calculating Dorbish and Bowley price index is given by

$$\text{Dorbish and Bowley's price index} = \frac{1}{2}\left\{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\right\} \times 100$$

3.2.4 Fisher's Ideal Method

Fisher's ideal method is the *geometric mean* of the Laspeyre's and Paasche's indexes. The formula for calculating price index is given by

$$\text{Fisher's ideal price index} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

Advantages and Disadvantages

Advantages: Fisher's ideal method (i) is based on geometric mean that is considered to be the best average for constructing index numbers, (ii) takes into account both base year and current year quantities as weights and hence avoids the bias associated with the Laspeyre's and Paasche's indexes, and (iii) satisfies essential tests such as time reversal test and factor reversal tests required for an index.

Disadvantages: Since calculation of Fisher's ideal index requires current quantity weights every time an index is calculated, therefore this method is not commonly used. Moreover, calculation of this index require more computation time as compare to other methods.

Example 3.3: Compute index number from the following data using Fisher's ideal index formula.

	2011		*2012*	
Commodity	*Price*	*Quantity*	*Price*	*Quantity*
A	12	10	15	12
B	15	7	20	5
C	24	5	20	9
D	5	16	5	14

Solution: Table 3.3 presents the information necessary for Fisher's method to calculate the index.

Table 3.3 *Calculations of Fisher Ideal Index*

Commodity	*Base Year, 2011*		*Current Year, 2012*					
	($q0$)	($p0$)	($q1$)	($p1$)	$p1q0$	$p0q0$	$p1q1$	$p0q1$
A	12	10	15	12	144	120	180	150
B	15	7	20	5	75	105	100	140
C	24	5	20	9	216	120	180	100
D	5	16	5	14	70	80	70	80
					505	425	530	470

$$\text{Fisher's ideal price index} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 = \sqrt{\frac{505}{425} \times \frac{530}{470}} \times 100$$
$$= \sqrt{1.3399} \times 100 = 1.1576 \times 100 = 115.76$$

NOTES

Hence, it may be concluded that the price level has increased by 15.76 per cent in the year 2012.

Example 3.4: Calculate from the following data, the Fisher's ideal index number for the year 2010:

Commodity	*2009*		*2010*	
	Price (₹)	*Expenditure on Quantity Consumed* (₹)	*Price* (₹)	*Expenditure on Quantity Consumed* (₹)
A	8	200	65	1950
B	20	1400	30	1650
C	5	80	20	900
D	10	360	15	300
E	27	2160	10	600

Solution: Table 3.4 presents the information necessary for Fisher's method to calculate the index.

Table 3.4 *Calculations of Fisher's Ideal Index*

Commodity	*Base Year, 2009*		*Current Year, 2010*					
	($p0$)	($q0$)	($p1$)	($q1$)	$p1\ q0$	$p0\ q0$	$p1\ q1$	$p0\ q1$
A	8	200/8 = 25	65	1950/65 = 30	200	1950	240	240
B	20	1400/20 = 70	30	1650/30 = 55	2100	1400	1650	1100
C	5	80/5 = 16	20	900/20 = 45	320	80	900	225
D	10	360/10 = 36	15	300/15 = 20	540	360	300	200
E	27	2160/27 = 80	10	600/10 = 60	800	2160	600	1620
					5385	4200	5400	3385

$$\text{Fisher's ideal price index} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 = \sqrt{\frac{5385}{4200} \times \frac{5400}{3385}} \times 100$$
$$= 1.430 \times 100 = 143.$$

Hence, it may be concluded that the price level has increased by 43 per cent in the year 2010.

3.2.5 Marshall-Edgeworth Method

In this method, the sum of quantities consumed of all commodities in base year and current year is considered as weight to calculate the index. The formula for constructing the index is

$$\text{Marshall-Edgeworth price index} = \frac{\Sigma(q_0 + q_1)p_1}{\Sigma(q_0 + q_1)p_0} \times 100 = \frac{\Sigma q_0 p_1 + \Sigma q_1 p_1}{\Sigma q_0 p_0 + \Sigma q_1 p_0} \times 100$$

The disadvantages of this method are the same as that of Paasche index and Fisher's ideal index because this method also uses current quantity weights to construct the index.

NOTES

3.2.6 Walsch's Method

In this method, the geometric mean of the base and current year quantities is used as weight. The formula for constructing the index is

$$\text{Walsch's price index} = \frac{\Sigma p_1 \sqrt{q_0 q_1}}{\Sigma p_0 \sqrt{q_0 q_1}} \times 100$$

This index satisfies the time reversal test but requires current quantity weight every time an index is constructed.

3.2.7 Kelly's Method

Kelly's method also called the *Fixed Weight Aggregate Method* uses fixed quantity weight both for base period and current period. The formula for constructing the index is

$$\text{Kelly's price index} = \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100$$

where q is the fixed weight.

The fixed weights and the base period prices may not be from the same period.

Advantages and Disadvantages

Advantages: For constructing this index different period may be chosen for fixed weight other than base period. The base period can also be changed without changing the fixed weight.

Disadvantages: For constructing this index the weight neither of the base period nor the current period is taken into consideration.

Example 3.5: It is stated that the Marshall-Edgeworth index number is a good approximation of the ideal index number. Verify this statement using the following data:

Commodity	*2008*		*2009*	
	Price	*Quantity*	*Price*	*Quantity*
A	2	74	3	82
B	5	125	4	140
C	7	40	6	33

Solution: Table 3.5 presents the information necessary to calculate Fisher and Marshall-Edgeworth indexes.

Table 3.5 *Calculations of Fisher's Ideal and Marshall-Edgeworth's Index*

Commodity	*Base Year, 2008*		*Current Year, 2009*					
	($p0$)	($q0$)	($p1$)	($q1$)	$p1\,q0$	$p0\,q0$	$p1\,q1$	$p0\,q1$
A	2	74	3	82	222	148	246	164
B	5	125	4	140	500	625	560	700
C	7	40	6	33	240	280	198	231
					962	1053	1004	1095

NOTES

$$\text{Fisher ideal price index} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 = \sqrt{\frac{962}{1053} \times \frac{1004}{1095}} \times 100$$
$$= \sqrt{0.836} \times 100 = 0.9144 \times 100 = 91.44$$

$$\text{Marshall-Edgeworth price index} = \frac{\Sigma p_1 (q_0 + q_1)}{\Sigma p_0 (q_0 + q_1)} \times 100 = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$
$$= \frac{962 + 1004}{1053 + 1095} \times 100 = 0.9152 \times 100 = 91.52$$

Hence, it may be concluded that Fisher's method and Marshall-Edgeworth method provide almost the same valve of the index.

Example 3.6: Compute Laspeyre's, Paasche's, Fisher's and Marshall-Edgeworth's index numbers from the following data:

Item	*1998*		*1999*	
	Price	*Quantity*	*Price*	*Quantity*
A	5	25	6	30
B	3	8	4	10
C	2	10	3	8
D	10	4	3	5

Solution: Table 3.6 presents the information necessary to calculate several indexes.

Table 3.6 *Calculations of Indexes*

Item	*Base Year, 1998*		*Current Year, 1999*					
	($p0$)	($q0$)	($p1$)	($q1$)	$p1\,q0$	$p0\,q0$	$p1\,q1$	$p0\,q1$
A	5	25	6	30	150	125	180	150
B	3	8	4	10	32	24	40	30
C	2	10	3	8	30	20	24	16
D	10	4	3	5	12	40	15	50
					224	209	259	246

$$\text{Laspeyre's price index} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{224}{209} \times 100 = 107.17$$

$$\text{Paasche's price index} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100 = \frac{259}{246} \times 100 = 105.28$$

$$\text{Fisher's ideal price index} = \sqrt{L \times P} = \sqrt{107.17 \times 105.28} = 106.22$$

$$\text{Marshall-Edgeworth's price index} = \frac{\Sigma p_1 (q_0 + q_1)}{\Sigma p_0 (q_0 + q_1)} \times 100 = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$
$$= \frac{244 + 259}{209 + 246} \times 100 = 110.55$$

NOTES

Weighted Average of Price Relative Index

For constructing weighted average of price relative, the quantity consumed in the base period is also used for weighting the items or commodities. The value (in rupees) of each item or commodity included in the calculation of composite index is determined by multiplying the price of each item by its quantity consumed.

The formula for constructing the weighted average of price relative index using base values is

$$\text{Weighted average of price relative index, } P_{01} = \frac{\Sigma\{(p_1/p_0)\times 100\}(p_0 q_0)}{\Sigma p_0 q_0} = \frac{\Sigma PV}{\Sigma V}$$

$$= \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

where V(= $p0q0$) = base period value P(= $(p_1/p_0) \times 100$ = price relative. This formula is same as Laspeyre's method for constructing unweighted price index.

If base period value is taken as V = $p0q1$ to compute a weighted average of price relative, then the above formula becomes

$$P_{01} = \frac{\Sigma\{(p_1/p_0)\times 100\}(p_0 q_1)}{\Sigma p_0 q_1} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

This formula is same as Paasche's method for constructing unweighted price index.

If instead of weighted arithmetic average the weighted geometric mean is used, then the above formula becomes

$$P_{01} = \frac{\Sigma V \times \log P}{\Sigma V}; \text{ where } P = \frac{p_1}{p_0} \times 100 \text{ and } V = p0q0$$

Advantages of Weighted Average Price Relatives

(i) Index numbers constructed using average price relative with same base can be combined to form a new index.

(ii) Weighted average of price relative method is suitable to construct an index by selecting one item from each subgroups of items. The values of each subgroup may then be used as weights.

Example 3.7: A large manufacturer purchases an identical component from three different suppliers that differ in unit price and quantity supplied. The relevant data for 2005 and 2006 are given below:

Supplier	*Quantity Index in (2005)*	*Unit Price* (₹)	
		2005	*2006*
A	20	18	20
B	40	12	14
C	10	15	16

NOTES

Construct a weighted average price relative index using (a) arithmetic mean and (b) geometric mean.

Solution: Calculations for the weighted average price relative index mean are shown in Table 3.7.

Table 3.7 *Calculations of Weighted Average of Price Relatives*

Supplier	*Prices in* 2005 $p0$	2006 $p1$	*Quantity in 2005* $q0$	*Percentage Price Relative* $P = \frac{p_1}{p_0} \times 100$	*Base Value* $V = p0q0$	*Weighted Percentage Relative PV*
A	18	20	20	(20/18) × 100 = 111.11	360	39,999.60
B	12	14	40	(14/12) × 100 = 116.67	480	56,001.60
C	15	16	10	(16/15) × 100 = 106.67	150	16,000.50
					990	1,12,001.70

(a) Weighted average of price relative index

$$P01 = \frac{[\Sigma(p_1 / p_0)100]\, p_0 q_0}{\Sigma p_0 q_0} = \frac{1,12,001.70}{990} = 113.13$$

The value of P01 implies that there has been 13.13 per cent increase in price from year 2005 to 2006.

(b) **C**alculations for the weighted geometric price relative index are shown in Table 3.8.

Table 3.8 *Calculations of Weighted Geometric Mean of Price Relatives*

Supplier	*Prices in* 2005 $p0$	2006 $p1$	*Quantity in 2005* $q0$	*Base Value* $V = p0q0$	*Percentage Price Relative* $P = \frac{p_1}{p_0} \times 100$	*Log P*	*V log P*
A	18	20	20	360	111.11	2.046	736.56
B	12	14	40	480	116.67	2.067	992.16
C	15	16	10	150	106.67	2.028	304.20
				990			2032.92

Weighted geometric mean of price relatives

$$P01 = \text{antilog}\left\{\frac{\Sigma V \times \log P}{\Sigma V}\right\} = \text{antilog}\left\{\frac{2032.92}{990}\right\}$$

$$= \text{antilog}\,(2.0535) = 113.11$$

Check Your Progress

1. State one advantage of Laspeyre's weighting method.
2. State one disadvantage of Paasche's weighting method.
3. What is Fisher's ideal method?
4. What is the other name for Kelly's method?

3.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

NOTES

1. Laspeyre's method uses quantities consumed only in the base period and therefore there is no need not keep record of quantities consumed in each period.
2. This method requires data on quantities consumed of several commodities in each period. Getting the data on the quantities for each period is either expensive or time consuming.
3. Fisher's ideal method is the geometric mean of the Laspeyre's and Paasche's indexes.
4. Kelly's method is also called the Fixed Weight Aggregate Method.

3.4 SUMMARY

- While computing weighted price index certain weights are assigned to all commodities in accordance of their relative importance. The weights are of two types: *quantity weights* and *value weights*.
- In a weighted aggregate price index, each item in the basket of items included for calculation of the index is assigned a weight according to its importance.
- Several methods (or approaches) to determine weights (value) to be assigned to each item in the basket are as follows:
  - Laspeyre's method
  - Marshall-Edgeworth's method
  - Paasche's method
  - Walsch's method
  - Dorbish and Bowley's method
  - Kelly's method
  - Fisher's ideal method
- *Laspeyre's price index*, named after the statistician Laspeyre's, treats quantities of usage as constant at base period level and is used for weighting price of each item both in base period and current period.
- *Paasche's price index* treats quantities of usage at current period as constant and is used for weighting price of each item both in base period and current period.

NOTES

- ***Dorbish and Bowley's method*** is the simple *arithmetic mean* of the Laspeyre's and Paasche's indexes and takes into account the influence of quantity weights of both base period and current period.
- Fisher's ideal method is the *geometric mean* of the Laspeyre's and Paasche's indexes.
- For constructing weighted average of price relative, the quantity consumed in the base period is also used for weighting the items or commodities.
- The value (in rupees) of each item or commodity included in the calculation of composite index is determined by multiplying the price of each item by its quantity consumed.

3.5 KEY WORDS

- **Weighted aggregate price index**: In a weighted aggregate price index, each item in the basket of items included for calculation of the index is assigned a weight according to its importance.
- **Weighted average of price relative**: For constructing weighted average of price relative, the quantity consumed in the base period is also used for weighting the items or commodities.

3.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. Briefly state the definition of weighted aggregate price index.
2. Write a short note on Fisher's ideal method.

Long-Answer Questions

1. Describe the advantages and limitations of Laspeyre's weighting method.
2. How is Marshall-Edgeworth method calculated? Discuss with the help of an example.

3.7 FURTHER READINGS

Chandan, J.S., Jagjit Singh and K.K. Khanna. 1995. *Business Statistics*, 2nd edition. New Delhi: Vikas Publishing House Pvt. Ltd.

Croxton, Frederick E., and Dudley J. Cowden. 1943. *Applied General Statistics*. New York: Prentice Hall.

Gupta, S.P. 2006. *Statistical Methods*. New Delhi: S. Chand & Company.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods*, 23rd edition. New Delhi: Vikas Publishing House Pvt.Ltd..

Levin, Richard I. and David S. Rubin. 1998. *Statistics for Management.* New Jersey: Prentice Hall.

NOTES

NOTES

BLOCK - II
CENSUS AND SAMPLING

UNIT 4 CENSUS AND SAMPLING: AN INTRODUCTION

Structure

4.0 INTRODUCTION

Research objectives are generally translated into research questions that enable the researchers to identify the information needs. Once the information needs are specified, the sources of collecting the information are sought. Some of the information may be collected through secondary sources (published material), whereas the rest may be obtained through primary sources. The primary methods of collecting information could be the observation method, personal interview with questionnaire, telephone surveys and mail surveys. Surveys are, therefore, useful in information collection, and their analysis plays a vital role in finding answers to research questions. Survey respondents should be selected using the appropriate procedures, otherwise the researchers may not be able to get the right information to solve the problem under investigation. The process of selecting the right individuals, objects or events for the study is known as sampling. Sampling involves the study of a small number of individuals, objects chosen from a larger group.

In this unit, you will study about the fundamental concept of sampling and census. In addition to this, the unit also discusses the use of sampling in everyday life along with the main features of the same. A comparison between a sample and a census has also been explained in the unit.

NOTES

4.1 OBJECTIVES

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

- Define the basic concept of sampling and census
- Understand the use of sampling in real life
- Differentiate between sample and census
- Analyse the meaning and features of sampling

4.2 MEANING AND FEATURES OF SAMPLING

Before we get into the details of various issues pertaining to sampling, it would be appropriate to discuss some of the sampling concepts.

Population: Population refers to any group of people or objects that form the subject of study in a particular survey and are similar in one or more ways. For example, the number of full-time MBA students in a business school could form one population. If there are 200 such students, the population size would be 200. We may be interested in understanding their perceptions about business education. If there are 200 class IV employees in an organization and we are interested in measuring their job satisfaction, all the 200 class IV employees would form the population of interest. If a TV manufacturing company produces 150 TVs per week and we are interested in estimating the proportion of defective TVs produced per week, all the 150 TVs would form our population. If, in an organization there are 1000 engineers, out of which 350 are mechanical engineers and we are interested in examining the proportion of mechanical engineers who intend to leave the organization within six months, all the 350 mechanical engineers would form the population of interest. If the interest is in studying how the patients in a hospital are looked after, then all the patients of the hospital would fall under the category of population.

Element: An element comprises a single member of the population. Out of the 350 mechanical engineers mentioned above, each mechanical engineer would form an element of the population. In the example of MBA students whose perception about the management education is of interest to us, each of the 200 MBA students will be an element of the population. This means that there will be 200 elements of the population.

Sampling frame: Sampling frame comprises all the elements of a population with proper identification that is available to us for selection at any stage of sampling. For example, the list of registered voters in a constituency could form a sampling frame; the telephone directory; the number of students

NOTES

registered with a university; the attendance sheet of a particular class and the payroll of an organization are examples of sampling frames. When the population size is very large, it becomes virtually impossible to form a sampling frame. We know that there is a large number of consumers of soft drinks and, therefore, it becomes very difficult to form the sampling frame for the same.

Sample: It is a subset of the population. It comprises only some elements of the population. If out of the 350 mechanical engineers employed in an organization, 30 are surveyed regarding their intention to leave the organization in the next six months, these 30 members would constitute the sample.

Sampling unit: A sampling unit is a single member of the sample. If a sample of 50 students is taken from a population of 200 MBA students in a business school, then each of the 50 students is a sampling unit. Another example could be that if a sample of 50 patients is taken from a hospital to understand their perception about the services of the hospital, each of the 50 patients is a sampling unit.

Sampling: It is a process of selecting an adequate number of elements from the population so that the study of the sample will not only help in understanding the characteristics of the population but will also enable us to generalize the results. We will see later that there are two types of sampling designs—probability sampling design and non-probability sampling design.

Census (or complete enumeration): An examination of each and every element of the population is called census or complete enumeration. Census is an alternative to sampling. We will discuss the inherent advantages of sampling over a complete enumeration later.

4.2.1 Uses of Sampling in Real Life

In our day-to-day life we make use of the concept of sampling. There is hardly any person who has not made use of the concept in a real-life situation. Consider the following examples:

- Suppose you go to a grocery shop to purchase rice. You have been instructed by your mother to purchase good quality rice. On reaching the grocery shop you have the choice of buying the rice from any one of three bags. What is generally done is that you pick up a handful of rice from each bag, examine its quality and then decide about which bag's rice is to be bought. The concept of sampling is being used here as a handpick from each bag is a sample and examining the quality is a process by which you are trying to assess the quality of all the rice in the bag.
- Suppose you have a guest for dinner at your residence. Your mother prepares a number of dishes and before the guest arrives, she may give

you a tablespoon of each of the dish to taste and tell her whether all the ingredients are in the right proportion or not. Again, a sample is being taken from each of the dish to know how each of them tastes.

- You go to a bookshop to buy a magazine. Before you decide to buy it, you may flip through its pages to know whether the contents of the magazines are of interest to you or not. Again, a sample of pages is taken from the magazine.

4.3 POPULATION AND SAMPLE

In a research study, we are generally interested in studying the characteristics of a population. Suppose in a town there are 2 lakh households and we are interested in estimating the proportion of those households who spend their summer vacations in a hill station. This information can be obtained by asking every household in that town. If all the households in a population are asked to provide information, such a survey is called a census. There is an alternative way of obtaining the same information by choosing a subset of all the two lakh households and asking them for the same information. This subset is called a sample. Based upon the information obtained from the sample, a generalization about the population characteristic could be made. However, that sample has to be representative of the population. For a sample to be a representative of the population, the distribution of sampling units in the sample has to be in the same proportion as the elements in the population. For example, if in a town there are 50, 35 and 15 per cent households in lower, middle and upper income groups, then a sample taken from this population should have the same proportions in for it to be representative. There are several advantages of sample over census.

- Sample saves time and cost. Consider as an example that we are interested in estimating the monthly average household expenditure on food items by the people of Delhi. It is known that the population of Delhi is approximately 1.2 crore. Now, if we assume that there are five members per household, it would mean that the population comprises approximately 24 lakh households. Collecting data on the expenditure of each of the 24 lakh households on food items would be a very time-consuming and expensive exercise. This is because you will need to hire a number of investigators and train them before you conduct the survey on the 24 lakh households. Instead, if a sample of, say, 2000 households is chosen, the task would not only be finished faster but will be inexpensive, too.
- Many times a decision-maker may not have too much of time to wait till all the information is available. Therefore, a sample could come to his rescue.

NOTES

- There are situations where a sample is the only option. When we want to estimate the average life of fluorescent bulbs, what is done is that they are burnt out completely. If we go for a complete enumeration there would not be anything left for use. Another example could be testing the quality of a photographic film. To test the quality, we need to expose it completely and the moment it is exposed it gets destroyed. Therefore, sample is the only choice.
- The study of a sample instead of complete enumeration may, at times, produce more reliable results. This is because by studying a sample, fatigue is reduced and fewer errors occur while collecting the data, especially when a large number of elements are involved.

A census is appropriate when the population size is small, e.g., the number of public sector banks in the country. Suppose the researcher is interested in collecting information from the top management of a bank regarding their views on the monetary policy announced by the Reserve Bank of India (RBI), in this case, a complete enumeration may be possible as the population size is not very large. As another example, consider a business school having a few students from Europe, East Africa, South East Asia and the Middle East. These students would have their own problems in settling down in the Indian environment because of the differences in social, cultural and environmental factors. To understand their concerns, a survey of population may be more appropriate. Therefore, a survey of population could be used when there is a lot of heterogeneity in the variables of interest and the population size is small.

Check Your Progress

1. Define sampling.
2. What does the term 'population' denote in sampling?
3. What is sampling frame?
4. What does census mean?
5. Distinguish between sample and census with the help of an example.
6. Mention any one situation where sampling is the only option for survey.

4.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The process of selecting the right individuals, objects or events for the study is known as sampling. The process involves the study of a small number of individuals, objects chosen from a larger group.

2. Population refers to any group of people or objects that form the subject of study in a particular survey and are similar in one or more ways. For example, the number of full-time MBA students in a business school could form one population. If there are 200 such students, the population size would be 200.
3. Sampling frame comprises all the elements of a population with proper identification that is available to us for selection at any stage of sampling. For example, the list of registered voters in a constituency could form a sampling frame; the telephone directory; the number of students registered with a university; the attendance sheet of a particular class and the payroll of an organization are examples of sampling frames.
4. Census or complete enumeration is an examination of each and every element of the population. Census is an alternative to sampling.
5. The difference between sample and census can be understood by simply studying the characteristics of a population. Suppose in a town there are 2 lakh households and we are interested in estimating the proportion of those households who spend their summer vacations in a hill station. This information can be obtained by asking every household in that town. If all the households in a population are asked to provide information, such a survey is called a census. There is an alternative way of obtaining the same information by choosing a subset of all the two lakh households and asking them for the same information. This subset is called a sample.
6. At times, there are situations where a sample is the only option for survey. For example, if we want to estimate the average life of fluorescent bulbs, what is done is that they are burnt out completely. If we go for a complete enumeration there would not be anything left for use. Another example could be testing the quality of a photographic film. To test the quality, we need to expose it completely and the moment it is exposed it gets destroyed. Therefore, sample is the only choice.

4.5 SUMMARY

- The process of selecting the right individuals, objects or events for the study is known as sampling. Sampling involves the study of a small number of individuals, objects chosen from a larger group.
- Population refers to any group of people or objects that form the subject of study in a particular survey and are similar in one or more ways. For example, the number of full-time MBA students in a business school could form one population. If there are 200 such students, the population size would be 200.

NOTES

- An element comprises a single member of the population. Out of the 350 mechanical engineers mentioned above, each mechanical engineer would form an element of the population.
- Sampling frame comprises all the elements of a population with proper identification that is available to us for selection at any stage of sampling.
- Sample is a subset of the population. It comprises only some elements of the population.
- A sampling unit is a single member of the sample. If a sample of 50 students is taken from a population of 200 MBA students in a business school, then each of the 50 students is a sampling unit.
- Sampling is a process of selecting an adequate number of elements from the population so that the study of the sample will not only help in understanding the characteristics of the population but will also enable us to generalize the results.
- Sample saves time and cost. Consider as an example that we are interested in estimating the monthly average household expenditure on food items by the people of Delhi. It is known that the population of Delhi is approximately 1.2 crore.
- A census is appropriate when the population size is small, e.g., the number of public sector banks in the country.

4.6 KEY WORDS

- **Census:** It refers to the procedure of systematically acquiring and recording information about the members of a given population.
- **Sampling:** It refers to the process or technique used to select a suitable sample for testing or analysing.
- **Sampling Frame:** It refers to a list of items or people forming a population from which a sample is taken.

4.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. What are the different methods of collecting information?
2. What is an element in research?
3. What is the difference between a sample and a sampling frame?
4. What do you understand by sampling unit?

Long Answer Questions

1. Describe the uses of sampling in real life.
2. Why is sampling preferred to the census? Explain.
3. Discuss the advantages of sample survey over census.

NOTES

4.8 FURTHER READINGS

Chandan, J.S., Jagjit Singh and K.K. Khanna. 1995. *Business Statistics*, 2nd edition. New Delhi: Vikas Publishing House Pvt. Ltd.

Croxton, Frederick E., and Dudley J. Cowden. 1943. *Applied General Statistics*. New York: Prentice Hall.

Gupta, S.P. 2006. *Statistical Methods*. New Delhi: S. Chand & Company.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods*, 23rd edition. New Delhi: Vikas Publishing House Pvt.Ltd..

Levin, Richard I. and David S. Rubin. 1998. *Statistics for Management.* New Jersey: Prentice Hall.

UNIT 5 SAMPLING: MEANING AND TYPES

NOTES

Structure

5.0 INTRODUCTION

As you have already studied in the previous unit, the process of selecting a sample or a portion of elements from a population or a process using a specific method is called sampling. Moving the discussion forward, in this unit you will study about different steps in sampling process, the concept of representative sample, sampling techniques, and the types of sampling.

5.1 OBJECTIVES

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

- Understand the meaning of sampling
- Discuss the steps involved in the sampling process
- Explain the concept of representative sample
- Comprehend the types of sampling

5.2 SAMPLING: MEANING AND DEFINITIONS

Usually, sampling involves determining a property or attribute to adhere to for the purpose of differentiating between items of a given population. These attributes, which are the objects of study, are called characteristics. The process of distinguishing the items is usually of two types, quantitative or qualitative. In quantitative sampling, characteristics pertaining to variables are dealt with. On the other hand, qualitative sampling is concerned with the characteristics related to attributes.

The basic idea behind sampling is to use the common characteristics of average items as samples for a larger entity. Thus, it involves choosing a subset of population elements for study. Thus, for example, if the population to be dealt with is, say that of roads, then the characteristics could be length, duration, roughness, carriage capacity, and so forth. Sampling proves to be a much cheaper and quicker mode of estimation where the population is absolutely huge.

However, it is absolutely necessary to take ample care while determining which characteristics should be sampled. Those characteristics, which are rare, should be avoided. Similarly, even if there are certain very common characteristics, which, however, do not contribute in any way to draw reliable estimates, then such characteristics should not be sampled.

5.2.1 Steps in Sampling

The sampling process involves the following seven tasks:

- **Defining the population:** It involves completely defining the population by specifying the following terms:
  - o Elements
  - o Sampling units
  - o Extent
  - o Time
- **Selecting the sampling frame:** The sampling frame should be selected in such a way that it consists of almost all the sampling units. A sample should be selected in such a way that it has all the characteristics of the population. Some of the popular sampling frames are census reports and electoral registers.
- **Specifying the sampling unit:** Sampling unit is the basic unit that contains elements of the target population.
- **Specifying the sampling method:** This method depicts how the sample units are selected. The most important decision in this method is to determine, which of the two—probability and non-probability—samples is to be chosen.
- **Determining the sample size:** This method includes decision-making about the number of elements to be chosen.
- **Specifying the sampling plan:** This method dictates that one should indicate how decisions made so far are to be implemented. All the expected issues in relation to the sampling survey must be answered by the sample plan.
- **Selecting the sample:** This is the final step in the sample process, which includes a good deal of fieldwork and office work. This is introduced

in the actual selection of the sample elements. It mainly depends on the sampling plan and the sample size required.

Representative Sample

NOTES

When a researcher carries out the research study, he may select a comparatively small number of subjects from the entire population. Thus, for example, he may choose salespersons of supermarkets from all supermarkets in the country. Literally, in this case, the researcher is using a representative sample. Thus, we can define representative sample as the sample which possesses the same characteristics as that of its parent population or variable. Thus, it factually represents the variation that exists in the parent variable on the general level.

The significance of a representative sample lies in the fact that it represents the population more accurately. For this, it is absolutely necessary that the sampling process be kept free from errors. Errors may occur when representative sampling is based on surveys that may be hampered with non-response errors or self-selection errors. By non-response errors, we mean that a survey is conducted in such a way that the researcher has targeted a large number of subjects, but only a small per cent has responded.

This can be explained with an example. Suppose that a supermarket tries to conduct a survey of its customers by offering feedback forms to every consumer and instructing them to put the filled-in form in a drop box. In this case, it is possible that some customers may fill the form, whereas some may just carry it and discard it outside. Suppose that the number of customers visiting every day is around 400. In case, just seventy-five of these have put in completely filled forms, the management cannot infer these seventy-five customers as representing the total 400. As such, if these seventy-five are used for the process of sampling, inaccurate generalizations are bound to occur.

In the case of self-selection error, there may be, for example, a few customers who may have chosen to fill only half the form. As can be seen, in this case too, the sampling is not possible on such a self-selected partial feedback of the customers. Another possibility is that the researcher may be tempted to conduct sampling by personal standards, which can greatly obstruct the generalization purpose of sampling. In this case, there is a possibility of the measurement being distorted or miscalculated as a result of subjective influence on the part of the surveyor.

Sampling Techniques

Sampling involves the application of a number of predefined concepts and types for conducting the survey meant for research. It is also necessary for the researcher to get acquainted with the various terms involved for effective application of the sampling method.

5.3 TYPES OF SAMPLING

NOTES

Primarily, there are two types of sampling, namely probability sampling and non-probability sampling which are further divided into sub types, as mentioned below:

1. **Probability Sampling:** It is a type of sampling technique wherein a sample from a larger population chosen with the help of a method based on the theory of probability. This sampling technique is further divided into five sub-types, such as:
   (a) Simple Random Sampling
   (b) Systematic Sampling
   (c) Stratified Random Sampling
   (d) Cluster Sampling
   (e) Multi-stage Sampling
2. **Non-probability Sampling:** In this type of sampling technique the samples are gathered in a process that does not give all the individuals in the population equal chances of being selected in the sample.
   (a) Convenience Sampling
   (b) Judgemental Sampling
   (c) Snowball Sampling
   (d) Quota Sampling

All of the above-mentioned sampling techniques have been discussed in detail in the forthcoming unit.

Check Your Progress

1. What is the difference between quantitative and qualitative sampling?
2. How can are select a sample and a sampling frame?
3. What is representative sample?
4. What are the two types of sampling?

5.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. In quantitative sampling, characteristics pertaining to variables are dealt with. On the other hand, qualitative sampling is concerned with the characteristics related to attributes.

NOTES

2. A sample should be selected in such a way that it has all the characteristics of the population. While a sampling frame should be selected in such a way that it consists of almost all the sampling units. Some of the popular sampling frames are census reports and electoral registers.
3. A representative sample is the sample which possesses the same characteristics as that of its parent population or variable. Thus, it factually represents the variation that exists in the parent variable on the general level. For example, a researcher may choose salespersons of supermarkets from all supermarkets in the country. Literally, in this case, the researcher here is using a representative sample.
4. The two types of sampling are: probability sampling and non-probability sampling.

5.5 SUMMARY

- Sampling involves determining a property or attribute to adhere to for the purpose of differentiating between items of a given population. These attributes, which are the objects of study, are called characteristics.
- In quantitative sampling, characteristics pertaining to variables are dealt with. On the other hand, qualitative sampling is concerned with the characteristics related to attributes.
- The sampling process involves the seven tasks, such as: defining the population, selecting the sampling frame, specifying the sampling unit, specifying the sampling method, determining the sample size, specifying the sampling plan, and selecting the sample.
- The representative sample as the sample which possesses the same characteristics as that of its parent population or variable. Thus, it factually represents the variation that exists in the parent variable on the general level.
- The representative sample represents the population more accurately. For this, it is absolutely necessary that the sampling process be kept free from errors.
- Sampling involves the application of a number of predefined concepts and types for conducting the survey meant for research. It is also necessary for the researcher to get acquainted with the various terms involved for effective application of the sampling method.
- There are two types of sampling, namely probability sampling and non-probability sampling which are further divided into sub types.

5.6 KEY WORDS

- **Representative sampling:** It refers to a subset of a population that seeks to accurately reflect the characteristics of the larger group.
- **Non-response error:** It occurs when sampling units selected for a sample are not interviewed.
- **Sampling techniques:** It refers to the name or other identification of the specific process by which the entities of the sample have been selected.

NOTES

5.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. What is the basic idea behind sampling?
2. Why is representative sampling important?
3. What do you mean by self-selection error?

Long Answer Questions

1. Explain the steps of the sampling process.
2. Discuss the two types of sampling along with their sub-types.

5.8 FURTHER READINGS

Chandan, J.S., Jagjit Singh and K.K. Khanna. 1995. *Business Statistics*, 2nd edition. New Delhi: Vikas Publishing House Pvt. Ltd.

Croxton, Frederick E., and Dudley J. Cowden. 1943. *Applied General Statistics*. New York: Prentice Hall.

Gupta, S.P. 2006. *Statistical Methods*. New Delhi: S. Chand & Company.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods*, 23rd edition. New Delhi: Vikas Publishing House Pvt.Ltd..

Levin, Richard I. and David S. Rubin. 1998. *Statistics for Management.* New Jersey: Prentice Hall.

UNIT 6 SAMPLING DESIGN: MEANING, TYPES AND CHALLENGES

Structure

6.0 INTRODUCTION

Sampling design refers to the process of selecting samples from a population. There are two types of sampling designs – probability sampling design and non-probability sampling design. Probability sampling designs are used in conclusive research. In a probability sampling design, each and every element of the population has a known chance of being selected in the sample. In case of non-probability sampling design, the elements of the population do not have any known chance of being selected in the sample. These sampling designs are used in exploratory research. The unit is all about the meaning, types and challenges of sampling designs.

6.1 OBJECTIVES

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

- Understand the meaning of sampling design
- Discuss the types and sub-types of sampling design
- Describe the challenges of sampling design

NOTES

6.2 PROBABILITY SAMPLING DESIGN

Under this, the following sampling designs would be covered—simple random sampling with replacement (SRSWR), simple random sampling without replacement (SRSWOR), systematic sampling, stratified random sampling and cluster sampling.

6.2.1 Simple Random Sampling with Replacement

Under this scheme, a list of all the elements of the population from where the samples to be drawn is prepared. If there are 1000 elements in the population, we write the identification number or the name of all the 1000 elements on 1000 different slips. These are put in a box and shuffled properly. If there are 20 elements to be selected from the population, the simple random sampling procedure involves selecting a slip from the box and reading of the identification number. Once this is done, the chosen slip is put back to the box and again a slip is picked up and the identification number is read from that slip. This process continues till a sample of 20 is selected. Please note that the first element is chosen with a probability of 1/1000, the second one is also selected with the same probability and so are all the subsequent elements of the population.

An alternative way of selecting the samples from the population is by using random number tables. Table 6.1 gives an illustrative example of random numbers.

Table 6.1 *Select four-digit random numbers*

I	II	III	IV	V
2807	0495	6183	7871	9559
8016	5732	3448	0164	2367
1322	4678	8034	1139	1474
0843	4625	7407	9987	5734
2364	1187	4565	2343	9786
4885	8755	4355	5465	0575
3406	4678	5950	7222	8494
5927	6010	7545	8979	1041
4447	3476	9140	0736	2332
4968	7553	1073	2493	4251
7489	1630	2330	4250	6170
4010	2707	3925	6007	8089
6531	9784	5520	7764	0008
7052	3861	7115	9521	2192

6573	2793	8710	2127	3846
8094	3205	2030	3035	5765
8615	6092	1900	4792	7684
9136	4016	3495	6549	9603
9656	5246	5090	8306	1522
2017	8323	1685	3006	3441

Table 6.1 gives four-digit random numbers arranged in 20 rows and five columns. These random numbers can be generated by a computer programmed to scramble numbers. The logic for generating random number is that any number can be constructed from numbers 0 to 9. The probability that any one digit from 0 through 9 will appear is the same as that for any other digit and the appearance of the numbers is statistically independent. Further, the probability of one sequence of digits occurring is the same as that for any other sequence of the same length.

The use of random number table for selecting samples could be illustrated through an example. Suppose there are 75 students in a class and it is decided to select 15 out of the 75 students. These students can be numbered from 01 to 75. Now, to pick up 15 students using random numbers and following the scheme of simple random sampling with replacement, we proceed as follows:

- With eyes closed, we place our finger on a number on the random number table. Suppose it is on the first row and the first column of our table. Now, we go down the first two columns and choose two-digit random numbers running from 01 to 75. If any number greater than 75 appears, it gets rejected. This way, the first number to be selected would be 28. The second number is 80, which would be rejected as we are choosing numbers from 01 to 75. The next selected number would be 13, followed by 08, 23, 48, 34, 59, 44, 49, 74, 40, 65, 70 and 65. Note that 65 has appeared twice. Since we are using the scheme of simple random sampling with replacement, we would retain it. This way we have selected 14 samples. The 15th number selected would be 20. In brief, the scheme explained above states that any number greater than the population size (in this case 75) is rejected and only the numbers from 01 to 75 are selected. A number may get repeated because simple random sampling scheme is done with replacement.

Simple Random Sampling Without Replacement

In the case of simple random sample without replacement, the procedure is identical to what was explained in the case of simple random sampling with replacement. The only difference here is that the chosen slip is not placed back in the box. This way, the first unit would be selected with the probability of 1/1000, second unit with the probability of 1/999, the third will be selected

with a probability of 1/998 and so on, till we select the required number of elements (in this case, 20) in our sample.

The simple random sampling (with or without replacement) is not used in a consumer research. This is because in a consumer research the population size is usually very large, which creates problems in the preparation of a sampling frame. For example, there is a large number of consumers of soft drinks, pizza, shampoo, soap, chocolate, etc. However, these (SRSWR and SRSWOR) designs could be useful when the population size is very small, for example, the number of steel/aluminum-producing companies in India and the number of banks in India. Since the population size is quite small, the preparation of a sampling frame does not create any problem.

Another problem with these (SRSWR and SRSWOR) designs is that we may not get a representative sample using such a scheme. Consider an example of a locality having 10,000 households, out of which 5,000 belong to low-income group, 3,500 belong to middle income group and the remaining 1,500 belong to high-income group. Suppose it is decided to take a sample of 100 households using the simple random sampling. The selected sample may not contain even a single household belonging to the high- and middle-income group and only the low-income households may get selected, thus, resulting in a non-representative sample.

6.2.2 Systematic Sampling

Systematic sampling takes care of the limitation of the simple random sampling that the sample may not be a representative one. In this design, the entire population is arranged in a particular order. The order could be the calendar dates or the elements of a population arranged in an ascending or a descending order of the magnitude which may be assumed as random. List of subjects arranged in the alphabetical order could also be used and they are usually assumed to be random in order. Once this is done, the steps followed in the systematic sampling design are as follows:

- First of all, a sampling interval given by K = N/n is calculated, where N = the size of the population and n = the size of the sample. It is seen that the sampling interval K should be an integer. If it is not, it is rounded off to make it an integer.
- A random number is selected from 1 to K. Let us call it C.
- The first element to be selected from the ordered population would be C, the next element would be C + K and the subsequent one would be C + 2K and so on till a sample of size n is selected.

This way we can get representation from all the classes in the population and overcome the limitations of the simple random sampling. To take an example, assume that there are 1,000 grocery shops in a small town. These shops could be arranged in an ascending order of their sales,

with the first shop having the smallest sales and the last shop having the highest sales. If it is decided to take a sample of 50 shops, then our sampling interval K will be equal to 1000 ÷ 50 = 20. Now we select a random number from 1 to 20. Suppose the chosen number is 10. This means that the shop number 10 will be selected first and then shop number 10 + 20 = 30 and the next one would be 10 + 2 × 20 = 50 and so on till all the 50 shops are selected. This way we can get a representative sample in the sense that it will contain small, medium and large shops.

It may be noted that in a systematic sampling the first unit of the sample is selected at random (probability sampling design) and having chosen this, we have no control over the subsequent units of sample (non-probability sampling). Because of this, this design at times is called mixed sampling.

The main advantage of systematic sampling design is its simplicity. When sampling from a list of population arranged in a particular order, one can easily choose a random start as described earlier. After having chosen a random start, every K^{th} item can be selected instead of going for a simple random selection. This design is statistically more efficient than a simple random sampling, provided the condition of ordering of the population is satisfied.

The use of systematic sampling is quite common as it is easy and cheap to select a systematic sample. In systematic sampling one does not have to jump back and forth all over the sampling frame wherever random number leads, and neither does one have to check for duplication of elements as compared to simple random sampling. Another advantage of a systematic sampling over simple random sampling is that one does not require a complete sampling frame to draw a systematic sample. The investigator may be instructed to interview every 10^{th} customer entering a mall without a list of all customers.

There may be situations where it may not be possible to get a representative sample. The design can create problems if the sampling interval is a whole number multiple of some cycle related to the problem. On this design there may be a problem that there is a high probability of systematic bias creeping into the sample resulting in a non-representative sample. Consider, for example, the case of a certain PVR cinema hall where there may be a couple of snack bars. We may be interested in estimating the average daily sales of a particular snack bar in that PVR. Now, using the daily data with the population and sample size known, we compute a sampling interval which may be a multiple of seven. Using this, we may select our first element which would reflect one of the seven days of the week, say Friday. The next element would also be Friday, as our sampling interval is a multiple of seven and so the subsequent elements of the population. Therefore, our sample would comprise only Fridays and the sample would

not reflect day of the week variation in the sales data, which could result in a non-representative sample. Therefore, while using daily data, care should be taken that our sampling interval is not a multiple of seven.

6.2.3 Stratified Random Sampling

Under this sampling design, the entire population (universe) is divided into strata (groups), which are mutually exclusive and collectively exhaustive. By mutually exclusive, it is meant that if an element belongs to one stratum, it cannot belong to any other stratum. Strata are collectively exhaustive if all the elements of various strata put together completely cover all the elements of the population. The elements are selected using a simple random sampling independently from each group.

There are two reasons for using a stratified random sampling rather than simple random sampling. One is that the researchers are often interested in obtaining data about the component parts of a universe. For example, the researcher may be interested in knowing the average monthly sales of cell phones in 'large', 'medium' and 'small' stores. In such a case, separate sampling from within each stratum would be called for. The second reason for using a stratified random sampling is that it is more efficient as compared to a simple random sampling. This is because dividing the population into various strata increases the representativness of the sampling as the elements of each stratum are homogeneous to each other.

There are certain issues that may be of interest while setting up a stratified random sample. These are:

What criteria should be used for stratifying the universe (population)?

The criteria for stratification should be related to the objectives of the study. The entire population should be stratified in such a way that the elements are homogeneous within the strata, whereas there should be heterogeneity between strata. As an example, if the interest is to estimate the expenditure of households on entertainment, the appropriate criteria for stratification would be the household income. This is because the expenditure on entertainment and household income are highly correlated. As another example, if the objective of the study is to estimate the amount of money spent on cosmetics, then, gender could be used as an appropriate criteria for stratification. This is because it is known that though both men and women use cosmetics, the expenditure by women is much more than that of their male counterparts. Someone may argue out that gender may no longer remain the appropriate criteria if it is not backed by income. Therefore, the researcher might have to use two or more criteria for stratification depending upon the problem in hand. This would only increase the number of strata thereby making the sampling difficult.

Generally stratification is done on the basis of demographic variables like age, income, education and gender. Customers are usually stratified

on the basis of life stages and income levels to study their buying patterns. Companies may be stratified according to size, industry, profits for analysing the stock market reactions.

How many strata should be constructed?

Going by common sense, as many strata as possible should be used so that the elements of each stratum will be as homogeneous as possible. However, it may not be practical to increase the number of strata and, therefore, the number may have to be limited. Too many strata may complicate the survey and make preparation and tabulation difficult. Costs of adding more strata may be more than the benefit obtained. Further, the researcher may end up with the practical difficulty of preparing a separate sampling frame as the simple random samples are to be drawn from each stratum.

What should be appropriate number of samples size to be taken in each stratum?

This question pertains to the number of observations to be taken out from each stratum. At the outset, one needs to determine the total sample size for the universe and then allocate it between each stratum. This may be explained as follows:

Let there be a population of size N. Let this population be divided into three strata based on a certain criterion. Let N_1, N_2 and N_3 denote the size of strata 1, 2 and 3 respectively, such that $N = N_1 + N_2 + N_3$. These strata are mutually exclusive and collectively exhaustive. Each of these three strata could be treated as three populations. Now, if a total sample of size n is to be taken from the population, the question arises that how much of the sample should be taken from strata 1, 2 and 3 respectively, so that the sum total of sample sizes from each strata adds up to n.

Let the size of the sample from first, second and third strata be n_1, n_2, and n_3 respectively such that $n = n_1 + n_2 + n_3$. Then, there are two schemes that may be used to determine the values of n_i, (i = 1, 2, 3) from each strata. These are proportionate and disproportionate allocation schemes.

Proportionate allocation scheme: In this scheme, the size of the sample in each stratum is proportional to the size of the population of the strata. As an example, if a bank wants to conduct a survey to understand the problems that its customers are facing, it may be appropriate to divide them into three strata based upon the size of their deposits with the bank. If we have 10,000 customers of a bank in such a way that 1,500 of them are big account holders (having deposits more than ₹10 lakh), 3,500 of them are medium sized account holders (having deposits of more than ₹2 lakh but less than ₹10 lakh), the remaining 5,000 are small account holders (having deposits of less than ₹2 lakh). Suppose the total budget for sampling is fixed at ₹20,000 and the cost of sampling a unit (customer) is ₹20. If a sample of 100 is to be chosen from all the three strata, the size of the sample from strata 1 would be:

$$n_1 = n \times \frac{N_1}{N} = 100 \times \frac{1500}{10000} = 15$$

The size of sample from strata 2 would be:

$$n_2 = n \times \frac{N_2}{N} = 100 \times \frac{3500}{10000} = 35$$

The size of sample from strata 3 would be:

$$n_3 = n \times \frac{N_3}{N} = 100 \times \frac{5000}{10000} = 50$$

This way the size of the sample chosen from each stratum is proportional to the size of the stratum. Once we have determined the sample size from each stratum, one may use the simple random sampling or the systematic sampling or any other sampling design to take out samples from each of the strata.

Disproportionate allocation: As per the proportionate allocation explained above, the sizes of the samples from strata 1, 2 and 3 are 15, 35 and 50 respectively. As it is known that the cost of sampling of a unit is ₹20 irrespective of the strata from where the sample is drawn, the bank would naturally be more interested in drawing a large sample from stratum 1, which has the big customers, as it gets most of its business from strata 1. In other words, the bank may follow a disproportionate allocation of sample as the importance of each stratum is not the same from the point of view of the bank. The bank may like to take a sample of 45 from strata 1 and 40 and 15 from strata 2 and 3 respectively. Also, a large sample may be desired from the strata having more variability.

6.2.4 Cluster Sampling

In the cluster sampling, the entire population is divided into various clusters in such a way that the elements within the clusters are heterogeneous. However, there is homogeneity between the clusters. This design, therefore, is just the opposite of the stratified sampling design, where there was homogeneity within the strata and heterogeneity between the strata. To illustrate the example of a cluster sampling, one may assume that there is a company having its corporate office in a multi-storey building. In the first floor, we may assume that there is a marketing department where the offices of the president (marketing), vice president (marketing) and so on to the level of management trainee (marketing) are there. Naturally, there would be a lot of variation (heterogeneity) in the amount of salaries they draw and hence a high amount of variation in the amount of money spent on entertainment. Similarly, if the finance department is housed on the second floor, we may find almost a similar pattern. Same could be assumed for third, fourth and other floors. Now, if each of the floors could be treated as a cluster, we find that there is homogeneity between the clusters but there is a lot of heterogeneity within the clusters. Now, a sample of, say, 2 to 3 clusters is chosen at random

and once having done so, each of the cluster is enumerated completely to be able to make an estimate of the amount of money the entire population spends on entertainment.

Examples of cluster sampling could include ad hoc organizational committees drawn from various departments to advise the CEO of a company on product development, new product ideas, evaluating alternative advertising programmes, budget allocations and marketing strategies. Each of the clusters comprises a heterogeneous collection of members with different interests, background, experience, value system and philosophy. The CEO of the company may be able to take strategic decisions based upon their combined advice.

Although the per unit costs of cluster sampling are much lower than those of other probability sampling, the applicability of cluster sampling to an organizational context may be questioned as a cluster may not contain heterogeneous elements. The condition of heterogeneity within the cluster and homogeneity between the clusters may not be met. As another example, the households in a block are to be similar rather than dissimilar and as a result, it may be difficult to form heterogeneous clusters.

Cluster sampling is useful when populations under a survey are widely dispersed and drawing a simple random sample may be impractical.

Check Your Progress

1. What do you understand by sampling design?
2. Name the sub-types of probability sampling design.
3. What is systematic sampling?
4. How does proportionate allocation scheme work?

6.3 NON-PROBABILITY SAMPLING DESIGN

Under the non-probability sampling, the following designs would be considered—convenience sampling, purposive (judgemental) sampling, snowball sampling and quota sampling.

6.3.1 Convenience Sampling

Convenience sampling is used to obtain information quickly and inexpensively. The only criterion for selecting sampling units in this scheme is the convenience of the researcher or the investigator. Mostly, the convenience samples used are neighbours, friends, family members, colleagues and 'passers-by'. This sampling design is often used in the pre-test phase of a

research study such as the pre-testing of a questionnaire. Some of the examples of convenience sampling are:

- People interviewed in a shopping centre for their political opinion for a TV programme.
- Monitoring the price level in a grocery shop with the objective of inferring the trends in inflation in the economy.
- Requesting people to volunteer to test products.
- Using students or employees of an organization for conducting an experiment.
- Interviews conducted by a TV channel of people coming out of a cinema hall, to seek their opinion about the movie.
- A researcher visiting a few shops near his residence to observe which brand of a particular product people are buying, so as to draw a rough estimate of the market share of the brand.

In all the above situations, the sampling unit may either be self-selected or selected because of ease of availability. No effort is made to choose a representative sample. Therefore, in this design the difference between the population value (parameters) of interest and the sample value (statistic) is unknown both in terms of the magnitude and direction. Therefore, it is not possible to make an estimate of the sampling error and researchers won't be able to make a conclusive statement about the results from such a sample. It is because of this, convenience sampling should not be used in conclusive research (descriptive and causal research).

Convenience sampling is commonly used in exploratory research. This is because the purpose of an exploratory research is to gain an insight into the problem and generate a set of hypotheses which could be tested with the help of a conclusive research. When very little is known about a subject, a small-scale convenience sampling can be of use in the exploratory work to help understand the range of variability of responses in a subject area.

6.3.2 Judgemental Sampling

Under judgemental sampling, experts in a particular field choose what they believe to be the best sample for the study in question. The judgement sampling calls for special efforts to locate and gain access to the individuals who have the required information. Here, the judgement of an expert is used to identify a representative sample. For example, the shoppers at a shopping centre may serve to represent the residents of a city or some of the cities may be selected to represent a country. Judgemental sampling design is used when the required information is possessed by a limited number/category of people. This approach may not empirically produce satisfactory results and, may, therefore, curtail generalizability of the findings due to the fact that we

are using a sample of experts (respondents) that are usually conveniently available to us. Further, there is no objective way to evaluate the precision of the results. A company wanting to launch a new product may use judgemental sampling for selecting 'experts' who have prior knowledge or experience of similar products. A focus group of such experts may be conducted to get valuable insights. Opinion leaders who are knowledgeable are included in the organizational context. Enlightened opinions (views and knowledge) constitute a rich data source. A very special effort is needed to locate and have access to individuals who possess the required information.

The most common application of judgemental sampling is in business-to-business (B to B) marketing. Here, a very small sample of lead users, key accounts or technologically sophisticated firms or individuals is regularly used to test new product concepts, producing programmes, etc.

6.3.3 Snowball Sampling

Snowball sampling is generally used when it is difficult to identify the members of the desired population, e.g., deep-sea divers, families with triplets, people using walking sticks, doctors specializing in a particular ailment, etc. Under this design each respondent, after being interviewed, is asked to identify one or more in the field. This could result in a very useful sample. The main problem is in making the initial contact. Once this is done, these cases identify more members of the population, who then identify further members and so on. It may be difficult to get a representative sample. One plausible reason for this could be that the initial respondents may identify other potential respondents who are similar to themselves. The next problem is to identify new cases.

6.3.4 Quota Sampling

In quota sampling, the sample includes a minimum number from each specified subgroup in the population. The sample is selected on the basis of certain demographic characteristics such as age, gender, occupation, education, income, etc. The investigator is asked to choose a sample that conforms to these parameters. Field workers are assigned quotas of the sample to be selected satisfying these characteristics.

A researcher wants to measure the job satisfaction level among the employees of a large organization and believes that the job satisfaction level varies across different types of employees. The organization is having 10 per cent, 15 per cent, 35 per cent and 40 per cent, class I, class II, class III and class IV, employees, respectively. If a sample of 200 employees is to be selected from the organization, then 20, 30, 70 and 80 employees from class I, class II, class III and class IV respectively should be selected from the population. Now, various investigators may be assigned quotas from each class in such a way that a sample of 200 employees is selected from various classes in the

same proportion as mentioned in the population. For example, the first field worker may be assigned a quota of 10 employees from class I, 15 from class II, 20 from class III and 30 from class IV. Similarly, a second investigator may be assigned a different quota such that a total sample of 200 is selected in the same proportion as the population is distributed. Please note that the investigators may choose the employees from each class as conveniently available to them. Therefore, the sample may not be totally representative of the population, hence the findings of the research cannot be generalized. However, the reason for choosing this sampling design is the convenience it offers in terms of effort, cost and time.

In the example given above, it may be argued that job satisfaction is also influenced by education level, categorized as higher secondary or below, graduation, and postgraduation and above. By incorporating this variable, the distribution of population may look as given in Table 6.2. From the table, we may note that there are 8 per cent class I employees who are postgraduate and above, there are 35 per cent class IV employees with a higher secondary education and below and so on. Now, suppose a sample of size 200 is again proposed. In this case, the distribution of sample satisfying these two conditions in the same proportion in the population is given in Table 6.3.

Table 6.2 *Distribution of population (percentage)*

Education	Category of Employees				
	Class I	Class II	Class III	Class IV	Total
Postgraduation and above	8	5	5	0	18
Graduation	2	10	20	5	37
Higher Secondary and below	0	0	10	35	45
Total	10	15	35	40	100

Table 6.3 *Distribution of sample (numbers)*

Education	Category of Employees				
	Class I	Class II	Class III	Class IV	Total
Postgraduation and above	16	10	10	0	36
Graduation	4	20	40	10	74
Higher Secondary and below	0	0	20	70	90
Total	20	30	70	80	200

Table 6.3 indicates that a sample of 20 class II employees who are graduates should be selected. Likewise, a sample of 10 employees who possess postgraduate and above education should be selected. In the above table, the sample to be taken from each of the 12 cells has been specified. Having done so, each of the investigators is assigned a quota to collect

information from the employees conforming to the above norms so that a sample of 200 is selected.

Quota sampling design may look similar to the stratified random sampling design. However, there are differences between the two. In the stratified sampling design, the selection of sample from each stratum is random but in the quota sampling, the respondents may be chosen at the convenience or judgement of the researchers. Further, as already stated, the results of stratified random sampling could be generalized, whereas it may not be possible in the case of quota sampling. Quota sampling has some advantages over the probabilistic techniques. This design is very economical and it does not take too much time to set it up. Also, the use of this design does not require a sampling frame.

However, quota sampling also has certain weaknesses like:

- The total number of cells depends upon the number of control characteristics associated with the objectives of the study. If the control characteristics are large, the total number of cells increases, which may result in making the task of the investigator difficult.
- The chosen control characteristics should be related to the objectives of the study. The findings of the study could be misleading if any relevant parameter is omitted for one reason or the other.
- The investigator may visit those places where the chances of getting the respondents with the required control characteristics are high. The investigator could also avoid some responses that appear to be unfriendly. All this could result in making the findings of the study less reliable.

Check Your Progress

5. What is convenience sampling?
6. What is the purpose of snowball sampling?

6.4 CHALLENGES OF SAMPLING

The most challenging task in sampling is to minimize errors. Researchers often face a dilemma in their attempt to reduce sampling errors by increasing the sample size. For, as one increases the sample size, the non-sampling errors tend to increase. Controlling these requires that the sample size be small, but only at the cost of sampling errors getting large. Hence, the remedy offered begets contradiction.

The small sample remedy for controlling the non-sampling errors is effective only when the degree of variability in the population observations is not large. Where the variability is large, it is advisable to control sampling

errors by increasing the sample size, and to minimize the non-sampling errors through a more rigorous planning and execution of the sample survey. Better coordination and effective control over different survey activities do help yield the desired results.

6.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Sampling design refers to the process of selecting samples from a population. There are two types of sampling designs – probability sampling design and non-probability sampling design.
2. Under the probability sampling design, the following sampling designs would be covered—simple random sampling with replacement (SRSWR), simple random sampling without replacement (SRSWOR), systematic sampling, stratified random sampling and cluster sampling.
3. Systematic sampling is a type of probability sampling method in which the entire population is arranged in a particular order according to a design.
4. In the proportionate allocation scheme, the size of the sample in each stratum is proportional to the size of the population of the strata. As an example, if a bank wants to conduct a survey to understand the problems that its customers are facing, it may be appropriate to divide them into three strata based upon the size of their deposits with the bank.
5. Convenience sampling is used to obtain information quickly and inexpensively. The only criterion for selecting sampling units in this scheme is the convenience of the researcher or the investigator. Mostly, the convenience samples used are neighbours, friends, family members, colleagues and 'passers-by'.
6. Snowball sampling is generally used when it is difficult to identify the members of the desired population, e.g., deep-sea divers, families with triplets, people using walking sticks, doctors specializing in a particular ailment, etc. Under this design each respondent, after being interviewed, is asked to identify one or more in the field. This could result in a very useful sample.

6.6 SUMMARY

- Under probability sampling design, the following sampling designs would be covered—simple random sampling with replacement (SRSWR), simple random sampling without replacement (SRSWOR), systematic sampling, stratified random sampling and cluster sampling.

- In simple random sampling with replacement, a list of all the population from where the samples to be drawn is prepared.
- In the case of simple random sample without replacement, the procedure is identical to what was explained in the case of simple random sampling with replacement. The only difference here is that the chosen slip is not placed back in the box.
- The simple random sampling (with or without replacement) is not used in a consumer research. This is because in a consumer research the population size is usually very large, which creates problems in the preparation of a sampling frame.
- Systematic sampling takes care of the limitation of the simple random sampling that the sample may not be a representative one. In this design, the entire population is arranged in a particular order.
- The use of systematic sampling is quite common as it is easy and cheap to select a systematic sample. In systematic sampling one does not have to jump back and forth all over the sampling frame wherever random number leads, and neither does one have to check for duplication of elements as compared to simple random sampling.
- Under the stratified random sampling, the entire population (universe) is divided into strata (groups), which are mutually exclusive and collectively exhaustive.
- In the cluster sampling, the entire population is divided into various clusters in such a way that the elements within the clusters are heterogeneous. However, there is homogeneity between the clusters. This design, therefore, is just the opposite of the stratified sampling design, where there was homogeneity within the strata and heterogeneity between the strata.
- Under the non-probability sampling, the following designs would be considered—convenience sampling, purposive (judgemental) sampling, snowball sampling and quota sampling.
- Convenience sampling is used to obtain information quickly and inexpensively. The only criterion for selecting sampling units in this scheme is the convenience of the researcher or the investigator.
- Under judgemental sampling, experts in a particular field choose what they believe to be the best sample for the study in question. The judgement sampling calls for special efforts to locate and gain access to the individuals who have the required information.
- Snowball sampling is generally used when it is difficult to identify the members of the desired population, e.g., deep-sea divers, families with triplets, people using walking sticks, doctors specializing in a particular ailment, etc.

- In quota sampling, the sample includes a minimum number from each specified subgroup in the population. The sample is selected on the basis of certain demographic characteristics such as age, gender, occupation, education, income, etc.

6.7 KEY WORDS

- **Sampling Interval:** It refers to the distance or time between which measurements are taken, or data is recorded.
- **Cluster sampling:** It refers to a type of sampling method wherein the researcher divides the population into separate groups, called clusters. Then, a simple random sample of clusters is selected from the population.
- **Snowball sampling:** It is a nonprobability sampling technique where existing study subjects recruit future subjects from among their acquaintances. Thus the sample group is said to grow like a rolling snowball.

6.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. What is the difference between simple random sampling with replacement and without replacement?
2. What are the steps to calculate systematic sampling?
3. What are the advantages of using stratified random sampling over simple random sampling?
4. Give some examples of convenience sampling.
5. List the advantages and disadvantages of quota sampling.

Long Answer Questions

1. Explain different types of probability sampling designs.
2. Discuss the advantages of systematic sampling design.
3. Describe various types of non-probability sampling designs.

6.9 FURTHER READINGS

Chandan, J.S., Jagjit Singh and K.K. Khanna. 1995. *Business Statistics*, 2nd edition. New Delhi: Vikas Publishing House Pvt. Ltd.

NOTES

Croxton, Frederick E., and Dudley J. Cowden. 1943. *Applied General Statistics*. New York: Prentice Hall.

Gupta, S.P. 2006. *Statistical Methods*. New Delhi: S. Chand & Company.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods*, 23rd edition. New Delhi: Vikas Publishing House Pvt.Ltd..

Levin, Richard I. and David S. Rubin. 1998. *Statistics for Management.* New Jersey: Prentice Hall.

UNIT 7 DESIGN OF QUESTIONNAIRE

NOTES

Structure

7.0 INTRODUCTION

The questionnaire is a structured technique for collecting primary data in a marketing survey. It is a series of written or verbal questions for which the respondent provides answers. A well-designed questionnaire motivates the respondent to provide complete and accurate information.

When it comes to designing a questionnaire, asking the right questions in the right way is crucial to collect exactly the data you need.

7.1 OBJECTIVES

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

- Describe the different types of questionnaires
- Explain the questionnaire design procedure
- Understand the method of administration
- Discuss the content of questionnaire

7.2 CRITERIA FOR QUESTIONNAIRE DESIGNING

When one is designing the questionnaire, there are certain criteria that must be kept in mind.

The first and foremost requirement is that the spelt-out research objectives must be converted into clear questions which will extract answers

NOTES

from the respondent. This is not as easy as it sounds, for example, if one wants to know something like what is the margin that a company gives to the retailer? This cannot be converted into a direct question as no one will give the correct figure. Thus, one will have to ask a disguised question like may be a range of percentage estimates—2–5 per cent, 6–10 per cent, 11–15 per cent, 16–20 per cent, etc., or the retailer might not go beyond a yes, no or 'industry standard'.

The second requirement is, like the Toyota questionnaire, it should be designed to engage the respondent and encourage a meaningful response. For example, a questionnaire measuring stress cannot have a voluminous set of questions which fatigue the subject. The questions, thus, should be non-threatening, must encourage response and be clear to understand. One needs to remember that the essential usage of the instrument is to administer the same to a large base, thus there must be clarity and interest that should be part of the measure itself.

Lastly, the questions should be self-explanatory and not confusing as then the answers one gets might not be accurate or usable for analysis.

Types of Questionnaire

There are many different types of questionnaire available to the researcher. The categorization can be done on the basis of a variety of parameters. The two which are most frequently used for designing purposes are the degree of construction or structure and the degree of concealment, of the research objectives. Construction or formalization refers to the degree to which the response category has been defined. Concealed refers to the degree to which the purpose of the study is explained or is clear to the respondent.

Instead of considering them as individual types, most research studies use a mixed format.

Formalized and unconcealed questionnaire: This is the one that is indiscriminately and most frequently used by all management researchers. For example, if a new brokerage firm wants to understand the investment behaviour of the population under study, they would structure the questions and answers as follows:

1.Do you carry out any investment(s)?

Yes _____ No _____

If yes, continue, else terminate.

2.Out of the following options, where do you invest (tick all that apply).

Precious metals _____, real estate _____, stocks _____,

government instruments _____, mutual funds _____,

any other _____.

3.Who carries out your investments?

Myself _____, agent _____, relative _____, friend _____,

any other _____.

In case the option ticked is self, please go to Q. 4, else skip.

4.What is your source of information for these decisions?

Newspaper _____, investment magazines _____, company records, etc. _____, trading portals _____, agent _____.

NOTES

This kind of structured questionnaire is easy to administer, as one can see that the questions are self-explanatory and, since the answer categories are defined as well, the respondent needs to read and tick the right answer. Another advantage with this form is that it can be administered effectively to a large number of people at the same time. Data tabulation and data analysis is also easier to compute than in other methods.

This format, as a consequence of its predefined composition, is able to produce relatively stable results and is reasonably high in its reliability. The validity, of course would be limited as the comprehensive meaning of the constructs and variables under study might not be holistic when it comes to structured and limited responses. In such cases, variables are made a part of the study and some open-ended questions as well as administration/ additional instructions/probing by the field investigator could help in getting better results.

Formalized and concealed questionnaire: The research studies which are trying to unravel the latent causes of behaviour cannot rely on direct questions. Thus, the respondent has to be given a set of questions that can give an indication of what are his basic values, opinions and beliefs, as these would influence how he would react to certain products or issues. For example, a publication house that wants to launch a newspaper wants to ascertain what are the general perceptions and current attitudes about newspapers. Asking a direct question would only reveal apparent information, thus, some disguised attitudinal questions would need to be asked in order to infer this.

Please indicate your level of agreement with the following statements:

SA – Strongly Agree; A – Agree; N – Neutral; D – Disagree; SD – Strongly Disagree

		SA	A	N	D	SD
1	The individual today is better informed about everything than before.					
2	I believe that one must live for the day and worry about tomorrow later.					

NOTES

3	An individual must at all times keep abreast of what is happening in the world around him/her.					
4	Books are the best friends anyone can have.					
5	I generally read and then decide what to buy.					
6	My lifestyle is so hectic that I do not have time for reading the newspaper.					
7	The advent of radio, television and Internet have made the traditional information sources-like newspapers, redundant.					
8	A man/woman is known by what he/she reads.					

The logic behind these tests of attitude is that the questions do not seem to be in a particular direction and are apparently non-threatening, thus the respondent gives an answer which would be in the general direction of his/her attitudes.

The advantage of these questions is that since these are structured, one can ascertain their impact and quantify the same through statistical techniques. Secondly, it has been found that psychographic questions like these increase the subject coverage and improve the validity of the instrument as well. Most studies interested in quantifying the primary response data make use of questions that are designed both as formalized unconcealed and formalized concealed.

Non-formalized unconcealed: Some researchers argue that the respondent is not really cognizant of his/her attitude towards certain things. Also, this method asks him to give structured responses to attitudinal statements that essentially express attitudes in a manner that the researcher or experts think is the correct way. This however might not be the way the person thinks. Thus, rather than giving them pre-designed response categories, it is better to give them unstructured questions where he has the freedom of expressing himself the way he wants to. Some examples of these kinds of questions are given below:

1. What has been the reason for the success of the 'lean management drive' that the organization has undertaken? Please specify FIVE most significant reasons according to YOU.
   (a) _____
   (b) _____
   (c) _____
   (d) _____
   (e) _____
2. Why do you think Maggi noodles are liked by young children? ____
   _____
   _____

3. How do you generally decide on where you are going to invest your money? _____

_____

4. Give THREE reasons why you believe that the Commonwealth 2010 Games have helped the country?

NOTES

The advantage of the method is that the respondent can respond in any way he/she believes is important. For example, for the last question, some people might respond by stating that it has boosted tourism in the country and contributed to the country's economy. Some might think it will encourage more international events to be held in the country. Some might also state that it is not a good idea and the government should instead be spending on improving the cause of the people who are below the poverty line.

Thus, one gets a comprehensive perspective on what the construct/product/policy means to the population at large; and at the micro level, what it means to people in different segments. The validity of these measures is higher than the previous two. However, quantification is a little tedious and one cannot go beyond frequency and percentages to represent the findings. The other problem is the researcher's bias which might lead to clubbing responses into categories which might not be homogenous in nature.

Non-formalized, concealed: If the objective of the research study is to uncover socially unacceptable desires and latent or subconscious and unconscious motivations, the investigator makes use of questions of low structure and disguised purpose. The presumption behind this is that if the argument, the situation or question is ambiguous, it is most likely that the revelation it would result in would be more rich and meaningful. The major weakness of projective types of questionnaires is that being of a low structure, the interpretation required is highly skilled. Cost, time and effort are additional elements which might curtail the use of these techniques. A study conducted to measure to which segment should men's personal care toiletries (especially moisturizers and fairness creams) be targeted, the investigator designed two typical bachelors' shopping lists. One with a number of monthly grocery products as well as the normal male toiletries like shaving blades, gels, shampoos, etc., and the other list had the same grocery products and male toiletries but it had two additional items—Fair and Handsome fairness cream and sensitive skin moisturizer. The list was given to 20 young men to conceptualize/describe the person whose list this is. The answers obtained were as follows:

List with Cream and Moisturizer	List without Cream and Moisturizer
65 per cent said this person was good looking	10 per cent said this man was good looking
5 per cent said typical male	39 per cent said 30 plus in age
25 per cent said a 20-year-old	90 per cent said rugged and manly

NOTES

48 per cent said has a girlfriend	38 per cent said has a girlfriend
46 per cent said has a boyfriend	No one spoke of boyfriend
26 per cent said spendthrift	21 per cent said thrifty
15 per cent said 'girly'	32 per cent said normal Indian male

Thus, as we can see, the normal Indian adult male is still going to take time to include beauty or cosmetic products into his normal personal care basket. Thus, it is wiser for the marketeers to target the younger metrosexual male who is a heavy spender.

Another useful way of categorizing questionnaires is on the method of administration. Thus, the questionnaire that has been prepared would necessitate a face-to-face interaction. In this case, the interviewer reads out each question and makes a note of the respondent's answers. This administration is called a *schedule*. It might have a mix of the questionnaire type as described in the section above and might have some structured and some unstructured questions. The investigator might also have a set of additional material like product prototypes or copy of advertisements. The investigator might also have a predetermined set of standardized questions or clarifications , which he can use to ask questions like 'why do you say that?' or 'can you explain this in detail' 'what I mean to ask is…….' The other kind is the *self-administered questionnaire*, where the respondent reads all the instructions and questions on his own and records his own statements or responses. Thus, all the questions and instructions need to be explicit and self-explanatory.

The selection of one over the other depends on certain study prerequisites.

Population characteristics: In case the population is illiterate or unable to write the responses, then one must as a rule use the schedule, as the questionnaire cannot be effectively answered by the subject himself.

Population spread: In case the sample to be studied is large and dispersed, then one needs to use the questionnaire. Also when the resources available for the study, time, cost and manpower are limited, then schedules become expensive to use and it is advisable to use self-administered questionnaire.

Study area: In case one is studying a sensitive topic, like organizational climate or quality of working life, where the presence of an investigator might skew the answers in a more positive direction, then it is better that one uses the questionnaire. However, in case the motives and feelings are not well-developed and structured, one might need to do additional probing and in that case a schedule is better. If the objective is to explore concepts or trace the reaction of the sample population to new ideas and concepts, a schedule is advisable.

Check Your Progress

1. What is the first and foremost requirement in designing a questionnaire?
2. What is the logic behind the tests of attitude?

NOTES

7.3 QUESTIONNAIRE DESIGN PROCEDURE

In the earlier section, the researcher must have understood the great advantage he has in case he uses a questionnaire for his research purpose. However, one of the most difficult steps in the entire research process is designing a well-structured instrument. A number of scholars have attempted to create structured and sequential guidelines to be used by a researcher, no matter what his/her interest area. While not following any particular school of thought, presented below is a standardized process that a researcher can follow.

These, of course, might need to be modified depending upon the objectives of research. The steps are indicative of what one needs to accomplish, however, the final document that emerges and the effectiveness of the measure in extracting the study-related information, depends entirely upon the individual understanding of the researcher to be able to:

- Effectively and comprehensively list out the research information areas.
- Convert these into meaningful research questions.
- Understand and use the language of the respondent.

The steps involved in designing a questionnaire are as follows (Figure 7.1): (1) Convert the research objectives into the information needed, (2) Method of administering the questionnaire, (3) Content of the questions, (4) Motivating the respondent to answer, (5) Determining the type of questions, (6) Question design criteria, (7) Determine the questionnaire structure, (8) Physical presentation of the questionnaire, (9) Pilot testing the questionnaire, (10) Standardizing the questionnaire.

Each of these would be discussed and illustrated in this section. The researcher needs to remember that these are not independent steps, where one needs to finish the first one to go on to the next one and so on. In the actual conduction, there might be a simultaneous conduction of some and one might not be able to draw clear cut boundaries between them. Also at times, the researcher might have to backtrack and modify an earlier task that he might have carried out.

NOTES

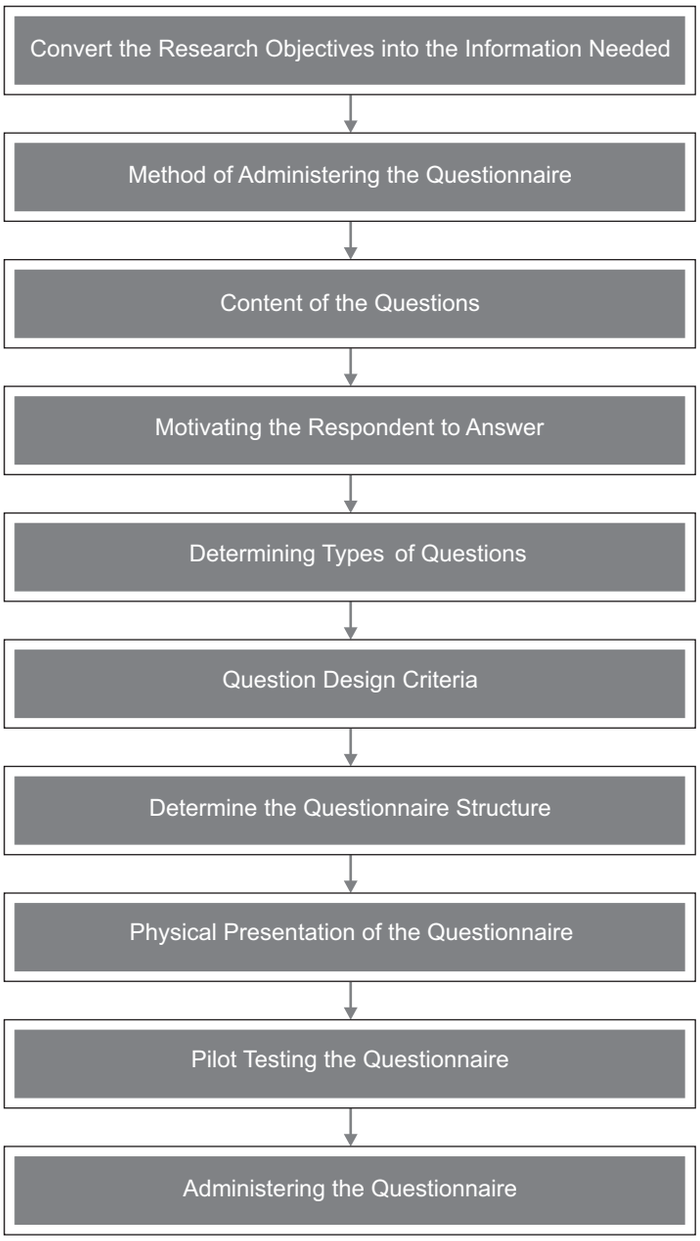


Fig. 7.1 *Questionnaire Design Process*

Convert the research objectives into information areas: This is the first step of the design process. As stated in the flowchart, this is the most critical stage and the researcher/investigator is assumed to have done considerable exploratory work to have crystallized objectives of the study. This is also the stage that requires formation of the research design of the study. Thus, by this stage one assumes that one has achieved the following tasks:

NOTES

- Spelt out clearly the specific research questions that the study will address.
- Converted these questions into statements of objectives.
- Operationalized the variables to be studied, i.e., the variables under study should have been clearly defined.
- Identified the direction of the relation or any other assumption one makes about the variables under study in the form of a hypothesis.
- Specified the information needed for the study, in this case one will look at the information needed from the primary data source.

Once these tasks are accomplished, one can prepare a tabled framework so that the questions which need to be developed become clear.

By this time, the respondent would have also developed a clear idea about the group that he would need to study. Thus, the characteristics of the population which might impact the constructs under study would also need to be studied in order to frame appropriate questions on these. At this stage, it might emerge that one needs to design separate questionnaires for the populations whose inputs are important, or have separate set of questions for those with different stands on the stated criteria. This stepwise process is explained in Table 7.1.

Table 7.1 *Framework for Identifying Information Needs*

Research Questions	Research Objectives	Variables to be Studied	Information (Primary Required)	Population to be Studied
What is the nature of plastic bag usage amongst people in the NCR (National Capital Region)?	To identify the different uses of plastic bags. To find out the method of disposal of plastic bags. To find out who uses plastic bags. To find out what is the level of consciousness that people have about the environment.	Usage behaviour Demographic details	Uses of plastic bags Disposal of plastic bags	Consumers Retailers

NOTES

What is the level of environment consciousness amongst them?	To find out whether they understand how plastic bags can be harmful to the environment. To identify strategies to discontinue plastic bag usage.	Environmental consciousness. Effect of plastic bag usage	Respondent attitudes and perceptions towards the environment Perception about the impact of plastic bags on the environment	Consumer Retailer
What measures can be taken to encourage people not to use plastic bags?		Corporation laws (if any) Attitudinal change strategies	Indicative measures for encouraging the general public to discontinue use of plastic bags	Policy maker Consumer Retailer

Method of administration: Once the researcher has identified his information area; he needs to specify how the information should be collected. The researcher usually has available to him a variety of methods for administering the study. The main methods are personal schedule self-administered questionnaire through mail, fax, e-mail and web-based. There are different preconditions for using one method over the other. Also once the decision has been taken about the method, one also needs to design different ways of asking the required information. Table 7.2 gives a template the researcher can use to take his administration decision and the kind of questions he must ask. As can be seen, a larger population can be covered by mail or fax. In case the population to be studied is computer literate, it is possible to use e-mail or web-designed surveys.

Table 7.2 *Mode of Administration and Design Implications*

	Schedule	**Telephone**	**Mail/Fax**	**E-mail**	**Web-Based**
Administrative control	high	medium	Low	low	low
Sensitive issues	high	medium	Low	low	low
New concept	high	medium	Low	low	low
Large sample	low	low	High	high	high
Cost/time taken	high	medium	Medium	low	low
Question structure	unstructured	either	structured	structured	structured
Sampling control	high	high	Medium	low	low
Response rate	high	high	Low	medium	low
Interviewer bias	high	high	low	low	low

NOTES

For a smaller population and more complex or sensitive issues, personal schedule is advisable. In computer-assisted dissemination (CAPI and CATI), complex skip and branching options are possible and randomization of questions to eliminate the order bias can be carried out with considerable ease. When the researcher wants to have a higher control over the way the questions are answered, i.e., the sequence and response time for answering, he should be using the schedule. By sampling control we mean who answers the questions. When one is interested in the decision maker's thought process and purchase process, one would not like to go to those users who might not always be the buyers, for example the housewife buying toothpaste for a toothpaste evaluation study is the respondent and not her son who might be using the toothpaste but who is, definitely, not the buyer. Sampling control, as we can see, is highest in schedule and lowest in a web-based survey.

As the researcher proceeds from one administration mode to another, the question structure and instructions change. The major reason for this is the presence or absence of the investigator. This has been illustrated in the example below.

Administration Mode and Question Structure

Schedule

Now I am going to give you a set of cards. Each card will have the name of one television serial (*Handover the cards to the respondent in a random order).* I want you to examine them carefully (*give her some time to read all the names)*. I would request you to hand over the card which has the name of the serial you like to watch the most. (*Record the serial and keep this card with you*). Now, of the remaining nine serials, name your next most favourite serial (*continue the same process till the person is left with the last card)*

	TV serial	Rank Order
1.	1	_____
2.	2	_____
3.	3	_____
4.	4	_____
5.	5	_____
6.	6	_____
7.	7	_____
8.	8	_____
9.	9	_____
10.	10	_____

NOTES

Telephone Questionnaire

Please listen very carefully; I am going to slowly read the names of ten popular TV serials. I want to know how much you prefer watching them. You need to use a 1 to 10 scale, where 1 means—I do not like watching it—and 10 means—I really like watching it. For those in between you may choose any number between 1 to 10. However, please remember that the higher the number, the more you like watching it. Now, I am going to name the serials one by one. In case the name is not clear, I will repeat the list again. So, the serial's name is _____. Please use a number between 1 to 10 as I had told you. Ok thank you, the next name is _____. And so on till all the 10 names have been read out and evaluated.

	Serial										
1.	*Balika Badhu*	1	2	3	4	5	6	7	8	9	10
2.	*Sathiya*	1	2	3	4	5	6	7	8	9	10
3.	*Sasural Genda Phool*	1	2	3	4	5	6	7	8	9	10
4.	*Bidai*	1	2	3	4	5	6	7	8	9	10
5.	*Pathshala*	1	2	3	4	5	6	7	8	9	10
6.	*Bandini*	1	2	3	4	5	6	7	8	9	10
7.	*Lapataganj*	1	2	3	4	5	6	7	8	9	10
8.	*Sajan Ghar Jaana Hai*	1	2	3	4	5	6	7	8	9	10
9.	*Tere Liye*	1	2	3	4	5	6	7	8	9	10
10.	*Uttaran*	1	2	3	4	5	6	7	8	9	10

Mail Questionnaire

In the next question you will find the names of ten popular Hindi serials that are being aired on television these days. You are requested to rank them in order of your preference. Start by identifying the serial which is your most favourite, to this you may give a rank of 1. Then from the rest of the nine, pick the second most preferred serial and give it a rank of 2. Please carry out this process till you have ranked all 10. The one you prefer the least should have a score of 10. You are also requested not to give two serials the same rank. The basis on which you decide to rank the serials is entirely dependent upon you. Once again, you are asked to rank all the 10 serials.

	Serial	**Rank Order**
1.	*Balika Badhu*	_____
2.	*Sathiya*	_____
3.	*Sasural Genda Phool*	_____
4.	*Bidai*	_____
5.	*Pathshala*	_____

6.	*Bandini*	_____
7.	*Lapataganj*	_____
8.	*Sajan Ghar Jaana Hai*	_____
9.	*Tere Liye*	_____
10.	*Uttaran*	_____

The pattern of instructions and the response structure for fax, e-mail and web surveys are similar. Thus, they have not been shown here separately.

NOTES

Content of the questionnaire: The next step, once the information needs and mode of administration has been decided, is to determine the matter to be included as questions in the measure. The decision to include or not include certain questions depends upon a certain criteria. Thus, the researcher needs to subject the questions designed by him to an objective quality check in order to ascertain what research objective/information need the question would be covering before using any of the framed questions.

How essential is it to ask the question? In the course of the research study, the researcher might formulate a number of questions which he thinks address the information needs of the study. Sometimes the researcher might find a particular question very intriguing or interesting and thus might decide to include it in the questionnaire. However, one needs to remember that the time of the respondent is precious and it should not be wasted. Unless a question is adding to the data required for reaching an answer to the formulated problem, it should not be included. For example, if one is studying the usage of plastic bags, then demographic questions on age group, occupation, education and gender might make sense but questions related to marital status, family size and the state to which the respondent belongs are not required as they have no direct relation with the usage or attitude towards plastic bags.

Sometimes, to gauge the information needs, the researcher might have to ask multiple questions, even though they might not seem to be related directly to the research objective. For example, instead of asking shopkeepers, who own a shop in a shopping centre, whether they would in the near future open an outlet in a mall, a set of questions were asked to understand the retailers' perception of shopping trends.

NOTES

Please indicate your level of agreement with the following statements:

SA – Strongly Agree; A – Agree; N – Neutral; D – Disagree; SD – Strongly Disagree

	Compared to the Past (5-10 years)	SA	A	N	D	SD
1	The individual customer today shops more					
2	The consumer is well-informed about market offerings					
3	The consumer knows what he/she wants to buy before he enters the store					
4	The consumer today has more money to spend					
5	There are more shopping options available to the consumer today					

There are also times, especially in self-administered questionnaires, when one may ask some neutral questions at the beginning of the questionnaire to establish an involvement and rapport. For example, for a biofertilizer usage study, the following question was asked:

Farming for you is a:

noble profession

ancestral profession

profession like any other

profession that is not lucrative

any other

Camouflaged or disguised questions are asked sometimes to keep the purpose or sponsorship of the project hidden. Here generally, the researcher might ask questions related to a set of brand names in the product category rather than asking questions only with reference to the company/brand one is interested in. For example, in a survey done on power drinks carried out by Gatorade, one might also have questions related to Powerade and Red Bull. Similar questions might be kept at different points in the study to assess the consistency of the respondent in answering. Questions like these add to the reliability of the scale.

Do we need to ask several questions instead of a single one? After deciding on the significance of the question, one needs to ascertain whether a single question will serve the purpose or should more than one question be asked. For example, in the TV serial study, assume that the second question after the ranking/rating question is:

'Why do you like the serial _____ (the one you ranked No. 1/prefer watching the most)?'

(Incorrect)

NOTES

Here, one lady might say, 'Everyone in my family watches it'. While another might say, 'It deals with the problems of living in a typical Indian joint family system' and yet another might say, 'My friend recommended it to me'. The first relates to joint decision-making by the family, the second relates to an attribute of the programme, while the third tells us what the information source was for her.

Thus, we need to ask her:

'What do you like about_____?'

'Who all in your household watch the serial?'

and

'How did you first hear about the serial?' (Correct)

Motivating the respondent to answer: The one thing the researcher must remember is that answering the questionnaire requires some effort on the part of the respondent. Thus, the questionnaire should be designed in a manner that it involves the respondent and motivates him/her to give comprehensive information. There might be two kinds of hindrances to active participation by the subject:

- The respondent might not be able to respond in the right manner.
- The respondent might be unwilling to part with the information.

We will discuss these situations and also understand how these need to be overcome, in order to be able to collect the data.

Assisting the respondent to provide the required information: There are three kinds of situations which might lead to inability to answer in a correct manner. Each of these is examined separately here:

Does the person have the required information? It has been found that once the respondents get into the rhythm of answering the questions, they answer questions even when they do not understand or have information about the construct being investigated. This is not because they are inherently dishonest; it is simply the result of confusion. For example, a young man whose personal care products are bought by his mother will not have any knowledge about the purchase process and decision. Yet, if asked, he will answer them based on his general understanding of the process.

Another situation might be when the person has had no experience with the issue being investigated. Look at the following question:

How do you evaluate the negotiation skills module, viz., the communication and presentation skill module?
(Incorrect)

In this case it might be that the person has not undergone one or even both the modules, so how can he compare? Thus, in situations where not all

NOTES

the respondents are likely to be informed about the research topic, certain qualifying or *filter questions* that measure the experience or knowledge must be asked before the questions about the topics themselves. Filter questions enable the researcher to filter out the respondents who are not adequately informed. Thus, the correct question would have been:

Have you been through the following training modules?

- Negotiation skills module Yes/no
- Communication and presentation skills Yes/no

In case the answer to both is yes, please answer the following question, or else move to the next question.

How do you evaluate the negotiation skills module, viz., the communication and presentation skill module? (Correct)

Does the person remember? Many a times, the question addressed might be putting too much stress on an individual's memory. All of us know that human memory might be short and yet sometimes while designing the questionnaire, one overlooks this. For example, consider the following questions:

How much did you spend on eating out last month? (Incorrect)

How many questions do you ask in a recruitment interview? (Incorrect)

As one can see, such questions far surpass any normal individual's memory bank. There have been a number of studies to demonstrate that people are generally not very good at remembering quantities. Usually, people forget significant events like birthdays or anniversaries. However, generally this is more related to pleasant days rather than bad days associated with accident or theft or even death anniversaries. Secondly, there is an element of the most recent events to remember. Thus, the employee will be able to better evaluate a training module that he attended last than those he attended in the whole year. A person remembers his recent big purchase details more than the last four major purchases.

Forgotten material can be drawn out by giving cues to stimulate the memory. These triggers are termed as *aided recall*. For example, unaided recall of TV serials could be measured by questions such as follows, 'Which TV serials did you watch last week?' The aided recall approach on the other hand would assist in recall by giving a list of serials aired in the last week and then ask. 'Which of these serials did you watch last week?'

Thus, the questions listed above could have been rephrased as follows:

When you go out to eat, on an average your bill amount is:

Less than ₹100

₹101–250

NOTES

₹251–500

More than ₹500

How often do you eat out in a week?

1–2 times.

3–4 times

5–6 times

Everyday (correct)

From the following, tick the areas on which you ask questions in a typical recruitment interview:

Educational background

Subject knowledge

Previous experience

General awareness

Individual information

Once the respondent ticks the relevant areas, then a number of questions from the indicated areas are asked. It is also possible to use the constant sum scale to indicate the percentage of questions asked from the area, so that the total adds up to 100 per cent.

Can the respondent articulate? The articulation does not refer to only enlisting the response. It also refers to not knowing what words to be used to articulate certain types of answers. For example, if you ask a respondent to:

- Describe a river rafting experience.
- The ambience of the new Levi's outlet. (Incorrect)

Most respondents would not know what phrases to use to give an answer. On the other hand, if the researcher uses a Semantic differential scale, the respondent can be provided adjectives to choose from. It must be remembered that if the person does not know what words to use or finds the task of description too tedious, the person will not fill in the answers. Thus, in the above case, one can provide answer categories to the person as follows:

Describe the river rafting experience. (Correct)

1	Unexciting		Exciting
2	Bad		Good
3	Boring		Interesting
4	Cheap		Expensive
5	Safe		Dangerous

NOTES

Assisting the respondent to answer: This is the second reason for not answering a question. It might happen that the person understands the question and also knows the answer, yet he is not willing to part with the information. We will discuss the situations which might result in this scenario.

The perspective is not clear: The questions that are being asked must possess face validity, i.e., they must not appear to be out of context with the other questions in the survey. Thus, a questionnaire which is measuring a person's quality of working life and poses questions as below will not be appreciated as the questions will seem to be suspicious and might be perceived as having a hidden agenda.

How many credit cards do you own?

When did you last go on a holiday?

How many movies do you watch in a fortnight?

People are not willing to answer questions they think do not make sense. Respondents are also hesitant about sharing personal demographic data such as age, income, and profession. Thus, the purpose of asking such questions has to be made explicit in the instructional note.

Thus, in the previous example, the researcher can justify that a spillover of a healthy quality of working life is also reflected in a person's way of living. Thus, we would like to know how you live.

In the second case of demographic data details, stating that 'We would like to determine which TV serials are preferred by people of different ages, incomes and professions, we need information on ...', will put the respondent at ease when sharing the data.

Sensitive information: There might be instances when the question being asked might be embarrassing to the respondents and thus they would not be comfortable in disclosing the data required. Sometimes, this might diminish the respondent's willingness to respond to the other questions as well. These topics could be related to income, family life, political and religious beliefs, and socially undesirable habits and desires. A number of techniques are available to reduce the respondent's hesitation.

- Make a generic statement to soothe the anxieties and state that 'these days most women consume alcoholic drinks at social gatherings, followed by a question on alcohol consumption. This technique is called *counter biasing.*
- Place the sensitive question in between some seemingly neutral questions and then ask the questions at a rapid speed.
- The best way to get answers on sensitive issues is to use *the third-person technique* and ask the question as related to other people.

For example, questions such as the following will not get any answers.

Have you ever used fake receipts to claim your medical allowance? (Incorrect)

Have you ever spit tobacco on the road (to tobacco consumers)? (Incorrect)

However, in case the socially undesirable habit is in the context of a third person, the chances of getting indicative correct responses are possible. Thus the questions should be rephrased as follows:

Do you associate with people who use fake receipts to claim their medical allowance? (Correct)

Do you think tobacco consumers spit tobacco on the road? (Correct)

- For certain demographic questions like income and age, instead of using the ratio scale one must use class intervals:

  'What is your household's annual income?' (Incorrect)

  'What is your household's annual income?'

  Under ₹25,000,
  ₹25,001–50,000,
  ₹50,001–75,000,
  Over ₹75,000. (Correct)

- For sensitive issues as stated earlier, it is much better to use unstructured questions and probe only after the respondent is comfortable with the investigator.

Check Your Progress

3. What are the steps involved in designing a questionnaire?
4. Why are camouflaged questions asked?

7.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The first and foremost requirement is that the spelt-out research objectives must be converted into clear questions which will extract answers from the respondent.
2. The logic behind these tests of attitude is that the questions do not seem to be in a particular direction and are apparently non-threatening, thus the respondent gives an answer which would be in the general direction of his/her attitudes.

NOTES

3. The steps involved in designing a questionnaire are as follows: (1) Convert the research objectives into the information needed, (2) Method of administering the questionnaire, (3) Content of the questions, (4) Motivating the respondent to answer, (5) Determining the type of questions, (6) Question design criteria, (7) Determine the questionnaire structure, (8) Physical presentation of the questionnaire, (9) Pilot testing the questionnaire, (10) Standardizing the questionnaire.
4. Camouflaged or disguised questions are asked sometimes to keep the purpose or sponsorship of the project hidden. Here generally, the researcher might ask questions related to a set of brand names in the product category rather than asking questions only with reference to the company/brand one is interested in.

7.5 SUMMARY

- When one is designing the questionnaire, there are certain criteria that must be kept in mind.
- The first and foremost requirement is that the spelt-out research objectives must be converted into clear questions which will extract answers from the respondent.
- There are many different types of questionnaire available to the researcher. The categorization can be done on the basis of a variety of parameters. The two which are most frequently used for designing purposes are the degree of construction or structure and the degree of concealment, of the research objectives.
- Formalized and unconcealed questionnaire is the one that is indiscriminately and most frequently used by all management researchers.
- This kind of structured questionnaire is easy to administer, as one can see that the questions are self-explanatory and, since the answer categories are defined as well, the respondent needs to read and tick the right answer.
- Formalized and concealed questionnaire tries to unravel the latent causes of behaviour cannot rely on direct questions.
- Another useful way of categorizing questionnaires is on the method of administration. Thus, the questionnaire that has been prepared would necessitate a face-to-face interaction.
- In this case, the interviewer reads out each question and makes a note of the respondent's answers. This administration is called a *schedule*.

- The steps involved in designing a questionnaire are as follows: (1) Convert the research objectives into the information needed, (2) Method of administering the questionnaire, (3) Content of the questions, (4) Motivating the respondent to answer, (5) Determining the type of questions, (6) Question design criteria, (7) Determine the questionnaire structure, (8) Physical presentation of the questionnaire, (9) Pilot testing the questionnaire, (10) Standardizing the questionnaire.

7.6 KEY WORDS

- **Questionnaire**: A questionnaire is a formalized set of questions for obtaining information from respondents. It must translate the information needed into a set of specific questions that the respondents can and will answer.
- **Schedule:** A schedule is a structure of a set of questions on a given topic which are asked by the interviewer or investigator personally.

7.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. Write a short note on formalised and unconcealed questionnaire.
2. State the advantage of concealed questionnaire.

Long Answer Questions

1. Analyse the criteria for questionnaire designing.
2. Describe the procedure of questionnaire design.

7.8 FURTHER READINGS

Chandan, J.S., Jagjit Singh and K.K. Khanna. 1995. *Business Statistics*, 2nd edition. New Delhi: Vikas Publishing House Pvt. Ltd.

Croxton, Frederick E., and Dudley J. Cowden. 1943. *Applied General Statistics*. New York: Prentice Hall.

Gupta, S.P. 2006. *Statistical Methods*. New Delhi: S. Chand & Company.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods*, 23rd edition. New Delhi: Vikas Publishing House Pvt.Ltd..

Levin, Richard I. and David S. Rubin. 1998. *Statistics for Management.* New Jersey: Prentice Hall.

UNIT 8 SAMPLING ERRORS

NOTES

Structure

8.0 INTRODUCTION

In statistics, sampling error is the error caused by observing a sample instead of the whole population. The sampling error is the difference between a sample statistic used to estimate a population parameter and the actual but unknown value of the parameter.

Sampling error is the deviation of the selected sample from the true characteristics, traits, behaviours, qualities or figures of the entire population.

8.1 OBJECTIVES

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

- Differentiate between sampling and non-sampling error
- Describe the reasons for non-sampling error
- Analyse the procedure for the calculation of standard error of mean

8.2 SAMPLING VS NON-SAMPLING ERROR

There are two types of errors that may occur while we are trying to estimate the population parameters from the sample. These are called sampling and non-sampling errors.

Sampling error: This error arises when a sample is not representative of the population. For example, if our population comprises 200 MBA students in a business school and we want to estimate the average height of these 200 students by taking a sample of 10 (say). Let us assume for the sake of

NOTES

simplicity that the true value of population mean (parameter) is known. When we estimate the average height of the sampled students, we may find that the sample mean is far away from the population mean. The difference between the sample mean and the population mean is called sampling error, and this could arise because the sample of 10 students may not be representative of the entire population. Suppose now we increase the sample size from 10 to 15, we may find that the sampling error reduces. This way, if we keep doing so, we may note that the sampling error reduces with the increase in sample size as an increased sample may result in increasing the representativeness of the sample.

Non-sampling error: This error arises not because a sample is not a representative of the population but because of other reasons. Some of these reasons are listed below:

- The respondents when asked for information on a particular variable may not give the correct answers. If a person aged 48 is asked a question about his age, he may indicate the age to be 36, which may result in an error and in estimating the true value of the variable of interest.
- The error can arise while transferring the data from the questionnaire to the spreadsheet on the computer.
- There can be errors at the time of coding, tabulation and computation.
- If the population of the study is not properly defined, it could lead to errors.
- The chosen respondent may not be available to answer the questions or may refuse to be part of the study.
- There may be a sampling frame error. Suppose the population comprises households with low income, high income and middle class category. The researcher might decide to ignore the low-income category respondents and may take the sample only from the middle and the high-income category people.

8.3 STANDARD ERROR

Standard error of the mean ($\sigma_{\bar{X}}$)is a measure of dispersion of the distribution of sample means and is similar to the standard deviation in a frequency distribution and it measures the likely deviation of a sample mean from the grand mean of the sampling distribution.

If all sample means are given, then ($\sigma_{\bar{X}}$) can be calculated as follows:

$$\sigma_{\bar{x}} = \sqrt{\frac{\Sigma(\bar{X} - \mu_{\bar{x}})}{N}}$$ where N = Number of sample means

NOTES

Thus we can calculate $\sigma_{\overline{X}}$ for Example 1 of the sampling distribution of the ages of 5 children as follows:

$\overline{X}$	$(\mu_{\overline{X}})$	$(\overline{X} - \mu_{\overline{X}})^2$
3	6	9
4	6	4
5	6	1
6	6	0
7	6	1
8	6	4
9	6	9
		$\Sigma (\overline{X} - \mu_{\overline{X}})^2 = 28$

Then,

$$\sigma_{\bar{x}} = \sqrt{\frac{\Sigma(\bar{x} - \mu_{\bar{x}})}{N}}$$

$$= \sqrt{\frac{28}{7}}$$

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

However, since it is not possible to take all possible samples from the population, we must use alternate methods to compute $\sigma_{\overline{X}}$.

The standard error of the mean can be computed from the following formula, if the population is finite and we know the population mean. Hence,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{(N-n)}{(N-1)}}$$ Where,

σ = Population standard deviation

N = Population size

n = Sample size

This formula can be made simpler to use by the fact that we generally deal with very large populations, which can be considered infinite, so that if the population size A' is very large and sample size n is small, as for example in the case of items tested from assembly line operations, then,

$\sqrt{\frac{(N-n)}{(N-1)}}$ would approach 1.

Hence,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The factor $\sqrt{\frac{(N-n)}{(N-n)}}$ is also known as the 'finite correction factor', and should be used when the population size is finite.

As this formula suggests, $\sigma_{\bar{X}}$ decreases as the sample size (w) increases, meaning that the general dispersion among the sample means decreases, meaning further that any single sample mean will become closer to the population mean, as the value of ($\sigma_{\bar{X}}$) decreases. Additionally, since according to the property of the normal curve, there is a 68.26 per cent chance of the population mean being within one $\sigma_{\bar{X}}$ of the sample mean, a smaller value of $\sigma_{\bar{X}}$ will make this range shorter; thus making the population mean closer to the sample mean (refer Example 1).

NOTES

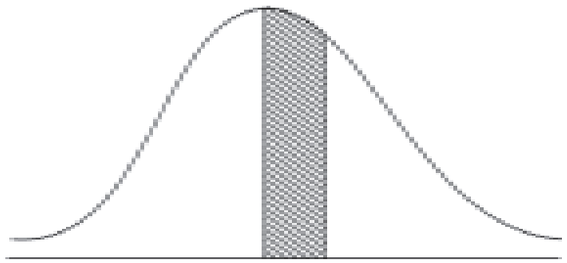
Example 1: The IQ scores of college students are normally distributed with the mean of 120 and standard deviation of 10.

(a) What is the probability that the IQ score of any one student chosen at random is between 120 and 125?

(b) If a random sample of 25 students is taken, what is the probability that the mean of this sample will be between 120 and 125.

Solution:

(a) Using the standardized normal distribution formula,



125

$\mu = 120$

$\sigma = 10$

$$Z = \frac{(X-\mu)}{\sigma}$$

$$Z = \frac{125-120}{10} = 5/10 = .5$$

The area for Z = .5 is 19.15.

This means that there is a 19.15 per cent chance that a student picked up at random will have an IQ score between 120 and 125.

(b) With the sample of 25 students, it is expected that the sample mean will be much closer to the population mean, hence it is highly likely that the sample mean would be between 120 and 125.

NOTES

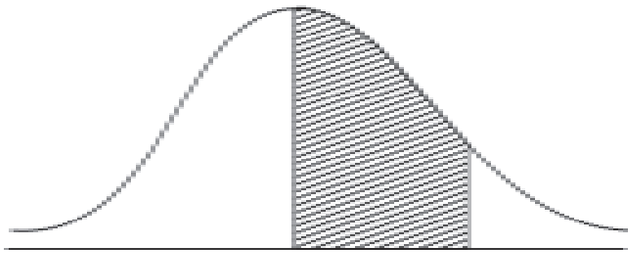
The formula to be used in the case of standardized normal distribution for sampling distribution of the means is given by,

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

where,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Hence,



125

μ =120

σ =10

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2$$

where,

Then,

$$Z = \frac{125 - 120}{2} = 5/2 = 2.5$$

The area for Z = 2.5 is 49.38.

This shows that there is a chance of 49.38 per cent that the sample mean will be between 120 and 125. As the sample size increases further, this chance will also increase. It can be noted that the probability of a sample mean being between 120 and 125 is much higher than the probability of an individual student having an IQ between 120 and 125.

Check Your Progress

1. State any one reason for non-sampling error.
2. What is standard error of mean?
3. Why does a sampling error arise?

8.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

NOTES

1. The error can arise while transferring the data from the questionnaire to the spreadsheet on the computer.
2. Standard error of the mean (σx) is a measure of dispersion of the distribution of sample means and is similar to the standard deviation in a frequency distribution and it measures the likely deviation of a sample mean from the grand mean of the sampling distribution.
3. Sampling error arises when a sample is not representative of the population.

8.5 SUMMARY

- There are two types of error that may occur while we are trying to estimate the population parameters from the sample. These are called sampling and non-sampling errors.
- Sampling error arises when a sample is not representative of the population.
- The difference between the sample mean and the population mean is called sampling error, and this could arise because the sample of 10 students may not be representative of the entire population.
- Non-sampling error arises not because a sample is not a representative of the population but because of other reasons.
- Standard error of the mean (σx)is a measure of dispersion of the distribution of sample means and is similar to the standard deviation in a frequency distribution and it measures the likely deviation of a sample mean from the grand mean of the sampling distribution.

8.6 KEY WORDS

- **Sampling error**: In statistics, sampling error is incurred when the statistical characteristics of a population are estimated from a subset, or sample, of that population.
- **Non-sampling error**: It is a catch-all term for the deviations of estimates from their true values that are not a function of the sample chosen, including various systematic errors and random errors that are not due to sampling.

NOTES

8.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. How is the standard error of mean calculated if all sample means are given?
2. How is the standard error of mean computed if the population is finite?

Long Answer Questions

1. Differentiate between sampling and non-sampling error.
2. With the help of examples, discuss the calculation of standard error of mean.

8.8 FURTHER READINGS

Chandan, J.S., Jagjit Singh and K.K. Khanna. 1995. *Business Statistics*, 2nd edition. New Delhi: Vikas Publishing House Pvt. Ltd.

Croxton, Frederick E., and Dudley J. Cowden. 1943. *Applied General Statistics*. New York: Prentice Hall.

Gupta, S.P. 2006. *Statistical Methods*. New Delhi: S. Chand & Company.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods*, 23rd edition. New Delhi: Vikas Publishing House Pvt.Ltd..

Levin, Richard I. and David S. Rubin. 1998. *Statistics for Management.* New Jersey: Prentice Hall.

BLOCK - III
COLLECTION AND TABULATION OF DATA

NOTES

UNIT 9 COLLECTION OF DATA

Structure

9.0 INTRODUCTION

Data collection is the process of gathering and measuring data, information or any variables of interest in a standardized and established manner that enables the collector to answer or test hypothesis and evaluate outcomes of the particular collection.

The data collection component of research is common to all fields of study including physical and social sciences, humanities, business, etc. While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same.

9.1 OBJECTIVES

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

- Differentiate between primary and secondary data
- Discuss the research applications of secondary data
- Explain internal and external data sources
- Differentiate between quantitative and qualitative data

NOTES

9.2 MEANING AND TYPES OF DATA

To understand the multitude of choices available to a researcher for collecting the project/study-specific information, one needs to be fully cognizant of the resources available for the study and the level of accuracy required. To appreciate the truth of this statement, one needs to examine the gamut of methods available to the researcher. The data sources could be either contextual and primary or historical and secondary in nature (Figure 9.1).

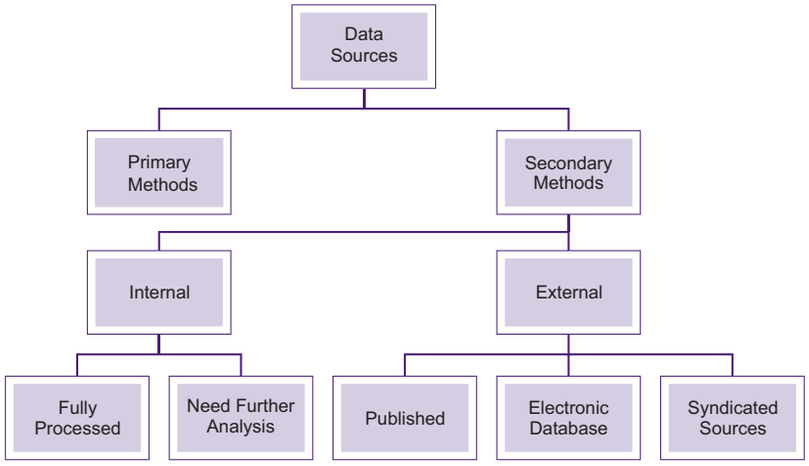


Fig. 9.1 *Sources of Data*

9.2.1 Primary and Secondary

Primary data as the name suggests is original, problem- or project-specific and collected for the specific objectives and needs spelt out by the researcher. The authenticity and relevance is reasonably high. The monetary and resource implications of this are quite high and sometimes a researcher might not have the resources or the time or both to go ahead with this method. In this case, the researcher can look at alternative sources of data which are economical and authentic enough to take the study forward. These include the second category of data sources—namely the secondary data.

Secondary data as the name implies is that information which is not topical or research- specific and has been collected and compiled by some other researcher or investigative body. The said information is recorded and published in a structured format, and thus, is quicker to access and manage. Secondly, in most instances, unless it is a data product, it is not too expensive to collect. As suggested in the opening vignette, the data to track consumer preferences is readily available and the information required is readily available as a data product or as the audit information which the researcher or the organization can procure and use it for arriving at quick decisions. In comparison to the original research-centric data, secondary data can be

economically and quickly collected by the decision maker in a short span of time. Also the information collected is contextual; what is primary and original for one researcher would essentially become secondary and historical for someone else.

Research Applications of Secondary Data

Secondary data can be used in multiple stages during the course of a business research study:

- **Problem identification and formulation stage:** Existing information on the topic under study is useful in giving a conceptual framework for the investigation. For example, if a researcher is interested in investigating the investor's perception of market risk, and he tracks investment behaviour of different quarters, alongside political, economic and social occurrences, he would be in a position to isolate the predictive variables he might wish to study.
- **Hypotheses designing:** Previous research studies done in the area as well as the industry trends and market facts could help in speculating on the expected directions of the study results. For example, the researcher in the above example might predict a positive, linear relationship between economic parameters like GDP and GNP and the choice of investment instruments and a linear negative relation between inflation rate and investment behaviour.
- **Sampling considerations:** There might be respondent related databases available to seek respondent statistics and relevant contact details. These would assist as the sampling frame for collection of primary information. For example, in the investment study, let us say the researcher wants to conduct study amongst upper income class individuals. He can then collect information on the size and spread through suitable census data.
- **Primary base:** The secondary information collected can be adequately used to design the primary data collection instruments, in order to phrase and design appropriate queries. Sometimes, the past studies done on the subject make the current study simpler, as the researcher can make use of the previously designed questionnaires. These have been standardized and validated earlier, thus the level of confidence and accuracy would be higher as compared to a new instrument.
- **Validation and authentication board:** Earlier records and studies as well as data pools can also be used to support or validate the information collected through primary sources.

Before we examine the wide range of the secondary sources available to the business researcher, it is essential that one is aware of the merits and demerits of using secondary sources.

NOTES

Benefits and Drawbacks of Secondary Data

Both benefits and drawbacks of secondary data have been discussed below:

Benefits

As we can observe, the usage of secondary data offers numerous advantages over primary data. This makes their inclusion in a research study almost mandatory. There are multiple reasons why we staunchly advocate their usage.

1. **Resource advantage:** The predominant and most important argument in support is the resource advantage. Any research or survey that is making use of secondary information will be able to save immensely in terms of both cost and time (Ghouri and Gronhaugh, 2002). VCare is a house maintenance company, located at Jaya Nagar, Bengaluru, and wants to assess the customer acceptance in the neighbouring areas. For this it wants to know: How many people reside in own houses/ apartments? How many have double income households? And how many are in the income bracket of 1 lakh+ per month?

   Thus, the latest city census data available can be accessed to arrive at these figures. Therefore, it is advocated that the investigator must first find out about the availability of probable, previously collected data, before venturing into primary data collection. The time saved in collecting information can be gainfully used for analysing and interpreting the data.

2. **Accessibility of data:** The other major advantage of secondary sources is that, once the information has been collected and compiled in a structured manner as a publication, accessing it for one's individual research purpose becomes much easier than collecting it for a singular study. Census data as the one mentioned above is generally available through a government source and is usually free of charge. However, in case VCare wants market data, in terms of size, players and volume—one might need to go to the commercial data sources which might be available for a cost, depending on the sample size and research agency repute. However, even when the data is purchased, the cost of the information would be much less as compared to collecting it on one's own.

3. **Accuracy and stability of data:** As stated in the above case, data that is collected by recognized bodies and on a large scale has the additional advantage of accuracy and reliability (Stewart and Kamins, 1993). Thus, any interpretation of primary findings or supportive logic for an implementation decision would be more precise. Moreover, since the data is collected and compiled by an outside body, it can be readily and easily accessed by other researchers as well (Denscombe, 1998).

4. **Assessment of data:** Another plus point of collecting secondary data is that the information can be used to compare and support the primary research findings of the investigators. In case the study was conducted on a representative sample of the population, the findings could be used to estimate the applicability on a larger population. Even if the findings of the earlier collected information are in contrast with the current findings, it is still useful as it might reveal the presence of certain moderator variables which might be operating in the two research conditions.

However, there is need for caution as well because in using secondary data, there might be some constraints and disadvantages as well.

Drawbacks

The drawbacks of secondary data are due to the following reasons:

1. **Applicability of data:** What one needs to remember in case of secondary data is the purpose for which the information was collected. It was unique to that study and thus cannot be an absolute fit for the current research. As a result of this, the information might not be applicable or relevant for the current objective. (Denscombe, 1998). The typical differences that emerge in such cases are with relation to the variables and the units being used to measure it. For example, market optimism or buoyancy by one researcher might be reflected by the consumer's spending in that quarter; while one might be interested in measuring buoyancy in terms of the investment in equity and growth funds.

   Another significant difference is in terms of the time period. The information that one might be using for the current research might have been collected in a different time coordinate or in a different environment. The implication of this divergence in the research base is that there might be multiple modifying variables, which might not be apparent like the socio-cultural environment, climatic effects and political factors. However, these might be responsible for skewing the direction of the findings.

2. **Accuracy of data:** While application of the data might be an issue, there is a sincere concern before one relies on the information gathered by another source—that is the level of trust one can have on the same. The concerns are three: Who, Why and How?

The first level of accuracy depends upon who was the investigator or the investigative agency. The reputation of the organization/person becomes extremely critical in establishing the truth of the findings as well as believing the inferences drawn in the quoted research. The second is the reason for collecting the data. For example, if a certain political party collects information on the potential voters and an independent market research agency collects

NOTES

information on the spread of the opinions—positive and negative—towards various political parties, one is more likely to rely on the second source. The reliability would be higher due to the reasons given below:

- Since the agency specializes in conducting opinion polls and has a vast experience as well as a respondent base, the chances of error would be minimized.
- The political party might have a hidden agenda of securing the campaign sponsorship through the survey conducted, while the independent body would be free from this bias.

Last but not the least is the data collection process of the study in terms of sample selection and sampling characteristics used to identify the respondent population. This is very important as this would be a clear indicator of the applicability of the results when extrapolating to the larger population.

Evaluation of Secondary Data—Research Authentication

Even though the data collected through other sources is valuable and critical to the research that one is undertaking, there must be certain quality checks that a researcher sometimes must undertake. On first reviewing the information, it may seem applicable and useful but on a closer examination, one might find either a mismatch with the framed research objectives or a doubt regarding the methodology or the analysis of the study. Thus, a set of evaluative measures can be employed before one decides to use it for the present study.

Methodology Check

The first evaluative criterion is the process or design used to collect the data so that in case there has been an element of skewed respondent selection or bias, one can detect it here. The verification one needs to attempt is for the following:

- **Sampling considerations:** This has to be done in terms of the defining criteria; the sampling frame; the respondent selection; response rate and the quality of data recording.
- **Methodology of data:** In terms of quality of instrument design and nature of fieldwork. This is critical as one might find that the variables measured are not as required by the current study (Jacob, 1994).
- **Analytical tools used and subsequent reporting and interpretation of results:** The problem that might occur here is that, while interpreting the findings the author might do so using his own personal judgement, which might not be based on any particular school of thought. Thus, taking the study report prima facie might be risky (Denscombe, 1998).

Further these checks also help the researcher establish whether the earlier assumptions and findings can be extrapolated on the present study.

Accuracy Check

NOTES

Dochartaigh et al. (2002) emphasize upon the significance of the source of information. The researcher must determine whether the data is accurate enough for the purpose of the present study. If the study has been conducted and the findings compiled by a reputed source, the reliability of using it as a base for further research is higher, viz., one conducted by a relative newcomer or on a small scale. In case information is from such a source, it would be advisable to collect similar data from multiple sources and then collate the findings. A related problem that might occur is when different studies/sources report contrary findings. In such a case, a short pilot study, supported by an expert opinion survey would help achieve the right perspective. This is termed as cross-check verification (Partzer, 1996).

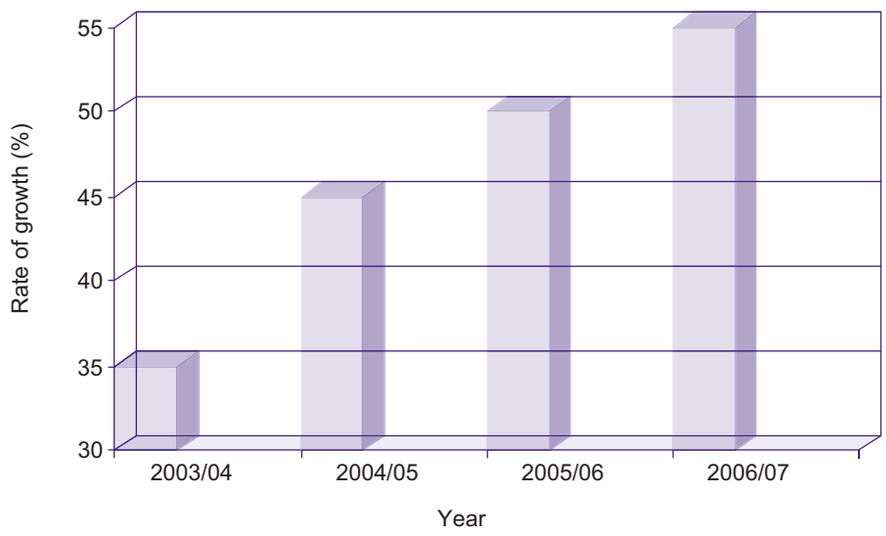
Another problem of accuracy is when the data is deliberately manipulated for the purpose of the study. This might happen in reporting of accidents and mishaps by supervisors and managers, in order to improve the safety records of the organization. Customer satisfaction surveys might decide to include only the consumer feedback data which was average to very good rather than very poor to very good thus presenting the findings demonstrating a high customer satisfaction.

The inaccuracy could also be in the presentation of the findings, i.e., the scale used might artificially enhance or play down the results. This is illustrated in the example below.

Misrepresentation of data—Bhagyshree evaluated the use of tabulated presentations in the company reports as part of her research study. Based on a sample of data collected from 53 companies' reports, she found that 29 per cent organizations made use of graphical data presentations, while 100 per cent made use of tables.

What was alarming was that 59 per cent of the figures made use of distorted graphical presentations. Either the size of the bar or the scale used was manipulated to do this. Thus, the interpretation might be misleading about the rate of change or growth. A frequently used mechanism was not to start the value axis at zero as is demonstrated in the following graph.

NOTES



Topical Check

Any information that is being used or cited in the research study needs also to be subjected to a topical check. It might happen that there is a considerable time lag between the earlier reported findings on the subject and the research being conducted now. A case in point is the census data, which is collected once in five years. However, if one is looking at the impact of variables such as age distribution and gender composition on the purchase patterns of personal care products, five years is a period where trends and fashions might have changed and presumptions or hypotheses made on the basis of such a data might be erroneous. To address these problems, a number of market research firms have started publishing syndicated sources (will be discussed later in the chapter) which are periodically updated.

Cost-benefit Analysis

Last but not the least is the financial check. Kervin (1999) states that before making use of secondary data, one needs to measure the cost of procuring the data, viz., the advantage of the information. This is applicable in the case of industry reports, market research data or readership surveys which might cost a considerable sum and the research funds might not be adequate for the purpose.

Secondary data—*Active Parenting* is a national magazine launched from Delhi. It published the results of a study conducted to find out the features parents consider most important when selecting a pre-nursery school for their child. In the order of importance, these characteristics are safety, cost, infrastructure, location, child care, teaching pedagogy, teacher attitude, and the number of admissions to reputed secondary schools. *Active Parenting* then ranked 20 schools in the NCR according to these characteristics.

NOTES

This article would be a useful source of secondary data for the pre-nursery school M Pride (MP) in conducting a market research study to identify aspects of school amenities that should be improved. However, before using the data, MP should evaluate according to several criteria.

First, the methodology used to collect the data for this survey needs to be evaluated in detail. As is the practice, *Active Parenting* has at the end of the survey indicated the methodology used in the study. A poll of 2,500 parents with children in the age group of 2–3 years was studied. The results of the survey had a 5 per cent error margin. The first thing MP needs to do is to determine whether 5 per cent is good enough to extrapolate the results to the NCR population.

Another issue that MP would need to consider is the time period of the study and the survey purpose in taking a decision on the utility of the survey findings. This survey was conducted before the Delhi government's directive on nursery admissions, which were more based on the school–residence distance. Thus, the features a parent might be looking at while evaluating a pre-nursery school might have changed. Secondly, the purpose of the survey was to acquaint the NCR parents with the options available and to build awareness on how to decide about the school for their child. Thus, the idea is to address the topical need of the hour and it is not really scientifically designed or conducted. The survey simply presents a perspective on parent opinion and is not necessarily aimed at addressing the need of the supplier—in this case the school.

The survey was conducted by CRB MR Agency for *Active Parenting* magazine. Thus, the reputation of the agency in conducting such surveys might need to be examined first. To validate the selection of the evaluative criteria, the school might look at some similar studies conducted by other MR agencies within the country or outside. Another related aspect about the methodology is the definition of the evaluation variables. For example, 'cost' in the survey was the cost inclusive of the school fees plus the transportation cost as well as the school uniform, while MP would like to evaluate 'cost' only in terms of the school fees.

However, despite all these drawbacks, the *Active Parenting* article is a cost-effective way of starting a customer expectation or a satisfaction study. For instance, it might be useful in formulating the problem's scope and objective, but, because of the article's limitations in regard to the time period, sampling, research design, and reliability, the researcher must look at some alternative studies as well as primary data collection methods.

Classification of Secondary Data

As we saw earlier in Figure 9.1, the information sources could be research-specific and primary or ex-post facto and secondary in nature. Secondary

NOTES

data can further be divided into either internal or external sources. Internal, as the name implies, is organization- or environment-specific source and includes the historical output and records available with the organization which might be the backdrop of the study. This would be directly accessible to the researcher in case he is part of the organization. However it might not be easily available to an investigator who is an outsider. The data that is independent of the organization and covers the larger industry-scape would be available through outside sources. This might be available to the researcher in the form of published material, computerized databases or data compiled by syndicated services.

Internal Sources of Data

Compilation of various kinds of information and data is mandatory for any organization that exists. Some sources of internal information are presented in Figure 9.2.

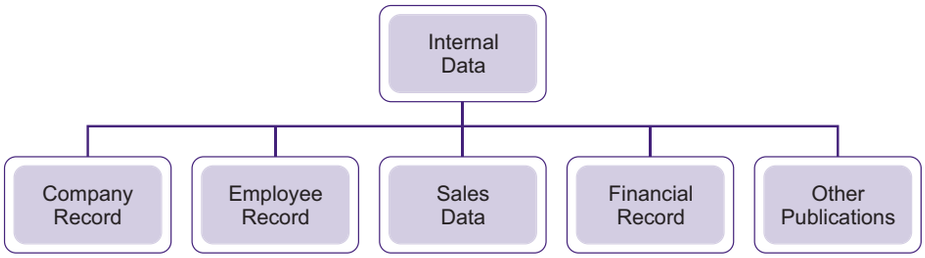


Fig. 9.2 *Internal sources of data*

The facts and information may be available (like the employee data) in a format where it can be directly used for data interpretation or analysis, however there might be certain studies for which the data from different heads would need to be processed before it can be further used. For example, in case one wants to calculate the capacity of the utilization and profitability of an organization then for this one needs the employee numbers, shift attendance, units made and sold as well as inventory figures. These have to be, then, evaluated against the financial statements.

1. **Company records:** This would entail all the data about the inception, the owners, and the mission and vision statements, infrastructure and other details including both the process and manufacturing (if any) and sales, as well as a historical timeline of the events. Policy documents, minutes of meetings and legal papers would come under this head. The access to some part of this data might be available on the public domains. However, there might be certain documents like corporate plans for the next year(s) which might not be available.
2. **Employee records:** All details regarding the employees (regular and part-time) would be part of employee records. This would include all the demographic information, as well as all the performance and discipline

NOTES

data available with reference to the individual. Performance appraisal records, satisfaction/dissatisfaction data as well as the exit interview data would also be available in the organization's annals. Sometimes, the decision maker can review the impact of certain policy changes, through performance data. Also, attrition and absenteeism data could serve as indicators for primary research required. For a service firm, employee records are more significant as people here are a part of the delivery process.

3. **Sales data:** This is an extremely valuable source and can be the most important part of the data collection process for a market research study. The data can take on different forms:
4. **Cash register receipt:** This is the simplest, most frequently recorded and available data. It would be used to reveal data under different conditions. For example, sales by product line, by major departments, by specific stores, geographical regions, by cash versus credit purchases, at specific time periods/days and the size of purchase bills.
5. **Salespersons' call records:** This is a document to be prepared and updated every day by each individual salesperson. This can reveal a wealth of information about the potential customer, classification of the customer in terms of product requirement/ company product purchase, as well as the popular products, the products that are hard to sell, information sought by the customer, customer's usage pattern and the demand analysis. The reports can also provide vital leads for a product's redesign or new product development. The data is also critical for creating job descriptions and building incentives into the system for motivating the sales force. The information needed and the presentation and negotiation required also help in designing more customized training and development initiatives.
6. **Sales invoices:** Customer who has placed an order with the company, his complete details including the size of the order, location, price by unit, terms of sale and shipment details (if any). This information set helps to forecast the annual demand for the product as well as evaluate the adequacy of sales and delivery.
7. **Financial records and sales reports:** These reveal total sales made against projected sales data, total sales by rupees and units, comparative sales performance across quarters, across regions, product categories, as well as subsequent to different sales promotion activities. Financial records in terms of sales expenses, sales revenue, sales overhead costs and profits are some of the most important output data recorded by an organization that are of critical importance as these are the dependent variables in most cases in a research for which the decision maker tries to establish the causation.

NOTES

Besides this, there are other published sources like warranty records, CRM data and customer grievance data which are extremely critical in evaluating the health of a product or an organization. There are also internal records of the published data about the organization; for example, newspaper or magazine coverage or articles published about the manufactured or a marketed product, e.g., business school ratings, harmful trans fats found in burgers and French fries as related to fast food burger chains.

There are some significant advantages of using internal data sources. First, they are readily accessible and economical to use. Secondly, they are topical and updated to the latest time period with a great amount of precision and details. However, despite these obvious advantages, most researchers do not explore the organizational archives in the first stage. A prime reason why this source is not actively sought is because it is a cumbersome task to collect information from multiple sources and then putting it together for the research study.

However, with the advent of technology, this task has been made simple and extremely fast with various data base techniques. Most organizations today maintain a data warehouse, which is essentially a computerized storehouse for the data bases that can organize large volumes of information into clusters of data based upon the user requirement. This process of organizing the data is termed as data mining. The researcher/investigator has the provision through this technique to create multi-dimensional analysis and reports based upon a unidimensional original data set. Various software programmes and languages are used to detect patterns and trends from the data like the neural networks, tree models, estimation, market basket analysis, genetic algorithms, clustering, classification, etc. In fact these techniques make the prediction of the outcome so effective and involving a minimal error that a lot of firms are actively relying on data mining of the internal data sources, viz., the external data or primary data for implementing planned strategies.

External Data Sources

As stated earlier, information that is collected and compiled by an outside source that is external to the organization is referred to as external source of data. Included under this head (Figure 9.1) are published sources, computer-based information sources and syndicated sources. Each of these would be discussed separately in this section.

Published data

The most frequently used and most easily available data information that is compiled by using public or private sources. There could be a plethora of information available on the same topic from varied sources. For the sake of the avid researcher who would like to explore these options, listed below are some potential information sources.

NOTES

There could be two kinds of published data—one that is from the official and government sources—this could include census data, policy documents and historical archives; the other kind of data is that which has been prepared by individuals or private agencies or organizations. This could be in the form of books, periodicals, industry data such as directories and guides.

1. **Government sources:** The Indian government publishes a lot of documents that are readily available and are extremely useful for the purpose of providing background data. This could be available on public domains or might be retrieved by special permission. The publications are usually available, for example the population or census data and other publications.
   - **Census data:** Considering the size of the Indian subcontinent, one needs to understand the magnitude of the data available and the intensity of effort required to record information from all parts of the country. Recently, the Census 2011 has been carried out and the quality of census data promises to be very high and the data has been collected in a much more detailed format.
   - **Other government publications:** In addition to the census, the Indian government collects and publishes a great deal of statistical data. The Planning Commission of India has in its archives all the details on economic planning and outcomes of the country. Other sources are budget and legislative documents and other economic surveys done related to the trade and culture of the country. The data could be further available at the micro level, that is the state level as well. Today, with the advent of technology, most of this is available in computerized form. Listed in Table 9.1 is an illustration of some of the sources. One may find that the list is neither complete nor exhaustive. The objective is to give the researcher a flavour of the kind of recorded information available to him for his study. Another point to be noted is that while we have listed the Indian sources, similar data is available for most countries.

Table 9.1 *Secondary Data—Government Publications*

	Sub-type	Sources	Data	Uses
1.	Census data conducted every ten years throughout the country	Registrar General of India conducting census survey http://censusindia.gov.in/	Size of the population and its distribution by age, sex, occupation and income levels. 2011 census took many more variables to get a better picture of the population	Population information is significant as the forecasts of purchase, estimates of growth and development, as well as policy decisions can be made on this basis

NOTES

	Sub-type	Sources	Data	Uses
2.	Statistical Abstract India – annually	CSO (Central Statistical Organization) for the past 5 years http://www.mospi.gov.in/cso_test1.htm	Education, health, residential information at the state level is part of this document	Making demand, estimations and a state-level assessment of government support and policy changes can be made
3.	White paper on national income	CSO http://www.mospi.gov.in/cso_test1.htm	Estimates of national income, savings and consumption	Significant indication of the financial trends; investment forecasts and monetary policy formulation
4.	Annual Survey of Industries – all industries	CSO – no. of units, persons employed, capital output ratio, turnover, etc. http://www.mospi.gov.in/cso_test1.htm		Information on existing units gives perspective on the Industrial development and helps in creating the employee profile
5.	Monthly survey of selected industries	CSO http://www.mospi.gov.in/cso_test1.htm	Production statistics in detail	Demand–supply estimations
6.	Foreign Trade of India Monthly Statistics	Director General of Commercial Intelligence http://www.dgciskol.nic.in/	Exports and imports countrywise and productwise	Forecast, manufacturing and trade estimations
7.	Wholesale price index – weekly all-India Consumer Price Index	Ministry of Commerce and Industry http://india.gov.in/sectors/commerce/ministry_commerce.php	Reporting of prices of products like food articles, foodgrains, minerals, fuel, power, lights, lubricants, textiles, chemicals, metal, machinery and transport	Establishing price bands of product categories; pricing estimations for new products; determining consumer spend
8.	Economic Survey – annual publication	Dept. of Economic Affairs, Ministry of Finance, patterns, currency and finance http://finmin.nic.in/the_ministry/dept_eco_affairs/	Descriptive reporting of the current economic status	Estimations of the future and evaluation of policy decisions and extraneous factors in that period
9.	National Sample Survey (NSS)	Ministry of Planning http://www.planningcommission.gov.in/	Social, economic, demographic, industrial and agricultural statistics	Significant for making policy decisions as well as studying sociological patterns

2. **Other data sources:** This source is the most voluminous and most frequently used, in every research study. The information could be in the form of books, periodicals, journals, newspapers, magazines, reports,

NOTES

and trade literature. The data could also be available as compilations in the form of guides, directories and indices.

- **Books and periodicals:** Books and periodicals are the simplest, easily accessible and user friendly form of documented material. The volumes could carry information ranging from constructs, technical details and cultural data to just a collection of views on the topic of interest to the researcher.
- **Guides:** These are an instructive source of standard or recurring information. A guide may subsequently lead into identifying other important sources of directories, trade associations and trade publications. In fact it is advisable to begin a study by exploring such guides.
- **Directories and indices:** Directories are useful as they may again lead to a source or a pool of specific information. Indices, on the other hand, serve as a collection of the location of information on a particular topic in several different publications.
- **Standard non-governmental statistical data:** Published statistical data are of great interest to researchers. Graphic and statistical analyses can be performed on these data to draw important insights. There are renowned private agencies which periodically compile and publish this kind of data and they are considered extremely significant in their contribution to understanding the market. Important sources of non-governmental statistical data include Standard and Poor's Statistical Service, Moody's Industrial manual and data from agencies such as NASSCOM & MAIT (IT Industry); SIAM (automobile industry); CETMA, IEEMA (electronics) and IPPAI (power). Reports and documents available from renowned bodies like the World Bank, United Nations and World Trade Organization are also valuable sources of secondary information. Some non-government data sources are presented in Table 9.2.

Table 9.2 *Secondary data—Non-government publications*

	Sub-type	Sources	Data	Uses
1.	Company Working Results – Stock Exchange Directory	Bombay Stock Exchange http://www.bseindia.com/	A complete database of the companies registered with the stock exchange and comprehensive details about stock policies and current share prices	Significant in determining the financial health of various sectors as well as assessment of corporate funding and predictions of outcomes

NOTES

	Sub-type	Sources	Data	Uses
2.	Status reports by various commodity boards	The commodity board or the industry associations like Jute Board, Cotton Industry, Sugar Association, Pulses Board, Metal Board, Chemicals, Spices, Fertilizers, Coir, Pesticides, Rubber, Handicrafts, Plantation Boards, etc.	Detailed information on current assets – in terms of units, current production figures and market condition	These are useful for individual sectors in working out their plans as well as evaluating causes of success or failure
3.	Industry associations on problems faced by private sector, etc.	FICCI, ASSOCHAM, AIMA, Association of Chartered Accountants and Financial Analysts, Indo-American Chamber of Commerce, etc. http://www.ficci.com/ http://www.assocham.org/ http://www.aima-ind.org/ www.iaccindia.com/	Cases/ comprehensive reports by the supplier or user or any other section associated with the sector	Cognizance of the gaps and problems in the effective functioning of the organization; trouble shooting
4.	Export-related data – commodity-wise	Leather Exports Promotion Council, Apparel Export Promotion Council, Handicrafts, Spices, Tea, Exim Bank, http://www.leatherindia.org/ http://www.aepcindia.com/	Product- and country-wise data on the export figures as well as information on existing policies related to the sector	To estimate the demand; gauge opportunities for trade and impetus required in terms of manufacturing and policy changes
5.	Retail Store Audit on pharmaceutical, veterinary, consumer products	ORG (Operations Research Group); Monthly reports on urban sector; Quarterly reports on rural sector	The touch point for this data is retailer, who provides the figures related to product sales; the data is very comprehensive and covers most brands. The data is region-specific and covers both inventory and goods sold	Market analysis and market structure mapping with estimations of market share of leading brands. The audit can also be used to study consumption trends at different time periods or subsequent to sales promotion or other activities

NOTES

	Sub-type	Sources	Data	Uses
6.	National Readership Survey (NRS)	IMRB survey of reading behaviour for different segments as well as different products http://www.imrbint.com/	Today these surveys are done by various bodies with different sample bases. Today the survey base has become younger, with the age of the reader lowered to 12+	Media planning and measuring exposure as well as reach for product categories
7.	Thompson Indices: Urban market index, rural market index	Hindustan Thompson Associates	All towns with population of more than one lakh are covered and information of demographic and socio-economic variables are given for each city with Mumbai as base. The rural index similarly covers about 400 districts with socio-economic indicators like value of agriculture output, etc.	The inclinations to purchase consumer products are directly related to socio-economic development of communities in general. The indices provide barometers to measure such potentials for each city and has implications for the researcher in terms of data collection sources

However, no matter how vast and differentiated is the published data source available to the researcher, hunting from huge volumes is truly a herculean task and can be extremely tedious. With the advent of computer technology, today, most published information is also available in the form of computerized databases.

9.2.2 Qualitative and Quantitative

To comprehend the distinction between the two approaches, one needs to appreciate the contribution of each to the research process.

Research objective

Qualitative research: It can be used to explore, describe or understand the reasons for a certain phenomena. For example, to understand what a low-cost car means to an Indian consumer, this kind of investigation would be required.

Quantitative research: When the data to be studied needs to be quantified and subjected to a suitable analysis in order to generalize the findings to the population at large or to be able to quantify and explain and predict the occurrence of a certain phenomena. For example, to measure the purchase

intentions for Nano as a function of the demographic variables of income, family size and distance travelled, one would need to use quantitative methods.

Research design

NOTES

Qualitative research: The design is exploratory or descriptive, loosely structured and open to interpretation and presumptions.

Quantitative research: The design is structured and has a measurable set of variables with a presumption about testing them.

Sampling plan

Qualitative research: Only a small sample is manageable as the information required needs to be extracted by a flexible and sometimes lengthy procedure.

Quantitative research: Large representative samples can be measured and the data collected can be based upon a shorter time span with a larger number. Chances of error in extrapolating it to a larger population are less and measurable.

Data collection

Qualitative research: The data collection is in-depth and collected through a more interactive and unstructured approach. Data collected includes both the verbal and non-verbal responses. Methodology requires a well-trained investigator.

Quantitative research: The data collected is formatted and structured. The nature of interrogation is more of stimulus-response type. The data collected is usually verbal and well- articulated. Interrogation does not need extensive training on the part of the investigator.

Data analysis

Qualitative research: Interpretation of data is textual and usually non-statistical.

Quantitative research: Interpretation of data entails various levels of statistical testing.

Research deliverables

Qualitative research: The initial and ultimate objective is to explain the findings from more structured sources.

Quantitative research: The findings must be conclusive and demonstrate clear indications of the decisive action and generalizations.

It is essential to remember that even though the information obtained is rich and extensive, it is diagnostic and not evaluative in nature, thus, should not be used for generalizations on to larger respondent groups. Secondly, because of the nature of the conduction, they always cover smaller sample groups or

individuals. Thus, they are indicative rather than predictive in nature. And lastly, they indicate the direction of respondent sentiments and should not be mistaken for the strength of the reactions. Thus, what is advocated is that the two approaches—qualitative and quantitative—are not to be treated as the extreme ends of a theoretical continuum. A business researcher should take them as complementary and supportive in order to get measurable as well as humanistic inputs for taking informed decisions.

Check Your Progress

1. Differentiate between primary and secondary data.
2. What are company records?
3. What are the advantages of using internal data sources?
4. What are the two kinds of published data?

9.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Primary data as the name suggests is original, problem- or project-specific and collected for the specific objectives and needs spelt out by the researcher. The authenticity and relevance is reasonably high. Secondary data as the name implies is that information which is not topical or research- specific and has been collected and compiled by some other researcher or investigative body. The said information is recorded and published in a structured format, and thus, is quicker to access and manage.
2. Company records would entail all the data about the inception, the owners, a the mission and vision statements, infrastructure and other details including both the process and manufacturing (if any) and sales, as well as a historical timeline of the events. Policy documents, minutes of meetings and legal papers would come under this head.
3. There are some significant advantages of using internal data sources. First, they are readily accessible and economical to use. Secondly, they are topical and updated to the latest time period with a great amount of precision and details.
4. .There could be two kinds of published data—one that is from the official and government sources—this could include census data, policy documents and historical archives; the other kind of data is that which has been prepared by individuals or private agencies or organizations. This could be in the form of books, periodicals, industry data such as directories and guides.

NOTES

9.4 SUMMARY

- To understand the multitude of choices available to a researcher for collecting the project/study-specific information, one needs to be fully cognizant of the resources available for the study and the level of accuracy required. To appreciate the truth of this statement, one needs to examine the gamut of methods available to the researcher.
- Primary data as the name suggests is original, problem- or project specific and collected for the specific objectives and needs spelt out by the researcher.
- Secondary data as the name implies is that information which is not topical or research- specific and has been collected and compiled by some other researcher or investigative body.
- Any information that is being used or cited in the research study needs also to be subjected to a topical check. It might happen that there is a considerable time lag between the earlier reported findings on the subject and the research being conducted now.
- Secondary data can further be divided into either internal or external sources. Internal, as the name implies, is organization- or environment-specific source and includes the historical output and records available with the organization which might be the backdrop of the study.
- There are some significant advantages of using internal data sources. First, they are readily accessible and economical to use. Secondly, they are topical and updated to the latest time period with a great amount of precision and details.
- There could be two kinds of published data—one that is from the official and government sources—this could include census data, policy documents and historical archives; the other kind of data is that which has been prepared by individuals or private agencies or organizations. This could be in the form of books, periodicals, industry data such as directories and guides.
- Books and periodicals are the simplest, easily accessible and user friendly form of documented material.
- The volumes could carry information ranging from constructs, technical details and cultural data to just a collection of views on the topic of interest to the researcher.

9.5 KEY WORDS

- **Primary data**: It is information collected through original or first-hand research. For example, surveys and focus group discussions.
- **Secondary data**: It is information which has been collected in the past by someone else. For example, researching the internet, newspaper articles and company reports.

NOTES

9.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. What are the research applications of secondary data?
2. State some of the drawbacks of secondary data.
3. State some of the government sources of data.
4. Differentiate between qualitative and quantitative data on the basis of research objectives.

Long Answer Questions

1. How is secondary data evaluated? Discuss.
2. Describe the classification of secondary data.
3. Distinguish between qualitative from quantitative data methods.
4. Discuss the different benefits of secondary data.

9.7 FURTHER READINGS

Chandan, J.S., Jagjit Singh and K.K. Khanna. 1995. *Business Statistics*, 2nd edition. New Delhi: Vikas Publishing House Pvt. Ltd.

Croxton, Frederick E., and Dudley J. Cowden. 1943. *Applied General Statistics*. New York: Prentice Hall.

Gupta, S.P. 2006. *Statistical Methods*. New Delhi: S. Chand & Company.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods*, 23rd edition. New Delhi: Vikas Publishing House Pvt.Ltd..

Levin, Richard I. and David S. Rubin. 1998. *Statistics for Management.* New Jersey: Prentice Hall.

UNIT 10 TABULATION OF DATA

NOTES

Structure

10.0 INTRODUCTION

This unit will introduce you to tabulation, its concepts and objectives. It refers to the tabulation of data into appropriate tables. Tables are generally one of the following types: single-column or single-row table, multiple-column or multiple-row table. The components of a table include table number, title of the table, headnotes, footnotes and sources. A statistical table, which is an orderly and systematic presentation of numerical data in columns and rows, also has the same components.

10.1 OBJECTIVES

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

- Explain the concept of tabular presentation and the types of tables
- Discuss the components of a table
- Analyse the framing of tables
- Describe the concept of statistical tables
- Differentiate between classification and tabulation

10.2 TABULATION: MEANING AND OBJECTIVES

Tabular presentation means tabulating the data in the form of appropriate tables. A table is a statistical table, containing data arranged into convenient

number of *rows* and/or *columns.* The numbers of rows or columns in which data may be classified (or distributed) help bring out the broad data features to the fore to be easily seen at a glance.

The basic function of a table is to simplify data and to present them in a manner that facilitates comparison. Simplifying data means that the information desired becomes easy to locate. Comparison involves bringing all related data together at one place such that a relational picture can be conveniently and efficiently drawn.

10.2.1 Types of Tables

Statistical tables can be laid in various ways. The form of a table must suit the data at hand and be convenient to achieve the objective(s) in mind. Generally, a table is of the following types:

(i) **Single-column or single-row tables:** Such tables are the simplest to construct. The data in these tables are arranged in a single row or a single column according to time, place, region of space, or an attribute of interest. The table is vertically laid when the data are arranged in a single column, and horizontally laid when the data are arranged in a single row.

In fact, laying the table either vertically or horizontally means the same thing. How the available space allows laying the table is perhaps the only important consideration that goes into deciding it. A horizontally laid table obviously consumes lesser space. Otherwise, these two ways of tabulating data constitute essentially a single type of table.

(ii) **Multiple-column and multiple-row tables:** As against single-column and single-row tables, the given data on a variable may also be arranged in multiple columns and rows. The data break-up and the kind of relational comparative picture intended determine the number of columns and rows required. If the number of *rows* is represented by *r* and of *columns* by *c,* such tables are known as *'r by c'* tables.

The intersection of each row with each column makes a *cell.* This means that any *'r by c'* table consists of *r* x *c* cells. A table so constructed is known as a cross-classification table, with format looking as in Table 10.1. It shows the following:

(a) There are three columns and four rows with 4 x 3 = 12 cells comprising the body of the table, each containing a figure.

(b) While columns describe one characteristic of the data, rows describe the other.

(c) Either columns or rows may represent time, place, region of space, or some other attribute of the data.

NOTES

Table 10.1
. ..... Title ..... .
... . Head Note .. ..

Stub	Box Head/Caption		
	Col. Head	Col. Head	Col.Head
Sub/Row Head	Cell	Cell	Cell
Stub/Row Head	Cell	Cell	Cell
Stub/Row Head	Cell	Cell	Cell
Stub/Row Head	Cell	Cell	Cell

Footnote(s): .....

Source(s): .....

(iii) **Reference vs summary tables:** Yet another classification of data may yield a statistical table known as a reference table or a summary table. Here the criterion of data classification is the quantum of information that a table contains.

Reference tables are the ones which present extensive information on any subject. For all practical purposes, these tables are the repository of basic data and work almost as data inventory. They are primarily meant for referencing, and serve as source material for summary tables. Accordingly, these tables are also known as basic or source tables.

For example, all data published in the *Annual Survey of Industries* are in the form of reference tables, which offer all relevant data on industries in detail.

Summary tables, on the contrary, provide only summarized data on one or more related aspects on a given subject. These are drawn from reference tables and are usually displayed in the course of running text. Usually, they are meant to be used as necessary support to inferences drawn in the text of the report. Accordingly, these are also known as text or analytical tables. Their basic function is to highlight comparisons and reveal possible relationships.

10.2.2 Presentation of a Table

The presentation of a table involves of few components. These components are functional parts that constitute the structure of the table. Almost invariably, there are eight (8) components of a statistical table. Each of these may be understood with reference to the typical format of Table 10.1.

- **Table number:** A table must be appropriately numbered, to allow making references and citing results. It makes sense to relate the numbering of the tables with serial number of the chapter. Generally, the digit occurring before the dot (.) indicates the chapter where the

table appears, and the digit appearing after the dot (.) tells the serial number of the table in the given chapter.

- **Title of the table:** Each table has to be given a suitable title. The title should be so framed and stated that it briefly tells all about the data tabulated. The title should be very short, but as complete and speaking as possible. It must also convey the subject, time, and place the data contained in the table refer to.
- **Headnotes:** A head note figures immediately below the title. It either offers some additional information about the title and/or qualifies the data presented in the table. *For example,* if the data are expressed in *thousand dollars,* it is mentioned as a head note. Importantly, a head note is a qualifier usually provided in brackets.
- **Stub and stub-head:** Stub refers to the main heading of rows, while the stub-heads/entries occur as *row-headings* against which data entries are made. The stub of a table consists of as many *stub-heads* as the number of rows.
- **Box head and sub-heads:** The box head describes the data provided in various columns. It is also called *caption,* being the title under which the column heads are provided. Since sub-heads specify the data occurring under various columns, these are parts of the *caption* and are provided thereunder.
- **Body of the table:** The body of the table consists of a number of *cells,* each containing a figure called *cell entry.* The body contains *r* x *c* cells, and thus equal number of cell entries. Each cell occurs at the intersection of a column and a row.
- **Footnote(s):** A footnote provides additional information, if any, about the functional parts of a table. Generally, it is by way of some clarification(s) that may be necessary about an entry made in the table. Or it may be by way of a qualification to the data presented in rows and/or columns.
- **Source(s):** A source mentions where the data presented in the table have come from. This is an important component of the table, since the source enables the reader to check and re-check the data from where these may have been borrowed. It may also help draw, if relevant and necessary, more information from the source. The source must indicate all information about itself, such as publication, place and year of publication, and page(s) and table(s) where the concerned data appear.

How to Frame Tables

There are no hard and fast rules governing how to frame a statistical table. It all depends on the kind of data available and the objective(s) one wishes to achieve. Experience of having been engaged in research is perhaps the

NOTES

only important factor that plays a decisive role in framing a table of high interpretative value.

There are, however, a few catch-points that do help construct a useful table.

- Where any two sets of data are to be compared, these should preferably be presented in columns. Column presentation of related data offers a more vivid comparative picture than when the same data are laid in rows.
- Where some figures provided in any column or a row are required to be brought into focus the same may be made bolder. *For example,* totals and sub-totals deserve more attention in drawing a comparison, or otherwise. This facilitates the desired data being easily noticed and/or distinguished. As a measure of data refining, it improves the value of data presentation and adds to the fineness of the table.
- Where availability of space is a constraint in deciding the size of a table, it should be so designed that the available space accommodates the table with all the information it is supposed to contain. Space limitation may at times be serious enough to require abridging the table either horizontally or vertically. Abridging must, however, ensure that the basic information needed for analysis, drawing inferences, and/or establishing relationship(s) is not lost in the process of reducing size. Otherwise, a table will suffer a serious handicap in achieving the objective(s) in mind.
- Where reducing the space requirement of a table is unavoidable, a useful way of doing so is to appropriately round off the figures, following the basic rules of rounding. Long figures expressed in many digits can be easily made short by expressing them in, say, thousands or millions, as may be necessary. Similarly, decimal figures can also be suitably adjusted up to the desired number of decimal points.

While the points made above do matter in constructing a meaningful statistical table, the basic ground rule is no different from applying common sense and imagination, keeping the use requirements in mind. Any table that we may intend to construct and lay should generally be an intelligent display of data so as to be conveniently read and understood.

A Contingency Table

A contingency table is an important form of presenting observed data. It is amenable to the application of a number of useful statistical tools of data analysis. It follows largely the same format as that of Table 10.1. Running

NOTES

into *r* number of rows and c number of columns, there are *'r* x c' cell entries which make the body structure of the table.

Consider for example, Table 10.2, which gives the distribution of 2,000 collegiate students according to sex and economic status. As a contingency table, it deviates from a normal multi-column and multi-row format in Table 10.2 as under:

Table 10.2 *Classification of 2000 Collegiate Students According to Sex and Economic Status (A 2 × 3 Contingency Table)*

Sex	Economic Status			
	High Income Means	Average Income Means	Low Income Means	Row Totals
Boys	120	700	380	1200
Girls	80	500	220	800
Column Totals	200	1200	600	2000

- A contingency table is a cross-section presentation of observed data in terms of any two attributes. Here, one is sex and the other economic status. Importantly, the data presented in any such table are the observed frequency, and not the continuous quantitative data on a variable.
- The data appearing as cell entries in a contingency table are essentially qualitative count data. To be more specific, the cell entries are observed frequencies/counts of an item or the outcome of an event possessing or not possessing a certain attribute.
- In addition to the cell entries being determined as *'r* x *c',* the last column in a contingency table provides row totals and the last row gives column totals. At the intersection of the last column (for row totals) and the last row (for column totals) lies a cell containing the total number of frequencies, or the number of subjects or objects/ items observed in terms of the two attributes of interest. The row totals and column totals are known as *marginal frequencies.*

A look at Table 10.2 shows that the data provided in the cells are count data. The row totals and column totals both add to 2000, the total number of students observed. The last column presents the row totals and the last row the column totals. All this is unlike a normal cross-classification table, where the data are the measurements of a continuous quantitative variable.

The cell frequencies in a contingency table are amenable to meaningful interpretations. *For example,* the first cell frequency (that is, 120) means that out of all the 2000 collegiate there are 120 boys who have high-income means.

Similarly, among 200 collegiate out of 2000 who have high income means, 120 are boys and 80 girls. And, so on. An important point that must weigh heavy in the construction of a contingency table is that the two classifying attributes are clearly and objectively defined. This helps

NOTES

stating the various column heads and row heads in unambiguous terms as to their meaning and coverage. Any ambiguity in defining the attributes and, consequently, the column and row heads seriously erodes an objective classification of the observed data. It also does not allow the cell frequencies to offer precise and meaningful interpretations.

Statistical Tables

A statistical table is an orderly and systematic presentation of numerical data in columns and rows. Columns are vertical arrangements; rows are horizontal. The main objective of a statistical table is to so arrange the physical presentation of numerical facts that the attention of the reader is automatically directed to the relevant information. Some of the main advantages of tabular presentation over descriptive statements are as follows:

- Tabulated data can be easily understood than facts stated in the form of descriptions.
- They leave a lasting impression.
- They facilitate quick comparison.
- Statistical tables make easier the summation of items and detection of errors and omissions.
- When data are tabulated all unnecessary details and repetitions are avoided.
- A tabular arrangements makes it unnecessary to repeat explanations, phrases and headings.

Parts of Tables

The following parts must be present in all tables:

- Title
- Caption
- Stubs
- Body

There are, however, other parts whose presence depends upon the specific purpose. They are Headnote (or prefatory note), footnote and source note.

- **Title:** A complete title explains in brief and concise language (a) what the data are, (b) where the data are, (c) time period of data and (d) how the data are classified.
- **Captions:** The title of the columns are given in captions. In case there is a sub-division of any column, there would be sub-caption headings also.

NOTES

- **Stubs:** The titles of the rows are called stubs. The box over the stub on the left of the table gives description of the stub contents, and each stub labels the data found in its row of the table.
- **Body:** The body of the table contains the numerical information.
- **Headnote** (or prefatory note) It is a statement, given below the title, which clarifies the contents of the table.
- **Footnote:** It is a statement which clarifies some specific items given in the table or explains the omission thereof. Thus, if we look into a table, giving yearly figures of wheat production in India, the sudden fall in the figure for 1947 relate to India after partition.
- **Source:** The source from where the data contained in the table has been obtained should be stated. This would permit the reader to check the figures and gather, if necessary, additional information.

Table 10.3 *Title (Description of Units and Year, Place etc) Headnote*

(Stub box) (D)	(A) Caption		(B) Caption	
	(1)	(2)	(3)	(4)
Sub X				
Y	B	O	D	Y
Z				
Total				

Notes: Any definition.

Any explanation.

Source from which derived.

Classification of Tables

Tables may be classified according to the number of characteristics used for tabulation. A simple or a one-way table use only one characteristic against which the frequency distributions given, as in Table 10.4 where the characteristic used is the age of student.

Table 10.4 *Age Wise Distribution of the Students of a College*

Age in Year	Students
16—17	—
17–18	—
18—19	—

In a two-way table, on the other hand, two characteristics are used. In this case, one characteristic is taken as column headings, and the other as row stubs. Example of a two-way table showing a two-way frequency distribution is shown in Table 10.5.

NOTES

Table 10.5 *Age and Sex Wise Distribution of the Students of a College*

Age in years	Students		Total
	Male	Female	
16–17	—	—	—
17–18	—	—	—
18 and on	—	—	—

When it is desired to represent three or more characteristics in a single table, such a table is called higher order table. Thus, if it is desired to represent the 'age', 'sex' and 'course', of the students, the table would take the form as shown on page 70 and would be called a higher order table.

Table 10.6 *Table Showing Distribution of the Students of a College According to 'Age', 'Sex', and 'Course'*

	Course					
Age in Years	Arts		Science		Commerce	
Total	Male	Female	Male	Female	Male	Female
16–17						
17–18						
18 and Over						
Total						

Illustration 10.1: Draft a form of tabulation to show:

(a) Sex,

(b) Three ranks–supervisors, assistants and clerks,

(c) Years–1970 and 1979

(d) Age group–18 years and under, over 18 but less than 55 years, over 55 years.

Solution: In the previous question, we have to prepare a table to show four characteristics, i.e., sex., three ranks of the employees, as given, for two different years and the data is to be divided according to age groups already given here. We can prepare a blank table to incorporate all these characteristics (Table 10.7).

Table 10.7 *Table Showing the Division of Three Ranks of Employees According to Sex and Age Group for 1976 and 1979*

		1976				1979			
	Age Group	**Supervisors**	**Assistants**	**Clerks**	**Total**	**Supervisors**	**Assistants**	**Clerks**	**Total**
Males	0–18								
	18–55								
	55 and above								
	Total								
Females	0–18								
	18–55								
	55 and above								
	Total								

Illustration 10.2: The city of Timbuktu was divided into three areas: the administrative district, other urban districts and rural districts. A survey of housing conditions was carried out and the following information was gathered:

There were 6,77,100 buildings of which 1,76,100 were in rural districts. Of the buildings in other urban districts 4,06,400 were inhabited and 4,500 were under construction in the administrative district 4,000 buildings were inhabited and 500 were under construction of the total of 61,600.

The total buildings in the city that are under construction are 6,200 and those uninhabited are 44,900.

Tabulate the above information so as to give the maximum possible information. How many buildings are under construction in rural areas?

Solution

Table 10.8 *Distribution of Building in the Three Districts of Timbuktu According to Inhabitation*

(IN HUNDREDS)

DISTRICT	**INHABITED**	**UNIHABITED**	**UNDER CONSTRUCTION**	**TOTAL**
ADMINISTRATIVE	571	40	5	616
OTHER URBAN	4064	285	45	4394
RURAL	1625	124	12	1761
TOTAL	6260	449	62	6771

The table clearly shows that there are 1,200 buildings under construction in rural areas.

Illustration 10.3: An investigation conducted by the education department in a public library revealed the following facts. You are required to tabulate the information as neatly and clearly as you can.

'In 1960, the total number of readers was 46,000 and they borrowed some 16,000 volumes. In 1965, the number of books borrowed increased by 4,000 and the borrowers by 50 per cent.'

NOTES

NOTES

The classification was on the basis of three sections: Literatures, Fiction and Illustrated News. There were 10,000 and 30,000 readers in the section Literature and Fiction, respectively, in the year 1960–Illustrated news and Fiction, respectively. Marked changes were seen in 1965. There were 7,000 and 42,000 readers in the Literature and Fiction section respectively. So also 4,000 and 13,000 books were lent in the section Illustrated News and Fiction respectively.

Solution:

Table 10.9 *Showing the Changes in the Number of Readers and Type of Books in the Year 1975 as Compared to 1970.*

	1970		*1975*			
Types of books	*Number of*	*Number of books readers*	*Number of borrowed*	*Number of books readers*	*Changes in 1975 over 1970*	*borrowed*
Fiction	30,000	10,000	42,000	13,000	+12,000	+3000
Literature	10,000	4,000	7,000	3,000	–3,000	–1,000
Illustrated news	6,000	2,000	20,000	4,000	+18,000	+2,000
Total	46,000	16,000	69,000	20,000	27,000	4,000

Illustration 10.4: Prepare a two-way frequency table and marginal frequency tables for 25 values of the two variables x and y given below. Take class interval of x as 10–20, 20–30, etc., and that of y as 100–200, 200–300 etc.

x	12	24	33	22	44	37	26	36	
y	140	256	360	470	470	380	280	315	
x	55	48	27	57	21	51	27	42	
y	420	390	440	390	590	250	550	360	
c	43	52	57	44	48	48	52	41	69
y	570	290	416	280	452	370	312	330	590

Solution:

Table 10.10 *Bivariate Frequency Table*

Y \ X	10–20	20–30	30–40	40–50	50–60	60–70	Total
100–300	1	–	–	–	–	–	1
200–300	–	2	–	–	2	–	4
300–400	–	–	3	5	2	–	10
400–500	–	2	–	2	2	–	6
500–600	–	2	–	1	–	1	4
Total	1	6	3	8	6	1	25

NOTES

Table 10.11 *Marginal Distribution of X*

x	f
10 – 20	1
20 – 30	6
30 – 40	3
40 – 50	8
50 – 60	6
60 – 70	1
Total	25

Table 10.12 *Marginal Distribution of Y*

y	f
100 – 200	1
200 – 300	4
300 – 400	10
400 – 500	6
500 – 600	4
Total	25

Illustration 10.5: In a trip organized by a college, there were 80 persons, each of who paid ₹ 15.50 on an overage. There were 60 students, each of who paid ₹ 16. Members of teaching staff were charged at a higher rate. The number of servants (all males) was six and they were not charged anything. The number of ladies was 20 per cent of the total and there was only one lady staff member. Tabulate this information.

Solution:

Total contribution = 80 × 15.50 = ₹ 1240.00

Table 10.13 *Showing Participants, Sex and Class wise*

Class	*Sex*		*Totals*	*Contribution*	*Contribution*
	Males	*Females*			
Students	45	15	60	16	960
Teaching Staff	13	1	14	20	280
Servants	6	–	6	–	–
Totals	64	16	80	15.50	1240

Illustration 10.6: Prepare a bivariate frequency distribution for the following data:

Marks in Law	10	11	10	11	11	14	12	12	13	10	13
Marks in Statistics:	20	21	22	21	23	23	22	21	24	23	24

Marks in Law:	12	11	12	10	14	12	13	10	14
Marks in Statistics:	23	22	23	22	22	20	24	23	24

NOTES

Solution:

Marks in Statistics / Law	20	21	22	23	24	Total
10	1	–	2	2	–	5
11	–	2	1	1	–	4
12	1	1	1	2	–	5
13	–	–	–	–	3	3
14	–	–	1	1	1	3
Total	2	3	5	6	4	20

Check Your Progress

1. What is the basic function of a table?
2. Which tables are known as '*r by c'* tables?
3. What is the function of a headnote in a table?

10.3 DIFFERENCE BETWEEN CLASSIFICATION AND TABULATION

Classification and tabulation, are both methods of summarizing data in statistics. It is used to draw further analysis of data or to draw inferences from the given data. Here below, are discussed the two methods of summarizing the data and the difference between classification and tabulation of data.

10.3.1 Classification of Data

Classification in statistics refers to the process of separation of data into various groups or classes with the help of properties in the data set. For example, the interests of particular class or group can be separated on the basis of gender. In this classification, the raw data condenses into suitable forms for statistical analysis and removes complex data patterns and highlights the core representatives of the raw data. Post classification, the data can be put to comparison or inferences. Classified data at some means can also provide relationships or correlative data patterns.

Data when it is raw, is classified using four key characteristics geographical, chronological, qualitative and quantitative properties. Considering that a data set is gathered for the analysis of the consumption of petrol per day around the world. The consumption of petrol can be classified on the basis of countries and types of vehicles. Here, geographical factors and vehicle types are the merits for classification. A further classification

as chronological, can include older vehicles which have a higher rate of consumption. The maintenance and serviceability of the vehicles can act as the qualitative base of classification and the gross average claimed by the manufacturer can act as the quantitative base for classification of the consumption.

10.3.2 Tabulation of Data

Tabulation in statistics is a method of summarising data by using a systematic arrangement of data into rows and columns. Tabulation is carried out as to investigate, compare, identify errors or omissions in data, to study a prevailing trend, to simplify the known raw data and to use the space economically and use it as future reference.

The following are the components of a statistical table:

Component	Description
Title	It is a brief explanation of the contents of the table
Table Number	It is a number assigned to a table for easy identification
Date	Date of the creation of the table should be indicated
Row Designations	Each row of the table is given a brief name, usually provided in the first column. Such a name is known as a "stub", and the column is known as the "stub column"
Column Headings	Each column is given a heading to explain the nature of the figures, these are known as "captions" or "headings".
Body of the table	Data is entered into the main body and should be created for easy identification of each data items. Numeric values are often ordered in either ascending or descending order.
Unit of Measurement	The unit of measurement of the values in the table body should be indicated.
Sources	The tables should provide the primary and secondary sources for the data below the body of the table.
Footnotes and References	These are additional details for clarifying the contents of the table.

Hence, in classification, data are separated and grouped based on a property of the data common to all values. Whereas in tabulation, data is arranged into columns and rows based on its characteristics or properties. Tabulation generally emphasizes on the presentation aspects of the data, while classification is used as a means of sorting of data for further analysis.

Check Your Progress

4. What is the common factor between classification and tabulation?
5. How is raw data classified?
6. What is the basic difference between tabulation and classification of data?

10.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

NOTES

1. The basic function of a table is to simplify data and to present them in a manner that facilitates comparison.
2. If the number of *rows* is represented by *r* and of *columns* by *c,* such tables are known as *'r by c'* tables.
3. A headnote either offers some additional information about the title or qualifies the data presented in the table.
4. Classification and tabulation, are both methods of summarizing data in statistics.
5. Raw data is classified using four key characteristics geographical, chronological, qualitative and quantitative properties.
6. Tabulation generally emphasizes on the presentation aspects of the data, while classification is used as a means of sorting of data for further analysis.

10.5 SUMMARY

- Tabular presentation means tabulating the data in the form of appropriate tables. A table is a statistical table, containing data arranged into convenient number of rows and/or columns.
- The basic function of a table is to simplify data and to present them in a manner that facilitates comparison. Simplifying data means that the information desired becomes easy to locate.
- Single-column or single-row tables are the simplest to construct. The data in these tables are arranged in a single row or a single column according to time, place, region of space, or an attribute of interest. The table is vertically laid when the data are arranged in a single column, and horizontally laid when the data are arranged in a single row.
- As against single-column and single-row tables, the given data on a variable may also be arranged in multiple columns and rows. The data break-up and the kind of relational comparative picture intended determine the number of columns and rows required.
- Summary tables, on the contrary, provide only summarized data on one or more related aspects on a given subject. These are drawn from reference tables and are usually displayed in the course of running text.

NOTES

- A table must be appropriately numbered, to allow making references and citing results. It makes sense to relate the numbering of the tables with serial number of the chapter.
- A head note figures immediately below the title. It either offers some additional information about the title and/or qualifies the data presented in the table.
- The body of the table consists of a number of cells, each containing a figure called cell entry. The body contains r x c cells, and thus equal number of cell entries. Each cell occurs at the intersection of a column and a row.
- A source mentions where the data presented in the table have come from. This is an important component of the table, since the source enables the reader to check and re-check the data from where these may have been borrowed. It may also help draw, if relevant and necessary, more information from the source.
- There are no hard and fast rules governing how to frame a statistical table. It all depends on the kind of data available and the objective(s) one wishes to achieve.
- Where availability of space is a constraint in deciding the size of a table, it should be so designed that the available space accommodates the table with all the information it is supposed to contain.
- A contingency table is an important form of presenting observed data. It is amenable to the application of a number of useful statistical tools of data analysis.
- The data appearing as cell entries in a contingency table are essentially qualitative count data. To be more specific, the cell entries are observed frequencies/counts of an item or the outcome of an event possessing or not possessing a certain attribute.
- A statistical table is an orderly and systematic presentation of numerical data in columns and rows. Columns are vertical arrangements; rows are horizontal. The main objective of a statistical table is to so arrange the physical presentation of numerical facts that the attention of the reader is automatically directed to the relevant information.
- The following parts must be present in all tables:
  - Title
  - Caption
  - Stubs
  - Body

NOTES

- Classification, in statistics refers to the process of separation of data into various groups or classes with the help of properties in the data set.
- Data when it is raw, is classified using four key characteristics- geographical, chronological, qualitative and quantitative properties. Considering that a data set is gathered for the analysis of the consumption of petrol per day around the world.
- Tabulation in statistics is a method of summarising data by using a systematic arrangement of data into rows and columns. Tabulation is carried out as to investigate, compare, identify errors or omissions in data, to study a prevailing trend, to simplify the known raw data and to use the space economically and use it as future reference.

10.6 KEY WORDS

- **Headnote:** It is a statement, given below the title, which clarifies the contents of the table.
- **Footnote:** It is a statement which clarifies some specific items given in the table or explains the omission thereof.
- **Stubs:** They are the titles of the rows.
- **Reference tables:** These tables present extensive information on any subject; for all practical purposes, these tables are the repository of basic data and work almost as data inventory.

10.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. What is tabular presentation of data? How does it facilitate comparison?
2. List the various types of tables.
3. What is the difference between footnote and headnote?
4. Differentiate between classification and tabulation of data.

Long Answer Questions

1. Enumerate the components of a table.
2. Discuss the steps involved in framing a table.
3. What is classification of data? Why is it necessary to classify data? Give an example where data is classified.

4. What are statistical tables? State the objectives of statistical tables. Also list the advantages of tabular presentation of data.

10.8 FURTHER READINGS

Creswell, John W. 2002. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* London: Sage Publications.

Booth, Wayne, Gregory G. Colomb and Joseph M. Williams. 1995. *The Craft of Research.* Chicago: University of Chicago Press.

Kumar, B. 2006. *Research Methodology.* New Delhi: Excel Books.

Paneerselvam, R. 2009. *Research Methodology.* New Delhi: Prentice Hall of India.

Gupta, D. 2011. *Research Methodology*. New Delhi: PHI Learning Private Limited.

NOTES

NOTES

BLOCK - IV
MEASURES OF CENTRAL TENDENCY, DISPERSION AND DIAGRAMMATICS

UNIT 11 MEASURES OF CENTRAL TENDENCY

Structure

11.0 INTRODUCTION

We have already discussed how raw data can be organized in the form of tables, charts and frequency distributions in order to understand its features. Although frequency distributions and graphical display make raw data more meaningful, yet fail to understand its three major properties. These three properties are as follows:

(i) ***Central tendency***: A numerical value around which most numerical values in the data set tend to cluster.

(ii) ***Variation***: The extent to which numerical values cluster or scatter around the central value.

(iii) ***Skewness***: The extent of dispension or variation from the symmetry of a distribution.

The statistical methods that are used to extract and measure these three features: *central tendency, variation and skewness* in the data set are called

descriptive (or summary) measures. There are three types of descriptive measures:

(i) Measures of central tendency

(ii) Measures of dispersion or variation

(iii) Measures of symmetry—skewness

These measures can also be used for comparing two or more populations with respect to central tendency, variation and skewness to draw useful inferences.

NOTES

11.1 OBJECTIVES

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

- Understand the role of descriptive statistics in summarization, description and interpretation of the data
- Understand the importance of summary measures to describe characteristics of a data set
- Explain several numerical methods belonging to measures of central tendency to describe the characteristics of a data set

11.2 UNDERSTANDING CENTRAL TENDENCY

The term 'central tendency' was coined because numerical values in most data sets show a distinct tendency to cluster around a value of an observation located somewhere in the middle of all numerical values (observations). It is necessary to know this *central value* (also called *average*) to describe characteristic of the data set. Statistical methods of computing the central value are called ***measures of central tendency***.

If the descriptive measures are computed using sample data, then these are called **sample statistic** but if these measures are computed using data of the population, they are called **population parameters**. The population parameter (also average or mean value) is represented by the Greek letter μ (read: mu) and sample statistic is represented by the Roman letter $\bar{x}$ (read: *x* bar).

11.2.1 Objectives of Averaging

Few of the objectives to calculate a central value (or average) in order to describe the characteristic of the given data set are given below:

1. It is useful to extract and summarize the characteristics of the data set in a precise form. For example, to understand individual families' need for water during summers, the knowledge of the average quantity of

NOTES

water needed for the entire population will help in planning for water resources.

2. Knowledge of 'average' value facilitates comparison between two or more populations with respect to a particular attribute. For example, average sale of any product for any month can be compared with the preceding months, or even with the sale of similar product of competitive companies for the same months.
3. It offers a base for computing various other measures such as dispersion, skewness and kurtosis that help in statistical analysis.

11.2.2 Requisites of a Measure of Central Tendency

The following are the few requirements to be satisfied by a measure of central tendency:

1. **Rigidly defined:** The definition of a measure of central tendency should be clear and rigidly defined by an algebraic formula to bring in uniformity in its interpretation by decision makers.
2. **Based on all the observations:** The value of a measure of central tendency should be calculated by taking into consideration the entire data set.
3. **Least variation in sample results:** The value of a measure of central tendency derived from independent random samples of the same size from a given population should not vary much from another. The amount of difference (if any) in the values is considered to be the sampling error.
4. **Capable of algebraic treatment:** The nature of the average should be such that it could be used for statistical analysis of the data set. For example, it should be possible to determine the average production in a particular year by the use of average production in each month of that year.
5. **Unaffected by extreme observations:** To represent truly the characteristics of the entire data set, the value of a measure of central tendency should not be affected by very small or large numerical value of observations in the data set.

Check Your Progress

1. What are the types of descriptive measures?
2. What do you understand by sample statistic and population parameters?

11.3 MEASURES OF CENTRAL TENDENCY: CHARACTERISTICS

NOTES

The various measures of central tendency or averages can be broadly classified in the following categories:

1. Mathematical Averages

(a) Arithmetic mean (also called mean or average)
- Simple
- Weighted

(b) Geometric mean

(c) Harmonic mean

2. Averages of Position

(a) Median

(b) Quartiles

(c) Deciles

(d) Percentiles

(e) Mode

In this unit, you will learn about, besides mathematical averages, averages of position, such as median and mode.

11.3.1 Mathematical Averages: Arithmetic, Geometric and Harmonic Means

Various methods of calculating mathematical averages in a data set are classified in accordance with the nature of data available, i.e., ungrouped (unclassified or raw) or grouped (classified) data.

1. Arithmetic Mean of Ungrouped Data

There are two methods for calculating **Arithmetic Mean (A.M.)** for ungrouped (or unclassified) data:

(i) Direct method

(ii) Indirect or short-cut method

Direct Method In this method, A.M. is calculated by adding numerical value of all observations and dividing the total by the number of observations. Thus, if x_1, x_2,..., x_N represent the numerical value of N observations in a population, then A.M. of population is:

$$\text{Population mean, } \mu \quad = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (11\text{-}1a)$$

NOTES

However, for a sample containing n observations x_1, x_2,..., x_n, the sample A.M. can be written as:

$$\text{Sample mean, } \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (11\text{-}1b)$$

The denominator in these formulae is different because in statistical analysis the uppercase letter N represents the number of observations in the population, while the lower case letter *n* represents the number of observations in the sample.

Example 11.1: In a survey of 5 cement companies, the profit (in ₹crore) earned during a year was 15, 20, 10, 35 and 32. Find the arithmetic mean of the profit earned.

Solution: Applying the formula (11-1b), we have

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{5} x_i = \frac{1}{5}\,[15 + 20 + 10 + 35 + 32] = 22.4$$

Thus, the arithmetic mean of the profit earned by these companies during a year was ₹22.4 crore.

Alternative Formula: In general, when numerical values, x_i (i = 1, 2,..., n) are arranged in a frequency distribution, then A.M. formula (11-1b) should be modified as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i x_i \quad (11\text{-}2)$$

where f_i represents the frequency (number of observations) with which variable x_i occurs in the data set, i.e. $n = \sum_{i=1}^{n} f_i$.

Example 11.2: If A, B, C and D are four chemicals costing ₹15, ₹12, ₹8 and ₹5 per 100 g, respectively, and are contained in a given compound in the ratio of 1, 2, 3 and 4 parts, respectively, then what should be the price of the resultant compound.

Solution: Using the formula (11-2), the sample arithmetic mean is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{4} f_i x_i = \frac{1\times 15 + 2\times 12 + 3\times 8 + 4\times 5}{1 + 2 + 3 + 4} = ₹8.30$$

Thus, the average price of the resultant compound should be ₹8.30 per 100 g.

Example 11.3: The number of new orders received by a company over the last 25 working days were recorded as follows: 3, 0, 1, 4, 4, 4, 2, 5, 3, 6, 4, 5, 1, 4, 2, 3, 0, 2, 0, 5, 4, 2, 3, 3, 1. Calculate the arithmetic mean for the number of orders received during these working days.

NOTES

Solution: Applying the formula (11-1b), the sample arithmetic mean ($\bar{x}$) is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{25} x_i = \frac{1}{25}[3 + 0 + 1 + 4 + 4 + 4 + 2 + 5 + 3 + 6 + 4 + 5 + 1 + 4 + 2 + 3 + 0 + 2 + 0 + 5 + 4 + 2 + 3 + 3 + 1]$$

$$= \frac{1}{25}(71) = 2.84 \cong 3 \text{ orders (approx.)}$$

Alternative Approach: Use of formula (11-2)

Table 11.1 *Calculation of Mean ($\bar{x}$) Value*

Number of Orders (x_i)	*Frequency (f_i)*	$f_i x_i$
0	3	0
1	3	3
2	4	8
3	5	15
4	6	24
5	3	15
6	1	6
	25	71

Arithmetic mean, $\bar{x} = \frac{1}{n}\sum f_i x_i = \frac{71}{25} = 2.84 \cong 3$ orders (approx.)

Example 11.4: From the following information on the number of defective components in 1000 boxes:

Number of defective components	:	0	1	2	3	4	5	6
Number of boxes	:	25	306	402	200	51	10	6

Calculate the arithmetic mean of defective components for the whole of the production line.

Solution: The calculations of mean defective components for the whole production line are shown in Table 11.2

Table 11.2 *Calculations of Mean $\bar{x}$ Value*

Number of Defective Components (x_i)	*Number of Boxes (f_i)*	$f_i x_i$
0	25	0
1	306	306
2	402	804
3	200	600
4	51	204
5	10	50
6	6	36
	1000	2000

Applying the formula (11-2), the arithmetic mean ($\bar{x}$) is

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{6} f_i x_i = \frac{1}{1000}(2000) = 2 \text{ defective components.}$$

NOTES

Indirect or Short-cut Method In this method, an arbitrary *assumed mean* is used as a basis for calculating deviations from individual values in the data set. Let A be the arbitrary assumed A.M. and let

$$d_i = x_i - \text{A or } x_i = \text{A} + d_i$$

Substituting the value of x_i in formula (11-1b), we have

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

$$= \text{A} + \frac{1}{n}\sum_{i=1}^{n} d_i \qquad (11\text{-}3)$$

If frequencies of the numerical values are also taken into consideration, then formula (11-3) becomes:

$$\bar{x} = \text{A} + \frac{1}{n}\sum_{i=1}^{n} f_i d_i \qquad (11\text{-}4)$$

where, $n = \sum_{i=1}^{n} f_i$ = total number of observations in the sample.

Example 11.5: The daily earnings (in rupees) of 175 employees working on a daily basis in a firm are:

Daily earnings (₹)	:	100	120	140	160	180	200	220
Number of employees	:	3	6	10	15	24	42	75

Calculate the average daily earning for all employees.

Solution: The calculations of average daily earning for employees are shown in Table 11.3.

Table 11.3 *Calculations of Mean* $\bar{x}$ *Value*

Daily Earnings (in ₹) (x_i)	*Number of Employees* (f_i)	$d_i = x_i - A = x_i - 160$	$f_i d_i$
100	3	−60	−180
120	6	−40	−240
140	10	−20	−200
160 ← A	15	0	0
180	24	20	480
200	42	40	1680
220	75	60	4500
	175		6040

Suppose assumed mean is A = 160. The required A.M. ($\bar{x}$) using the formula (11-4) is given by

$$\bar{x} = \text{A} + \frac{1}{n}\sum_{i=1}^{7} f_i d_i = 160 + \frac{6040}{175} = ₹194.51$$

Example 11.6: The human resource manager at a city hospital began a study of the overtime hours of the registered nurses. Twenty five nurses were selected at random, and their overtime hours during a month were recorded:

13 13 12 15 7 15 5 12 6 7 12 10 9 13 12
5 9 6 10 5 6 9 6 9 12

Calculate the arithmetic mean of overtime hours during the month.

Solution: Calculations of arithmetic mean of overtime hours are shown in Table 11.4.

Table 11.4 *Calculations of Mean ($\bar{x}$) Value*

Overtime Hours (x_i)	*Number of Nurses* (f_i)	$d_i = x_i - A = x_i - 10$	$f_i d_i$
5	3	–5	–15
6	4	–4	–16
7	2	–3	–6
9	4	–1	–4
10 ← A	2	0	0
12	5	2	10
13	3	3	9
15	2	5	10
	25		–12

Supposed assumed mean is, A=10. The required arithmetic mean ($\bar{x}$) of overtime using the formula (11-4) is as follows:

$$\bar{x} = A + \frac{1}{n}\sum_{i=1}^{25} f_i d_i = 10 - \frac{12}{25} = 9.52 \text{ hours}$$

2. Arithmetic Mean of Grouped Data

Arithmetic mean for grouped or classified data set can also be calculated by applying any of the following methods:

(i) Direct method

(ii) Indirect or step-deviation method

For calculating arithmetic mean for a grouped data set, the following assumptions are made:

(i) The class intervals must be closed.

(ii) The width of each class interval should be equal.

(iii) The mid-value of each class interval must represent the average of all values in that class. That is, it is assumed that all values of observations are evenly distributed between the lower and upper class limits.

NOTES

Direct Method The formula used in this method is same as formula (11-2) except that x_i is replaced with the mid-point value of class intervals. The new formula becomes:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i m_i \qquad (11\text{-}5)$$

where m_i is mid-value of ith class interval, f_i is frequency of ith class interval and $n = \Sigma f_i$ is the sum of all observations (or frequencies).

Example 11.7: A company is planning to improve plant safety. For this, accident data for the last 50 weeks was compiled. These data are grouped into the frequency distribution as shown below. Calculate the A.M. of the number of accidents per week.

Number of accidents	:	0–4	5–9	10–14	15–19	20–24
Number of weeks	:	5	22	13	8	2

Solution: The calculations of A.M. are shown in Table 11.5 using formula (11-5).

Table 11.5 *Arithmetic Mean of Accidents*

Number of Accidents	*Mid-value* (m_i)	*Number of Weeks* (f_i)	$f_i m_i$
0–4	2	5	10
5–9	7	22	154
10–14	12	13	156
15–19	17	8	136
20–24	22	2	44
		50	500

The A.M. of the number of accidents per week is

$$\text{Arithmetic mean, } \bar{x} = \frac{1}{n}\sum_{i=1}^{5} f_i m_i = \frac{500}{50} = 10 \text{ accidents per week.}$$

Step-deviation Method The formula (11-5) for calculating A.M. can be improved as formula (11-6). This improved formula is also known as the s*tep-deviation method:*

$$\bar{x} = A + \left\{\frac{1}{n}\sum f_i d_i\right\} \times h \qquad (11\text{-}6)$$

where h is the width of the class intervals and $d_i = \frac{m_i - A}{h}$ is deviation from the assumed mean.

Example 11.8: A company is planning to improve plant safety. For this, accident data for the last 50 weeks was compiled. These data are grouped into the frequency distribution as shown below.

Number of accidents	:	0–4	5–9	10–14	15–19	20–24
Number of weeks	:	5	22	13	8	2

NOTES

Calculate the arithmetic mean of accidents per week by using s*tep-deviation method*.

Solution: The calculations of the average number of accidents are shown in the Table 11.6.

Table 11.6 *Arithmetic Mean of Accidents*

Number of Accidents	*Mid-value* (m_i)	$d_i = (m_i - A)/h = (m_i - 12)/5$	*Number of Weeks* (f_i)	$f_i d_i$
0–4	2	–2	5	–10
5–9	7	–1	22	–22
10–14	12 ← A	0	13	0
15–19	17	1	8	8
20–24	22	2	2	4
			50	–20

$$\text{Arithmetic mean, } \bar{x} = A + \left\{\frac{1}{n}\sum f_i d_i\right\} \times h$$

$$= 12 + \left\{\frac{1}{50}(-20)\right\} 5 = 10 \text{ accidents per week}$$

Example 11.9: The following distribution gives the pattern of overtime work done by 100 employees of a company. Calculate the average overtime work done per employee.

Overtime hours	:	10–15	15–20	20–25	25–30	30–35	35–40
Number of employees	:	11	20	35	20	8	6

Solution: The calculations of the average overtime work done per employee with assumed mean, A = 22.5 and class width, h = 5 are given in Table 11.7.

Table 11.7 *Calculations of Average Overtime*

Overtime (hrs) x_i	*Number of Employees*, (f_i)	*Mid-value* (m_i)	$d_i = (m_i - 22.5)/5$	$f_i d_i$
10–15	11	12.5	–2	–22
15–20	20	17.5	–1	–20
20–25	35	22.5 ← A	0	0
25–30	20	27.5	1	20
30–35	8	32.5	216	
35–40	6	37.5	3	18
	100			12

$$\text{The required A.M. is } (\bar{x}) = A + \left\{\frac{1}{n}\sum f_i d_i\right\} \times h = 22.5 + \frac{12}{100} \times 5 = 23.1 \text{ hrs}$$

NOTES

Example 11.10: The following is the age distribution of 1000 persons working in an organization

Age Group	*Number of Persons*	*Age Group*	*Number of Persons*
20–25	30	45–50	105
25–30	160	50–55	70
30–35	210	55–60	60
35–40	180	60–65	40
40–45	145		

Due to continuous losses, it is desired to bring down the manpower strength to 30 per cent of the present number according to the following scheme:

(a) Retrench the first 15 per cent from the lower age group.

(b) Absorb the next 45 per cent in other branches.

(c) Make 10 per cent from the highest age group retire permanently, if necessary.

Calculate the age limits of the persons retained and those to be transferred to other departments. Also, find the average age of those retained.

Solution: (a) The first 15 per cent persons to be retrenched from the lower age groups are (15/100)×1000 = 150. But the lowest age group 20–25 has only 30 persons and therefore the remaining, 150 – 30 = 120 will be taken from next higher age group, that is, 25–30, which has 160 persons.

(b) The next 45 per cent, that is, (45/100)×1000 = 450 persons who are to be absorbed in other branches belong to the following age groups:

Age Groups	*Number of Persons*
25–30	(160 – 120) = 40
30–35	210
35–40	180
40–45	(450 – 40 – 210 – 180) = 20
	450

(c) Those who are likely to be retired are 10 per cent, that is, (10/100)×1000 = 100 persons and belong to the following highest age groups:

Age Group	*Number of Persons*
55–60	60
60–65	40

Hence, the calculations of the average age of those retained and/or to be transferred to other departments are shown in Table 11.8:

Table 11.8 *Calculations of Average Age*

Age Group (x_i)	*Mid value,* (m_i)	*Number of Persons* (f_i)	$d_i = (x_i - 47.5)/5$	$f_i d_i$
40–45	42.5	145 – 20 = 125	–1	–125
45–50	47.5 ← A	105	0	0
50–55	52.5	70	1	70
		300		–55

The required average age is, $\bar{x} = A + \left\{\frac{1}{n}\sum f_i d_i\right\} \times h = 47.5 - \frac{55}{300} \times 5 =$ 46.58 = 47 years (approx.).

3. Weighted Arithmetic Mean

While calculating arithmetic mean, as discussed earlier, equal importance (or weight) is given to each observation in the data set. However, there are situations in which values of individual observations in the data set are not of equal importance. If such values occur with different frequencies, then computing A.M. of values (as opposed to the A.M. of observations) may not be a true representative of the data set characteristic and thus may be misleading. Under these circumstances, we may attach to each observation value a 'weight' $w_1, w_2, ..., w_N$ as an indicator of their importance within the data set and compute a weighted mean or average denoted by $\bar{x}_w$ as follows:

$$\mu_w \text{ or } \bar{x}_w = \frac{\sum x_i w_i}{\sum w_i}$$

Remark: The **weighted arithmetic mean** should be used

- When the importance of all the numerical values in the given data set is not equal.
- When the frequencies of various classes are widely varying.
- Where there is a change either in the proportion of numerical values or in the proportion of their frequencies.
- When ratios, percentages or rates are being averaged.

Example 11.11: An examination was held to decide for awarding of a scholarship. The weights of various subjects were different. The marks obtained by 3 candidates (out of 100 in each subject) are given in the following table:

NOTES

Subject	Weight	Students		
		A	B	C
Mathematics	4	60	57	62
Physics	3	62	61	67
Chemistry	2	55	53	60
English	1	67	77	49

Calculate the weighted A.M. to award the scholarship.

Solution: The calculations of the weighted arithmetic mean are shown in Table 11.9.

Table 11.9 *Calculations of Weighted Arithmetic Mean*

Subject	Weight	Students					
	(w_i)	Student A		Student B		Student C	
		Marks (x_i)	$x_i w_i$	Marks (x_i)	$x_i w_i$	Marks (x_i)	$x_i w_i$
Mathematics	4	60	240	57	228	62	248
Physics	3	62	186	61	183	67	201
Chemistry	2	55	110	53	106	60	120
English	1	67	67	77	77	49	49
	10	244	603	248	594	238	618

Applying the formula for weighted mean, we get

$$\bar{x}_{wA} = \frac{603}{10} = 60.3 \; ; \; \bar{x}_A = \frac{244}{4} = 61$$

$$\bar{x}_{wB} = \frac{594}{10} = 59.4 \; ; \; \bar{x}_B = \frac{248}{4} = 62$$

$$\bar{x}_{wC} = \frac{618}{10} = 61.8 \; ; \; \bar{x}_C = \bar{x}_C = 59.5$$

From above calculations, it may be noted that student B should get the scholarship as per simple A.M. values, but according to weighted A.M., student C should get the scholarship because all the subjects of examination are not of equal importance.

Example 11.12: The owner of a general store was interested in knowing the mean contribution (sales price minus variable cost) of his stock of 5 items. The data is given below:

Product	Contribution per Unit	Quantity Sold
1	6	160
2	11	60
3	8	260
4	4	460
5	14	110

Solution: If the owner ignores the values of the individual products and gives equal importance to each product, then the mean contribution per unit sold will be

$$\bar{x} = (1/5)\ \{6 + 11 + 8 + 4 + 14\} = ₹8.6$$

However, ₹8.60, may not necessarily be the mean contribution per unit of different quantities of the products sold. In this case, the owner has to take into consideration the number of units of each product sold as different weights. Computing weighted A.M. by multiplying units sold (w) of a product by its contribution (x). That is,

$$\bar{x}_w = \frac{6(160) + 11(60) + 8(260) + 4(460) + 14(110)}{160 + 60 + 260 + 460 + 110} = \frac{7,080}{1,050} = ₹6.74$$

This value, ₹6.74, is different from the earlier value, ₹8.60. The owner must use the value ₹6.74 for decision-making purpose.

Example 11.13: A management consulting firm, has four types of professionals on its staff: managing consultants, senior associates, field staff and office staff. Average rates charged to consulting clients for the work of each of these professional categories are ₹3150/hour, ₹1680/hour, ₹1260/hour and 630/hour, respectively. Office records indicate the following number of hours billed last year in each category: 8000, 14,000, 24,000 and 35,000, respectively. If the firm is trying to come up with an average billing rate for estimating client charges for next year, what would you suggest they do and what do you think is an appropriate rate?

Solution: The data given in the problem are as follows:

Staff	*Consulting Charges (₹per hour)* x_i	*Hours Billed* w_i
Managing consultants	3150	8000
Senior associates	1680	14,000
Field staff	1260	24,000
Office staff	630	35,000

Applying the formula for weighted mean, we get

$$\bar{x}_w = \frac{\Sigma x_i w_i}{\Sigma w_i} = \frac{3150(8000) + 1680\,(14,000) + 1260\,(24,000) + 630\,(35,000)}{8000 + 14,000 + 24,000 + 35,000}$$

$$= \frac{2,52,00,000 + 2,35,20,000 + 3,02,40,000 + 2,20,50,000}{81,000}$$

= ₹1247.03 per hour

However, the firm should cite this as an average rate for clients who use the four professional categories for approximately 10 per cent, 17 per cent, 30 per cent and 43 per cent of the total hours billed.

NOTES

Advantages and Disadvantages of Arithmetic Mean

Advantages

1. The calculation of arithmetic mean is simple and is unique in every data.
2. The calculation of arithmetic mean is based on all values given in the data set.
3. The arithmetic mean is reliable single value that represents all values in the data set.
4. The arithmetic mean is least affected by variations in the sample size. In other words, arithmetic mean determined from various samples drawn from a population, varies by the least possible amount.
5. Some of the algebraic properties of arithmetic mean are as follows:

(a) The algebraic sum of deviations of all the observations x_i *(i = 1, 2..., n)* from A.M. is always zero, that is,

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

The difference $x_i - \bar{x}\,(i = 1, 2,..., n)$ is usually referred to as *deviation from the arithmetic mean*. Thus, mean is characterized as a *point of balance,* i.e. sum of the positive deviations from mean is equal to the sum of the negative deviations from mean.

(b) The sum of the squares of the deviations of all the observations from A.M. is less than the sum of the squares of all the observations from any other quantity.

If $x_i\,(i = 1, 2,..., n)$ are the given observation and $\bar{x}$ be their arithmetic mean, then this property implies that

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

where '*a*' is any constant quantity.

This property of A.M. is also known as the *least square property* and is helpful in calculating standard deviation.

(c) It is possible to calculate the combined (or pooled) arithmetic mean of two or more than two data sets of the same nature.

Let $\bar{x}_1$ and $\bar{x}_2$ be arithmetic means of two data sets of the same nature, of size n_1 and n_2, respectively. Then their combined A.M. can be calculated as

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \qquad (11\text{-}7)$$

The result (11-7) can also be generalized in the same way for more than two data sets.

(d) If observations in any data set are wrongly recorded, then in such a case, the correct value of A.M. is calculated first by subtracting the sum of observations wrongly recorded from the total of all observations, Σx_i and then adding the sum of the correct observations to it. The result is then divided by the total number of observations.

Disadvantages

1. The value of A.M. cannot be calculated for unequal and open-ended class intervals at beginning or end of frequency distribution.
2. The value of A.M. is affected by the extreme observations (or outliers) present in the data set. Outliers at the high end increase the mean value, while outliers at the lower end decrease it.
3. The mean cannot be calculated for qualitative characteristics such as intelligence, honesty, beauty or loyalty.

4. Geometric Mean

In many business and economics problems, such as calculation of compound interest and inflation, quantities (variables) change over a period of time. In such cases, a decision maker may like to know an average percentage change rather than simple average value to represent the average growth or declining rate in the variable value over a period of time. Thus, another measure of central tendency called **geometric mean** (G.M.) is calculated.

For example, consider the annual growth rate of output of a company in the last five years.

Year	*Growth Rate (Per cent)*	*Output at the End of the Year*
2006	5.0	105.00
2007	7.5	112.87
2008	2.5	115.69
2009	5.0	121.47
2010	10.0	133.61

The simple arithmetic mean of the growth rate is

$$\bar{x} = \frac{1}{5}(5 + 7.5 + 2.5 + 5 + 10) = 6$$

This value of 'mean' implies that if 6 per cent is the growth rate, then output at the end of year 2012 should be 133.81, which is slightly more than the actual value, 133.61. Thus, the correct growth rate should be less than 6.

To find the correct growth rate, we apply the formula of geometric mean:

$$\text{G.M.} = \sqrt[n]{\text{Product of all the } n \text{ values}}$$

$$= \sqrt[n]{x_1 \cdot x_2 \ldots x_n} = (x_1 \cdot x_2 \cdot \ldots x_n)^{\frac{1}{n}} \qquad (11\text{-}8)$$

NOTES

In other words, G.M. of a set of n observations is the nth root of their product.

Substituting the values of growth rate in the given formula, we have

$$\text{G.M.} = \sqrt[5]{5 \times 7.5 \times 2.5 \times 5 \times 10} = \sqrt[5]{4687.5}$$

= 5.9 per cent average growth.

Calculation of Geometric Mean If the number of observations are more than three, then G.M. can be calculated by taking logarithm on both the sides of the equation. The formula (11-8) for G.M. for ungrouped data can be expressed in terms of logarithm as shown below:

$$\text{Log (G.M.)} = \frac{1}{n} \log (x_1 \cdot x_2 \cdot \ldots \cdot x_n)$$

$$= \frac{1}{n} \{\log x_1 + \log x_2 + \ldots + \log x_n\} = \frac{1}{n} \sum_{i=1}^{n} \log x_i$$

and therefore $$\text{G.M.} = \text{antilog} \left\{ \frac{1}{n} \sum \log x_i \right\} \quad (11\text{-}9)$$

If the observations $x_1, x_2, \ldots, x_n$ occur with frequencies $f_1, f_2, \ldots, f_n$, respectively, and the total of frequencies are, $n = \Sigma f_i$ then the G.M. for such data is given by

$$\text{G.M.} = \left(x_1^{f_1} \cdot x_2^{f_2} \cdot \ldots \cdot x_n^{f_n} \right)^{1/n}$$

or $$\log (\text{G.M.}) = \frac{1}{n} \{f_1 \log x_1 + f_2 \log x_2 + \ldots + f_n \log x_n\}$$

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

or $$\text{G.M.} = \text{Antilog} \left(\frac{1}{n} \sum f_i \log x_i \right) \quad (11\text{-}10)$$

Example 11.14: The rate of increase in population of a country during the last three decades is 5 per cent, 8 per cent and 12 per cent. Find the average rate of growth during the last three decades.

Solution: Since the data is given in terms of percentage, therefore geometric mean is a more appropriate measure. The calculations of geometric mean are shown in Table 11.10:

Table 11.10 *Calculations of G.M.*

Decade	*Rate of Increase in Population (%)*	*Population at the End of Decade (x) Taking Preceding Decade as 100*	$\log_{10} x$
1	5	105	2.0212
2	8	108	2.0334
3	12	112	2.0492
			6.1038

Using the formula (11-10), we have

$$\text{G.M.} = \text{Antilog}\ \frac{1}{n}\sum \log x = \text{Antilog}\ \frac{1}{3}(6.1038)$$

$$= \text{Antilog}\ (2.0346) = 108.2$$

Hence, the average rate of increase in population over the last three decades is 108.2 – 100 = 8.2 per cent.

Example 11.15: A given machine is assumed to depreciate 40 per cent in value in the first year, 25 per cent in the second year and 10 per cent per year for the next three years, each percentage being calculated on the diminishing value. What is the average depreciation recorded on the diminishing value for the period of five years?

Solution: The calculations of geometric mean are shown in Table 11.11.

Table 11.11 *Calculations of G.M.*

Rate of Depreciation (x_i) (*in percentage*)	*Number of Years* (f_i)	$\log_{10} x_i$	$f_i \log_{10} x_i$
40	1	1.6021	1.6021
25	1	1.3979	1.3979
10	3	1.0000	3.0000
			6.0000

Using formula (3-11), we have

$$\text{G.M.} = \text{Antilog}\left\{\frac{1}{n}\sum f_i \log_{10} x_i\right\} = \text{Antilog}\ \frac{1}{5}(6.0000)$$

$$= \text{Antilog}\ (1.2) = 15.85$$

Hence, the average rate of depreciation for first five years is 15.85 per cent.

Combined Geometric Mean

The combined geometric mean of observations is obtained by pooling the geometric means of different data set as follows:

$$\log \text{G.M.} = \frac{\sum_{i=1}^{n} n_i \log G_i}{\sum_{i=1}^{n} n_i} \qquad (11\text{-}11)$$

where G_i is the geometric mean of the *i*th data set having n_i number of observations.

NOTES

Weighted Geometric Mean

If different observations $x_i (i = 1, 2, ..., n)$ are given different weights (importance), say $w_i (i = 1, 2, ..., n)$, respectively, then their weighted geometric mean is defined as

$$\text{G.M. } (w) = \text{Antilog} \left[\left(\frac{1}{\sum w} \right) \sum w \log x \right] \tag{11-12}$$

Example 11.16: Three sets of data contain 8, 7 and 5 observations and their geometric means are 8.52, 10.12 and 7.75, respectively. Find the combined geometric mean of these 20 observations.

Solution: Applying the formula (11-11), the combined geometric mean can be obtained as follows:

$$\text{G.M.} = \text{Antilog} \left[\frac{n_1 \log G_1 + n_2 \log G_2 + n_3 \log G_3}{n_1 + n_2 + n_3} \right]$$

$$= \text{Antilog} \left[\frac{8 \log (8.52) + 7 \log (10.12) + 5 \log (7.75)}{8 + 7 + 5} \right]$$

$$= \text{Antilog} \left[\frac{(8 \times 0.9304) + (7 \times 1.0051) + (5 \times 0.8893)}{20} \right]$$

$$= \text{Antilog} \left(\frac{18.9254}{20} \right) = \text{Antilog } (0.94627) = 8.835$$

Hence, the combined G.M. of 20 observations is 8.835.

Example 11.17: The weighted geometric mean of four numbers 8, 25, 17 and 30 is 15.3. If the weights of the first three numbers are 5, 3 and 4, respectively, find the weight of fourth number.

Solution: Let weight of fourth number be w. Then the weighted geometric mean of four numbers can be calculated as shown in Table 11.12.

Table 11.12 *Calculations of Weighted G.M.*

Numbers (x)	*Weight of Each Number* (w)	$log_{10} x$	$w \log_{10} x$
8	5	0.9031	4.5155
25	3	1.3979	4.1937
17	4	1.2304	4.9216
30	w	1.4771	1.4771w
	$12 + w$		$13.6308 + 1.4771w$

Thus the weighted G.M. is

$$\log \{\text{G.M. } (w)\} = \left[\left(\frac{1}{\sum w} \right) \sum w \log x \right]$$

$$\text{or} \quad \log (15.3) = \left[\left(\frac{1}{12 + w} \right) (13.6308 + 1.4771 w) \right]$$

$$(1.1847)\ (12 + w) = 13.6308 + 1.4771w$$

$$14.2164 + 1.1847w = 13.6308 + 1.4771w$$

$$0.5856 = 0.2924\ w$$

or $$w = \frac{0.5856}{0.2924} = 2 \text{ (approx.)}$$

NOTES

Thus, the weight of fourth number is 2.

Advantages, Disadvantages and Applications of G.M.

Advantages

1. The value of G.M. is not much affected by extreme observations and is computed by taking all the observations into account.
2. In the calculation of G.M. more weight is given to smaller values and less weight to higher values. For example, it is useful in the study of price fluctuations where the lower limit can touch zero whereas the upper limit may go up to any number.
3. Algebraic manipulations of the original formula of geometric mean is possible. The calculation of weighted G.M. and combined G.M. are two examples.

Disadvantage

The value of G.M. cannot be calculated in case any of the observations in the data set is either negative or zero.

Applications

1. The concept of G.M. is used in the construction of index numbers.
2. Since G.M. ≤ A.M., G.M. is useful in those cases where smaller observations are to be given importance. Such cases usually occur in the study of social and economic problems.
3. The G.M. of a data set is useful in estimating the average rate of growth per unit per period such as percentage increase in sales, profit, production, population and so on. Also in calculating the amount of money accumulated at the end of n periods, with principal amount, P_0 as follows:

$$P_n = P_0(1 + r)^n$$

or $$r = \left(\frac{P_n}{P_0}\right)^{\frac{1}{n}} - 1$$

where r is the interest rate per unit period and n is the length of the period.

NOTES

5. Harmonic Mean

The **harmonic mean** (H.M.) of a set of observations is defined as the reciprocal of the arithmetic mean and is calculated by taking reciprocal of the individual observations, i.e.,

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

or $$\text{H.M.} = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)} \qquad \text{(For ungrouped data)} \qquad (11\text{-}13)$$

If $f_1, f_2, ..., f_n$ are the frequencies of observations $x_1, x_2, ..., x_n$, then the harmonic mean is defined as

$$\text{H.M.} = \frac{n}{\sum_{i=1}^{n} f_i\left(\frac{1}{x_i}\right)} \qquad \text{(For grouped data)} \qquad (11\text{-}14)$$

where $n = \sum_{i=1}^{n} f_i$.

Example 11.18: An investor buys ₹20,000 worth of shares of a company each month. During the first 3 months he bought the shares at a price of ₹120, ₹160 and ₹210. After 3 months what is the average price paid by him for the shares?

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

$$\text{H.M.} = \frac{3}{(1/120)+(1/160)+(1/210)} = \frac{3}{0.008+0.006+0.004}$$

$$= 3/0.018 = ₹166.66$$

Example 11.19: Find the harmonic mean of the following distribution of data

Dividend yield (per cent)	:	2–6	6–10	10–14
Number of companies	:	10	12	18

Solution: The calculation of harmonic mean is shown in Table 11.13.

Table 11.13 *Calculation of H.M.*

Class Intervals (*Dividend yield*)	*Mid-value* (m_i)	*Number of Companies* (*frequency*, f_i)	*Reciprocal* $\left(\frac{1}{m_i}\right)$	$f_i\left(\frac{1}{m_i}\right)$
2 – 6	4	10	1/4	2.5
6 – 10	8	12	1/8	1.5
10 – 14	12	18	1/12	1.5
		40		5.5

The harmonic mean is H.M. $= \frac{n}{\sum_{i=1}^{3} f_i\left(\frac{1}{m_i}\right)} = \frac{40}{5.5} = 7.27$

Hence, the average dividend yield of 40 companies is 7.27 per cent.

NOTES

Advantages, Disadvantages and Applications of H.M.

Advantages

1. The H.M. is computed based on every observation in the data set.
2. Higher weight age is given to smaller values in a data set because the reciprocal of numerical values is taken for the calculation of H.M.
3. Certain algebraic changes can be made in the formula of H.M. for further analysis of data.

Disadvantages

1. The H.M. is not often used for analysing business problems.
2. The H.M. cannot be calculated if a data set has negative and/or zero elements.
3. For calculating H.M., the higher weight is given to smaller values in the data set. Thus, it does not represent the true characteristic of the data set.

Applications The harmonic mean is particularly useful for computation of average rates and ratios. Such rates and ratios are generally used to define relations between two different types of measuring units expressed reciprocally. For example, distance (in km) and time (in hours).

Relationship among A.M., G.M. and H.M.

For any data set the relationship between A.M., G.M., and H.M. is A.M. ≥ G.M. ≥ H.M. If numerical values in a data set can be expressed in a form: a, ar, ar^2,..., ar^{n-1}, then relationship becomes $(\text{G.M.})^2 = \text{A.M.} \times \text{H.M.}$

11.3.2 Averages of Position: Median and Mode

Mathematical averages – arithmetic mean, geometric mean and harmonic mean, measure quantitatively characteristics of a data set, such as, income, profit, level of production, rate of growth, etc. However, to guard against the influence of outliers, and/or to measure qualitative characteristics of a data set, such as honesty, intelligence, beauty, consumer acceptance, and so on, other measures of central tendency namely *median*, *quartiles*, *deciles*, *percentiles* and *mode* are used. These measures are also called *positional averages*. The term 'position' refers to the place of the value of an observation in the data set.

NOTES

1. Median

Median may be defined as the *middle value* (half of the observations are smaller and half are larger than this value) in the data set when elements are arranged in a sequential (either ascending or descending) order of magnitude. Thus, **median** is a measure of the *location* or *centrality* of the observations.

The median can be calculated for both ungrouped and grouped data sets. The median is helpful in understanding the characteristic of a data set when

- Observations are qualitative in nature.
- Extreme values (outliers) are present in the data set.
- At a glance estimate of an average is desired.

Methods of Calculating Median

Ungrouped Data: Arrange elements in the data set in a sequential (either ascending or descending order) of magnitude.

(i) If the number of observations (n) are *odd number*, then the median (Med) value is

$$\text{Med} = \text{Size or value of } \left(\frac{n+1}{2}\right)\text{th observation in the data array.}$$

(ii) If the number of observations (n) are *even number*, then the median value is the arithmetic mean of the numerical values of (n/2)th and (n/2 + 1)th observations in the data array. That is,

$$\text{Med} = \frac{\frac{n}{2}\text{th observation} + \left(\frac{n}{2}+1\right)\text{th observation}}{2}$$

Example 11.20: Calculate the median of the following data that relates to the service time (in minutes) per customer for 7 customers at a railway reservation counter: 3.5, 4.5, 3, 3.8, 5.0, 5.5, 4.

Solution: The data are arranged in ascending order as follows:

Observations in the data array	:	1	2	3	4	5	6	7
Service time (in minutes)	:	3	3.5	3.8	4	4.5	5	5.5

Since number of observations are odd, the median for this data would be

$$\text{Med} = \text{value of } (n+1)/2\text{th observation in the data array}$$
$$= \{(7+1)/2\}\text{th} = 4\text{th observation in the data array} = 4$$

Thus, the median service time is 4 minutes per customer.

Example 11.21: Calculate the median of the following data that relates to the number of patients examined per hour in the outpatient ward (OPD) in a hospital: 10, 12, 15, 20, 13, 24, 17, 18.

Solution: The data are arranged in ascending order as follows:

Observations in the data array	:	1	2	3	4	5	6	7	8
Patients examined per hour	:	10	12	13	15	17	18	20	24

Since the number of observations are even, the average of $(n/2)$th = 4th observation, i.e. 15 and $(n/2) + 1$ = 5th observation, i.e. 17, will be the median, that is,

$$\text{Med} = (15 + 17)/2 = 16$$

Thus, median number of patients examined per hour in OPD in a hospital is 16.

Grouped Data: For grouped data, first find

(i) less than type cumulative frequency,

(ii) identify the median class interval or $(n/2)$th observation of the data set, i.e., cumulative frequency equal to or greater than the value of $(n/2)$th observation, and

(iii) apply the following formula to determine the median value:

$$\text{Med} = l + \frac{(n/2) - cf}{f} \times h$$

where l = lower class limit (or boundary) of the median class interval;

cf = cumulative frequency of the class prior to the median class interval, i.e., the sum of all the class frequencies up to, but not including, the median class interval;

f = frequency of the median class;

h = width of the median class interval; and

n = total number of observations in the distribution.

Remark To find median value by using interpolation, it is assumed that the numerical values of observations are evenly spaced over the entire class interval.

Example 11.22: A survey was conducted to determine the age (in years) of 120 automobiles. The result of such a survey is as follows:

Age of auto	:	0–4	4–8	8–12	12–16	16–20
Number of autos	:	13	29	48	22	8

What is the median age for the autos?

Solution: Calculations required to find median age of autos are shown in Table 11.14.

Table 11.14 *Calculations for Median Value*

Age of Auto (in years)	*Number of Autos (f)*	*Cumulative Frequency .(cf)*	
0–4	13	13	
4–8	29	42	
8–12	48	90	← Median class
12–16	22	112	
16–20	8	120	
	$n = 120$		

NOTES

The total number of observations (frequencies) in the data set are, $n = 120$. Median is the size of $(n/2)$th = 120/2 = 60th observation in the data set. This observation lies in the class interval 8–12. Applying the formula (11-16), we have

$$\text{Med} = l + \frac{(n/2) - cf}{n} \times h$$

$$= 8 + \frac{60 - 42}{48} \times 4 = 8 + 1.5 = 9.5$$

Example 11.23: In a factory employing 3000 persons, 5 per cent earn less than ₹150 per day, 580 earn from ₹151 to ₹200 per day, 30 per cent earn from ₹201 to ₹250 per day, 500 earn from ₹251 to ₹300 per day, 20 per cent earn from ₹301 to ₹350 per day and the rest earn ₹351 or more per day. What is the median wage?

Solution: Calculations for median wage per day are shown in Table 11.15.

Table 11.15 *Calculations of Median Wage*

Earnings (₹)	*Percentage of Workers* (*Per cent*)		*Number of Persons* (*f*)	*Cumulative Frequency* (*cf*)
Less than 150	5		150	150
151–200	—		580	730
201–250	30		900	1630 ← Median class
251–300	—	500	2130	
301–350	20	600	2730	
351 and above	—	270	3000	
		$n = 3000$		

Median is the size of $(n/2)$th = (3000)/2 = 1500th observation in the data set. This observation lies in the class interval 201 – 250. Applying the formula (11-16), we have

$$\text{Med} = l + \frac{(n/2) - cf.}{f} \times h$$

$$= 201 + \frac{1500 - 730}{900} \times 50 = 201 + 42.77 = ₹243.77$$

Hence, the median wage is ₹243.77 per day.

Advantages and Disadvantages of Median

Advantages

1. Median is unique, i.e. there is only one median for a set of data.
2. The value of median is easy to understand and may be calculated from any type of data.
3. The sum of the absolute difference of all observations from the median is less than from any other value in the distribution.

4. The extreme values in the data set do not affect the median value, and therefore it is the useful measure of central tendency when extreme values in the data set occur.
5. The median is useful to study the qualitative attribute of an observation in the data set.
6. The median value can also be calculated for open-ended class intervals in the data set.

Disadvantages

1. The median is not capable of algebraic operations. For example, pooled median of two populations cannot be determined.
2. The value of median is affected more by sampling variations, i.e., it is affected by the number of observations rather than the values of the observations.
3. Since median is an average of position, therefore arranging the data in ascending or descending order of magnitude is time consuming in case of a large number of observations.
4. The calculation of median in case of grouped data is based on the assumption that values of observations are evenly spaced over the entire class interval.

2. Mode

The **mode** is that value of an observation which occurs with highest frequency in the raw data or in classified data set.

The concept of mode is of great use to large-scale manufacturers of consumable items such as readymade garments, shoe makers, and so on. In all such cases, it is important to know the size that fits most consumers rather than 'mean' size.

The knowledge of arithmetic mean alone is not sufficient to understand the characteristic of numerical values of observations in the data set due to the presence of extreme values such as:

(i) 'average man prefers ... brand of cigarettes',

(ii) 'average production of an item in a month',

(iii) 'average service time at the service counter', etc.,

Also when observations in data set are uneven, median value may not represent the characteristics of numerical values of observations in data set. For example, in a distribution where values in the lower half vary from 10 to 100, while the same number of observations in the upper half varies from 100 to 7000 with most of them close to the higher value, the median value will not reflect true nature of the data. Such disadvantages of mean and median are taken care by using *mode*—third measure of central tendency.

However, mode is a poor measure of central tendency when more frequently occurring values of an observation do not appear close to the center of the data. Also, a frequency distribution may have more than one mode value. For example, frequency distribution shown in Fig. 11.1(a) has its mode at the lowest class and certainly cannot be considered representative of central location. The frequency distribution shown in Fig. 11.1(b) has two modes. Neither of these values appears to be representative of the central location of the data. For these reasons the mode has limited use as a measure of central tendency for decision making.

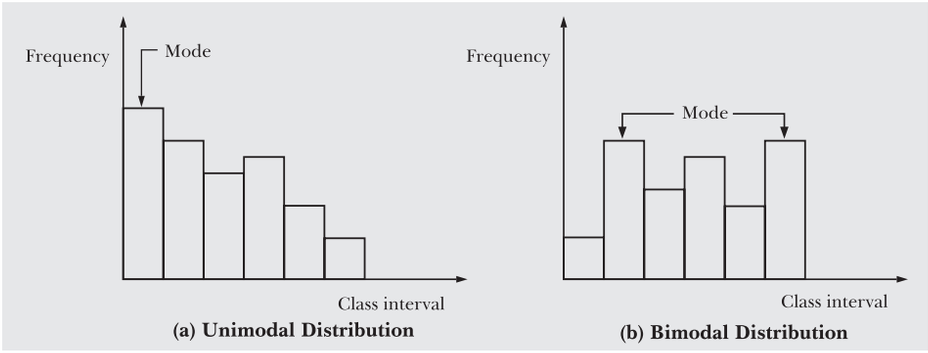


Fig. 11.1 *Frequency Distribution*

Illustration: Sales per day of an item for 20 days period is shown in Table 11.16. The mode value of this data is 71 since this value occurs more frequently (four times than any other value). However, it fails to reveal the fact that most of the values are under 70.

Table 11.16 *Sales During 20 Days Period*

(Data arranged in ascending order)

53,	56,	57,	58,	58,	60,	61,	63,	63,	64
64,	65,	65,	67,	68,	71,	71,	71,	71,	74

Converting this data into a frequency distribution as shown in Table 11.17:

Table 11.17 *Frequency Distribution of Sales Per Day*

Sales volume (*Class interval*) :	53–56	57–60	61–64	65–68	69–72	72 and above
Number of days (*Frequency*) :	2	4	5	4	4	1

Table 11.17 shows that a sale of 61–64 units of the item was achieved in 5 days. Thus, this class is more representative of the sales per day.

NOTES

In the case of grouped data, the following formula is used for calculating mode:

$$\text{Mode} = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h$$

where l = lower limit of the mode class interval;

f_{m-1} = frequency of the class preceding the mode class interval;

f_{m+1} = frequency of the class following the mode class interval; and

h = width of the mode class interval.

Example 11.24: The following are the data on sales per day of an item for 20 days period.

Sales volume (*Class interval*)	:	53–56	57–60	61–64	65–68	69–72	72 and above
Number of days (*Frequency*)	:	2	4	5	4	4	1

Calculate the mode of sales distribution of the units of item during the 20 days period.

Solution: Since the largest frequency corresponds to the class interval 61–64, therefore it is the mode class. Then we have, $l = 61, f_m = 5, f_{m-1} = 4, f_{m+1} = 4$ and $h = 3$. Applying the formula:

$$M_o = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h$$

$$= 61 + \frac{5-4}{10-4-4} \times 3 = 61 + 1.5 = 62.5$$

Hence, the modal sale is of 62.5 units.

Example 11.25: In 500 small-scale industrial units, the return on investment ranged from 0 to 30 per cent; no unit sustaining loss. Five per cent of the units had returns ranging from zero per cent to (and including) 5 per cent, and 15 per cent of the units earned returns exceeding 5 per cent but not exceeding 10 per cent. The median rate of return was 15 per cent and the upper quartile 2 per cent. The uppermost layer of returns exceeding 25 per cent was earned by 50 units.

(a) Present the information in the form of a frequency table as follows:

Exceeding 0 per cent but not exceeding 5 per cent

Exceeding 5 per cent but not exceeding 10 per cent

Exceeding 10 per cent but not exceeding 15 per cent and so on.

(b) Find the rate of return around which there is maximum concentration of units.

NOTES

Solution: (a) The given information is summarized in the form of a frequency distribution as shown in Table 11.18.

Table 11.18

Rate of Return	*Industrial Units*
Exceeding 0 per cent but not exceeding 5 per cent	$500 \times \frac{5}{100} = 25$
Exceeding 5 per cent but not exceeding 10 per cent	$500 \times \frac{15}{100} = 75$
Exceeding 10 per cent but not exceeding 15 per cent	250 – 100 = 150
Exceeding 15 per cent but not exceeding 20 per cent	375 – 250 = 125
Exceeding 20 per cent but not exceeding 25 per cent	500 – 375 – 50 = 75
Exceeding 25 per cent but not exceeding 30 per cent	50

(b) Calculating mode to find out the rate of return around which there is maximum concentration of the units. The mode lies in the class interval 10–15. Thus,

$$M_o = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}}$$

$$= 10 + \frac{150 - 75}{2 \times 150 - 75 - 125} \times 5 = 10 + 3.75 = 13.75 \text{ per cent}$$

Graphical Method for Calculating Mode Value

The procedure of calculating mode using the graphical method is summarized below:

- Draw a histogram of the data, the tallest rectangle will represent the modal class.
- Draw two diagonal lines from the top right corner and left corner of the tallest rectangle to the top right corner and left corner of the adjacent rectangles.
- Draw a perpendicular line from the point of intersection of the two diagonal lines on the *x*-axis. The value on the *x*-axis marked by the line will represent the modal value.

Example 11.26: Calculate mode value using the graphical method for the following distribution of data:

Sales (in units) :	53–56	57–60	61–64	65–68	69–72	73–76
Number of days :	2	4	5	4	4	1

Solution: Construct a histogram of the data shown in Fig. 11.2 and draw diagonal lines for the calculation of mode value. The mode value from Fig. 11.2 is 62.5.

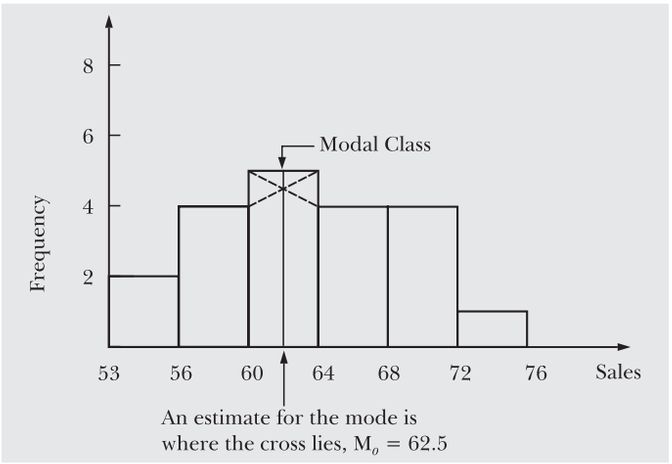


Fig. 11.2 *Graph for Modal Value*

Advantages and Disadvantages of Mode Value

Advantages

1. Mode value is easy to understand and calculate. Mode class can also be located by inspection.
2. The mode value is not affected by the extreme values in the distribution and can also be calculated for open-ended frequency distributions.
3. The mode value can be used to describe quantitative and/or qualitative characteristic of a variable in the data set.

Disadvantages

1. Mode is not a rigidly defined measure as there are several methods for calculating its value.
2. Locating a modal class in multi-modal frequency distributions is difficult.
3. Algebraic changes in the formula of mode are not possible.

11.3.3 Relationship between Mean, Median and Mode

If values of mean, median and mode are equal, then distribution of numerical values in the data set is symmetrical as shown in Fig. 11.3. But, if these three values are not equal, then distribution of numerical values in the data set is not symmetrical as shown in Figs. 11.3(b) and 11.3(c).

NOTES

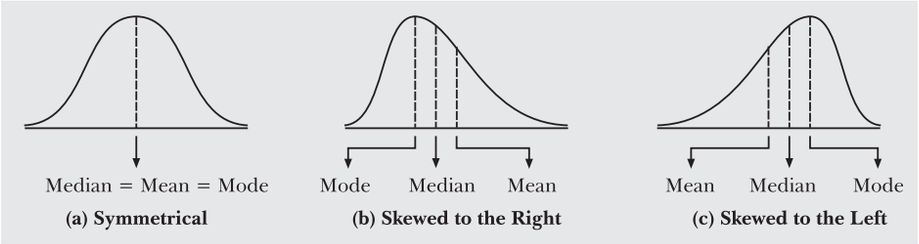


Fig. 11.3 *A Comparison of Mean, Median and Mode for Three Distributional Shapes*

If most of the values fall either to the right or to the left of the mode, then such a distribution is said to be *skewed*. In such cases, a relationship between these three measures of central tendency as suggested by Karl Pearson is as follows:

$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median}) \qquad (11\text{-}18)$$

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

If most of the values of observations in a distribution fall to the right of the mode as shown in Fig. 11.3(b), then it is said to be skewed to the right *or* **positively skewed** (i.e. values of higher magnitude are concentrated more to the right of the mode). In this case, mode remains under the peak (i.e., representing highest frequency) but the median (value that depends on the number of observations) and mean move to the right (value that is affected by extreme values). The order of magnitude of these measures will be

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

But if the distribution is skewed to the left or **negatively skewed** (i.e., values of lower magnitude are concentrated more to the left of the mode), then mode is again under the peak whereas median and mean move to the left of mode. The order of magnitude of these measures will be

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

In both these cases, the difference between mean and mode is three times the difference between mean and median.

In general, for a single mode skewed distribution (non-symmetrical), the median is preferred to the mean for measuring location because it is neither influenced by the frequency of occurrence of a single observation value as mode nor it is affected by extreme values.

Comparison between Measures of Central Tendency

The choice to use a method for describing a distribution of numerical values in a data set is mainly guided by its characteristics. The characteristics of these three measures of central tendency differ from each other with regard to three factors:

1. ***Presence of outlier data values:*** Certain values that are much higher/ smaller than other values in a data set are known as ***outliers***. Since median is not sensitive to outlier values because its value depend only on the number of observations and the value always lies in the middle of the ordered set of values, whereas mean which is calculated using all data values is sensitive to the outlier values in a data set. Obviously, less the number of observations in a data set, more the influence of outliers on the mean. This implies that median is not influenced by the presence of outlier data values but mean is.
2. ***Shape of frequency distribution:*** In general, the median is preferred to the mean for single peaked, skewed distributions to measure a characteristic in the data set. Because it satisfies the criterion that the *sum of absolute difference* of median from other values in the data set is minimum. But for multi-modal distributions neither mean, median nor mode will serve the purpose to locate central value and the mode can vary from one sample to another, particularly in case of small sample size.
3. ***Status of theoretical development:*** In *inferential statistics*, the objective of any statistical analysis is to minimize the *sum of squared deviations* (*errors, also called least squares criterion*) taken from mean, median or mode to every value in the data set. Since A.M. satisfies the least squares criterion, it is mathematically consistent with several techniques of statistical inference. Whereas median is only used for basic descriptive purposes.

Check Your Progress

3. What are the three types of averages? Name them.
4. What do understand by geometric mean?
5. Define harmonic mean.
6. What does median tell about data?
7. What does mode reperesent?

11.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The statistical methods that are used to extract and measure these three features: *central tendency, variation and skewness* in the data set are called descriptive measures. There are three types of descriptive measures:
   (i) Measures of central tendency

NOTES

(ii) Measures of dispersion or variation

(iii) Measures of symmetry—skewness

2. If the descriptive measures are computed using sample data, then these are called sample statistic but if these measures are computed using data of the population, they are called population parameters.
3. The three types of averages are mean, mode and median.
4. The geometric mean is a type of average and generally it is used for growth rates, like population growth or interest rates. While the arithmetic mean **adds** items, the geometric mean **multiplies** items.
5. The harmonic mean (H.M.) of a set of observations is defined as the reciprocal of the arithmetic mean and is calculated by taking reciprocal of the individual observations.
6. The median can be calculated for both ungrouped and grouped data sets. The median is helpful in understanding the characteristic of a data set when
   - Observations are qualitative in nature.
   - Extreme values (outliers) are present in the data set.
   - At a glance estimate of an average is desired.
7. The **mode** is that value of an observation which occurs with highest frequency in the raw data or in classified data set.

11.5 SUMMARY

- The statistical methods that are used to extract and measure these three features: *central tendency, variation and skewness* in the data set are called ***descriptive (or summary) measures***. There are three types of descriptive measures:

  (i) Measures of central tendency

  (ii) Measures of dispersion or variation

  (iii) Measures of symmetry—skewness
- If the descriptive measures are computed using sample data, then these are called sample statistic but if these measures are computed using data of the population, they are called population parameters.
- The **harmonic mean** (H.M.) of a set of observations is defined as the reciprocal of the arithmetic mean and is calculated by taking reciprocal of the individual observations.
- Median may be defined as the *middle value* (half of the observations are smaller and half are larger than this value) in the data set when

elements are arranged in a sequential (either ascending or descending) order of magnitude.

- The **mode** is that value of an observation which occurs with highest frequency in the raw data or in classified data set.
- Summation of n numbers

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_n$$

Simplified expression for the summation of n numbers

$$\sum x_i = x_1 + x_2 + ... + x_n$$

- Sample mean, $\bar{x} = \frac{\sum x_i}{n}$

Population mean, $\mu = \frac{\sum x_i}{N}$

Sample mean for grouped data, $\bar{x} = \frac{\sum f_i m_i}{n}$

where $n = \sum f_i$ and m_i = mid-value of class intervals

- Weighted mean for a population or a sample,

$$\bar{x}_w \text{ or } \mu_w = \frac{\sum w_i x_i}{\sum w_i}$$

where w_i = weight for observation i

- Position of the median in an ordered set of observation belong to a population or a sample is Med= $x_{(n/2)+(1/2)}$

Median for grouped data,

$$\text{Med} = l + \left[\frac{(n/2) - cf}{f}\right] h$$

- Mode for a grouped data

$$M_o = l + \left[\frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}}\right] h$$

Mode for a multimode frequency distribution

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

11.6 KEY WORDS

- **Central tendency:** It refers to the tendency for the values of a random variable to cluster round its mean, mode, or median.
- **Geometric mean:** It refers to the mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values.

- **Arithmetic mean:** It refers to the mean or average when the context is clear. It is the sum of a collection of numbers divided by the count of numbers in the collection.
- **Median:** It is a statistical term that is one way of finding the 'average' of a set of data points.
- **Mode:** It is a type of average that refers to the most-common or most-frequently occurring value in a series of data.

11.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. What are the requisites of the measures of central tendency?
2. What is the method of calculating median?
3. Briefly mention the advantages and disadvantages of median.
4. Mention the advantages and disadvantages of mode value.
5. The following is the data on profit margin (in per cent) of three products and their corresponding sales (in ₹) during a particular period.

Product	*Profit Margin* (*Per cent*)	*Sales* (₹*1000*)
A	12.5	2,000
B	10.3	6,000
C	6.4	10,000

   (a) Determine the mean profit margin.

   (b) Determine the weighted mean considering the rupee sales as weight for each product.

   (c) Which of the means calculated in part (a) and (b) is the correct one?

6. The number of cars sold by each of the 10 car dealers during a particular month, arranged in ascending order, is 12, 14, 17, 20, 20, 20, 22, 22, 24, 25. Considering this scale to be the statistical population of interest, determine the mean, median, and mode for the number of cars sold.

   (a) Which value calculated above best describes the 'typical' sales volume per dealer?

   (b) For the given data, determine the values at the (i) quartile Q_1 and (ii) percentile P30 for these sales amounts.

7. Calculate the mean, median, and mode for the following data pertaining to marks in statistics. There are 80 students in a class and the test is of 140 marks.

Marks more than :	0	20	40	60	80	100	120
Number of students :	80	76	50	28	18	9	3

8. A company invests one lakh rupees at 10 per cent annual rate of interest. What will be the total amount after 6 years if the principal is not withdrawn?

Long Answer Questions

1. Discuss the advantages and disadvantages of arithmetic mean.
2. Discuss the merits, demerits and applications of geometric mean.
3. Analyse the relationship between mean, median and mode.
4. A quality control inspector tested nine samples of each of three designs A, B and C of certain bearing for a new electrical winch. The following data are the number of hours it took for each bearing to fail when the winch motor was run continuously at maximum output, with a load on the winch equivalent to 1.9 times the intended capacity.

A	:	16	16	53	15	31	17	14	30	20
B	:	18	27	23	21	22	26	39	17	28
C	:	31	16	42	20	18	17	16	15	19

   Calculate the mean and median for each group and suggest which design is best and why?

5. In the production of light bulbs, many bulbs are broken. A production manager is testing a new type of conveyor system in the hope of reducing the percentage of bulbs broken each day. For ten days he observes bulb breakage with the current conveyor. He then records bulb breakage for ten days with the new system, after allowing a few days for the operator to learn to use it. His data are as follows:

Conveyor System	*Percentage of Bulbs Broken Daily*									
Old :	8.7	11.1	4.4	3.7	9.2	6.6	7.8	4.9	6.9	8.3
New :	10.8	6.2	3.2	4.6	5.3	6.5	4.6	7.1	4.9	7.2

   (a) Compute the mean and median for each conveyor system.

   (b) Based on these results, do you think this test establishes that the new system lowers the breakage rate? Explain.

6. The following are the weekly wages in rupees of 30 workers of a firm:

140	139	126	114	100	88	62	77	99
103	108	129	144	148	134	63	69	148
132	118	142	116	123	104	95	80	85
106	123	133						

NOTES

The firm gave bonus of ₹10, 15, 20, 25, 30, and 35 for individuals in the respective salary slabs: exceeding 60 but not exceeding 75; exceeding 75 but not exceeding 90; and so on up to exceeding 135 and not exceeding 150. Find the average bonus paid.

7. The mean monthly salaries paid to 100 employees of a company was ₹5,000. The mean monthly salaries paid to male and female employees were ₹5,200 and ₹4,200 respectively. Determine the percentage of males and females employed by the company.

11.8 FURTHER READINGS

Creswell, John W. 2002. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* London: Sage Publications.

Booth, Wayne, Gregory G. Colomb and Joseph M. Williams. 1995. *The Craft of Research.* Chicago: University of Chicago Press.

Kumar, B. 2006. *Research Methodology.* New Delhi: Excel Books.

Paneerselvam, R. 2009. *Research Methodology.* New Delhi: Prentice Hall of India.

Gupta, D. 2011. *Research Methodology*. New Delhi: PHI Learning Private Limited.

UNIT 12 MEASURES OF DISPERSION-I

NOTES

Structure

12.0 INTRODUCTION

The measures of central tendency describe that the values in the data set tend to spread (cluster) around a central value called *average*. But these measures do not reveal how these values are spread (dispersed or scattered) on each side of the central value. Just as central tendency can be measured by a number in the form of an average, the amount of variation (dispersion, spread or scatter) among the values in the data set can also be measured. The dispersion of values is indicated by the extent to which these values tend to spread over an interval rather than cluster closely around an average.

The statistical techniques to measure the extent to which values in the data set tend to spread are of two types:

(i) Techniques that are used to measure the extent of variation or deviation of each value in the data set from a measure of central tendency, usually the mean or median. Such statistical techniques are called *measures of dispersion* (or *variation*).

(ii) Techniques that are used to measure the direction (away from uniformity or symmetry) of variation in the distribution of values in the data set. Such statistical techniques are called *measures of skewness*.

NOTES

In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed. Common examples of measures of statistical dispersion are the variance, standard deviation, and interquartile range.

Dispersion is contrasted with location or central tendency, and together they are the most used properties of distributions.

12.1 OBJECTIVES

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

- Discuss the importance of the concept of variability (dispersion)
- Measure the spread of dispersion, understand it, and identify its causes to provide a basis for action
- Describe Quartile, Mean and Standard Deviation

12.2 UNDERSTANDING DISPERSION

Identifying the causes and then measuring the dispersion is useful to draw statistical inference (estimation of parameter, hypothesis testing, forecasting and so on). A small dispersion among values in the data set indicates that values in the data set are clustered closely around the mean, implying that the mean is a reliable average. Conversely, if values in the data set are widely clustered around the mean, then this implies that the mean is not a reliable average, i.e. mean is not representative of the data.

The symmetrical distribution of values in two or more data sets may have same variation but differ in terms of A.M. as shown in Fig 12.1. On the other hand, two or more data sets may have the same A.M. values but differ in variation as shown in Fig. 12.2.

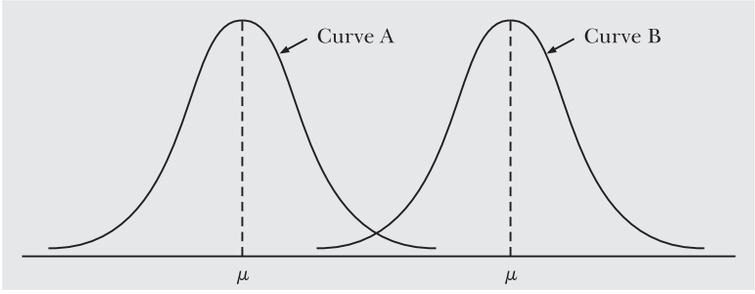


Fig. 12.1 *Symmetrical Distributions with Unequal Mean and Equal Standard Deviation*

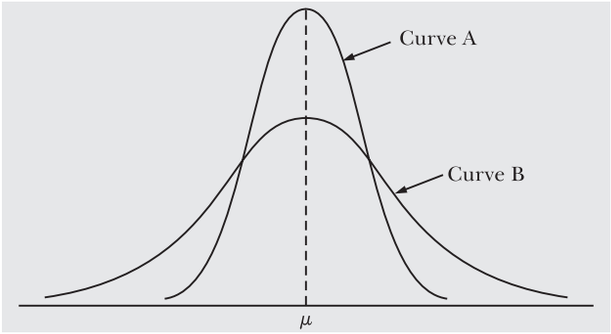


Fig. 12.2 *Symmetrical Distributions with Equal Mean and Unequal Standard Deviation*

Illustration: Suppose over the six-year period the net profits (in percentage) of two firms are as follows:

Firm 1	:	5.2,	4.5,	3.9,	4.7,	5.1,	5.4
Firm 2	:	7.8,	7.1,	5.3,	14.3,	11.0,	16.1

Since average amount of profit is 4.8 per cent for both firms, therefore both the firms are equally good and that a choice for investment purposes must depend on other considerations. However, in case of Firm 2, net profit values are varying from 5.3 to 16.1 per cent, i.e., difference among the values is more while the net profit values of Firm 1 are varying from 3.9 to 5.4 per cent, i.e., difference among the values is less as compared to Firm 2. In other words, net profit values in data set 2 are spread more than those in data set 1. This implies that the performance of Firm 1 is consistent as compared to Firm 2. Consequently, for investment, a comparison of the average (mean) profit values alone should not be sufficient.

12.2.1 Significance of Measuring Dispersion

The following are some of the purposes for which measures of variation are needed.

1. ***Test the reliability of an average:*** Measures of variation help to understand the extent an average represents the characteristic of a data set. If the extent of dispersion of values is less on each side of an average value, then it indicates high uniformity among values in the distribution. On the other hand, if the variation is large, then it indicates a lower degree of uniformity among values in the data set, and the average value may be unreliable.
2. ***Control the variability:*** Measuring variation helps to identify the nature and causes of variation. Such information is useful in controlling the variations. According to Spurr and Bonini, *In matters of health, variations in, body temperature, pulse beat and blood pressure are the basic guides to diagnosis. Prescribed treatment is designed to*

NOTES

control their variation. In industrial production, efficient operation requires control of quality variation, the causes of which are sought through inspection and quality control programmes. In social science, the measurement of 'inequality' of distribution of income and wealth requires the measurement of variability.

3. ***Compare two or more sets of data with respect to their variability:*** Measures of variation help in comparing variation in two or more sets of data with respect to their uniformity or consistency. For example, (i) measurement of variation in share prices and their comparison with respect to different companies over a period of time, and (ii) measurement of variation in the length of stay of patients in a hospital helps to set staffing levels, number of beds, number of doctors, and other trained staff, patient admission rates and so on.
4. ***Facilitate the use of other statistical techniques:*** Measures of variation facilitate the use of other statistical techniques such as correlation and regression analysis, hypothesis testing, forecasting, quality control and so on.

12.2.2 Requisites for a Good Measure of Variation

Certain essential requisites that help in identifying the merits and demerits of individual measure of variation are as follows:

(i) Should be based on all the values (elements) in the data set.

(ii) Should be calculated easily, quickly and accurately.

(iii) Should be unaffected by the fluctuations in sample size and also by outliers.

(iv) Should be further mathematical or algebraic changes are possible.

Check Your Progress

1. What are the requisites for a good measure of variation?
2. What does a small dispersion among values in data sets indicate?

12.3 CLASSIFICATION OF MEASURES OF DISPERSION: QUARTILE, MEAN AND STANDARD DEVIATION

Measures of dispersion (or variation) based on the purpose of measuring are classified into two categories:

1. **Absolute measures:** These measures are described by a number (or value) to represent the amount of variation (or difference) among values

in a data set. Such a value is expressed in the same unit of measurement such as rupee, inch, foot, kilogram, ton, etc., as the values in the data set. Such measures help in comparing two or more sets of data in terms of absolute magnitude of variation, provided variable values are expressed in the same unit of measurement and have almost the same average value.

2. **Relative measures:** These measures are described as the ratio of a measure of absolute variation to an average and is termed as *coefficient of variation*. The word 'coefficient' means a number that is independent of any unit of measurement. While computing the relative variation, the average value used as base should be the same from which the absolute deviations were calculated.

Another classification of the measures of variation is based on the method used for their calculations:

(i) Distance measures

(ii) Average deviation measures

The **distance measures** describe the spread or dispersion of values of a variable in terms of difference among values in the data set. The **average deviation measures** describe the average deviation for a given measure of central tendency.

The classification of various measures of dispersion (variation) is shown in Fig. 12.3.

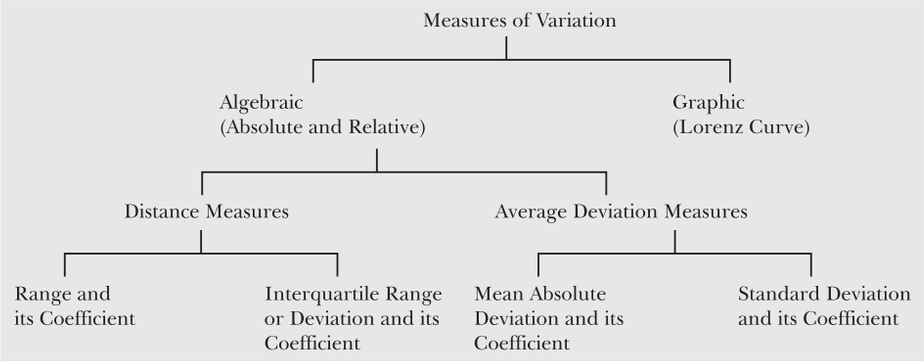


Fig. 12.3 *Classification of Measures of Variation*

12.3.1 Interquartile Range or Deviation

The limitations or disadvantages of range can partially be overcome by using another measure of variation called **Interquartile Range or Deviation (IQR)**. The IQR measures the spread within middle half of the values in the data set so as to minimize the influence of outliers (extreme values) in the calculation

NOTES

of range. Since a large number of values in the data set lie in the central part of the frequency distribution, it is necessary to study the **Interquartile Range** (also called mid-spread).

To compute IQR, data set is divided into four parts each of which contains 25 per cent of the observed values. Thus, *interquartile range* is *a measure of dispersion or spread of values in the data set between the third quartile, Q_3 and the first quartile*, Q_1. In other words, the *interquartile range or deviation* is the range for the middle 50 per cent of the data set. The concept of IQR is shown in Fig. 12.4:

Interquartile range (IQR) = $Q_3 - Q_1$

Half the distance between Q_1 and Q_3 is called the *semi-interquartile range* or the *quartile deviation* (QD).

$$\text{Quartile deviation (QD)} = \frac{Q_3 - Q_1}{2}$$

The median is not necessarily midway between Q_1 and Q_3, although it is true for a symmetrical distribution. The median and quartiles divide the data set into equal numbers of values but do not necessarily divide the data into equally wide intervals.

The ***quartile deviation (QD)***) measures the average range of 25 per cent of the values in the data set. It is computed by taking an average of the middle 50 per cent of the observed values rather than 25 per cent part of the values in the data set.

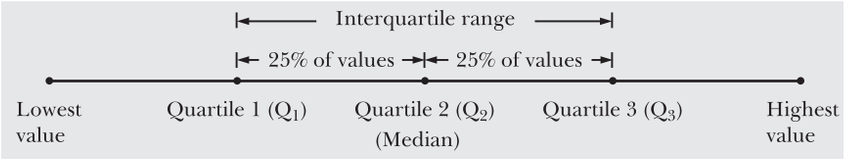


Fig. 12.4 *Interquartile Range*

In a symmetrical distribution, the two quartiles Q_1 and Q_3 are at equal distance from the median, i.e., Median – Q_1 = Q_3 – Median. Thus, *Median ± Quartile Deviation* covers exactly 50 per cent of the observed values in the data set.

A smaller value of quartile deviation indicates high uniformity (or less variation) among the middle 50 per cent values around the median value. On the other hand, a high value of quartile deviation indicates large variation among the middle 50 per cent values.

The median and quartiles divide the data set into equal parts of values but not necessarily into equally wide intervals.

NOTES

Coefficient of Quartile Deviation

Since quartile deviation is an absolute measure of variation, therefore its value gets affected by the size and number of observed values in the data set. Thus, Q.D. of two or more than two data sets may differ. Due to this reason, to compare the degree of variation in different data sets, we compute the relative measure corresponding to Q.D., called the coefficient of Q.D. as follows:

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

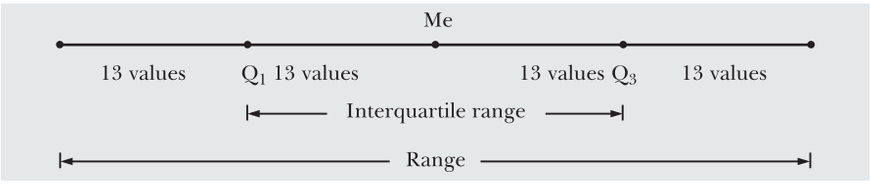
Example 12.1: Following are the responses from 55 students to the question about how much money they spent every day.

55	60	80	80	80	85	85	85	90	90	90
90	92	94	95	95	95	95	100	100	100	100
100	100	105	105	105	105	109	110	110	110	110
112	115	115	115	115	115	120	120	120	120	120
124	125	125	125	130	130	140	140	140	145	150

Calculate the range and interquartile range and interpret your result.

Solution: Since number of responses are 55 — an odd numbe**r,** therefore median of the given values in the data set is: (55+1)/2 = 28th value which is 105. This means there are 27 values at or below 105 and another 27 at or above 105.

The lower quartile Q_1 = (27+1)/2 = 14th value from bottom of the data, i.e. Q_1 = 94 and upper quartile is the 14th value from the top, i.e. Q_3 = 120. The 55 values have been partitioned as follows:



The interquartile range (IQR) is: 120 – 94 = ₹26, while the range is, R = 150 – 55 = ₹95. The middle 50 per cent values of data set fall in a narrow range of only ₹26. This means responses are densely clustered near the centre of the data and more spread towards the extremes. For instance, lowest 25 per cent of the students had responses in the interval 55 to 94, i.e. ₹39, while the next 25 per cent had responses in the interval 94 to 105, i.e. ₹11. Similarly, the third quarter had responses in the interval 105 to 110, i.e. ₹5, while the top 25 per cent had responses in the interval 120 to 150, i.e. ₹30.

NOTES

Example 12.2: Use an appropriate measure to evaluate the variation in the following data:

Farm Size (*acre*)	*No. of Farms*	*Farm Size* (*acre*)	*No. of Farms*
below 40	394	161–200	169
41–80	461	201–240	113
81–120	391	241 and above	148
121–160	334		

Solution: Since first and last intervals in the frequency distribution are open-end class intervals, Q.D. is an appropriate measure to evaluate variation. The computation of Q.D. is shown in Table 12.1.

Table 12.1 *Calculations of Quartile Deviation*

Farm Size (*acre*)	*No. of Farms*	*Cumulative Frequency* (*cf*) (*less than*)	
below 40	394	394	
41–80	461	855	← Q_1 class
81–120	391	1246	
121–160	334	1580	← Q_3 class
161–200	169	1749	
201–240	113	1862	
241 and above	148	2010	
	2010		

Q_1= Value of (n/4)th observation = 2010/ 4 or 502.5th observation

This observation lies in the class interval 41–80. Therefore,

$$Q_1 = l + \frac{(n/4) - cf}{f} \times h$$

$$= 41 + \frac{502.5 - 394}{461} \times 40 = 41 + 9.41 = 50.41 \text{ acres}$$

Q_3= Value of (3n/4)th observation = (3 × 2010)/4 or 1507.5th observation

This observation lies in the class interval 121–160. Therefore,

$$Q_3 = l + \frac{(3n/4) - cf}{f} \times h$$

$$= 121 + \frac{1507.5 - 1246}{334} \times 40 = 121 + 31.31 = 152.31 \text{ acres}$$

Thus, the quartile deviation is given by

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{152.31 - 50.41}{2} = 50.95 \text{ acres}$$

and Coefficient of Q.D. $= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{50.95}{202.72} = 0.251$

NOTES

Advantages and Disadvantages of Quartile Deviation

Advantages

1. Easy to calculate but useful only to evaluate variation among observed values within the middle of the data set.
2. Value is not affected by the extreme (highest and lowest) values in the data set.
3. An appropriate measure of variation for a data set having open-ended class intervals.
4. Since it is a positional measure of variation, it is useful in case of highly skewed distributions, where other measures of variation get affected by extreme values in the data set.

Disadvantages

1. Instead of all observations, the value of Q.D. is based on the middle 50 per cent values in the data set, it cannot be considered as a good measure of variation.
2. The value of Q.D. is not affected by the distribution of the individual values within the interval of the middle 50 per cent values in the data set.

12.3.2 Average (Mean) Deviation measures

Since two measures of variation, range and quartile deviation discussed earlier do not indicate how values in a data set are scattered around a central value or disperse throughout the range, therefore it is important to measure the amount (degree) by which these values in a data set deviate from a measure of central value—usually mean or median.

To understand the nature of spread of values in the data set, two more measures of dispersion that are useful to measure the average deviation from a measure of central value—usually mean or median are:

(i) Mean Absolute Deviation or Average Deviation

(ii) Variance and Standard Deviation

1. Mean Absolute Deviation

Since average (mean) deviation of individual values in the data set from their actual arithmetic mean (A.M.) is always zero, therefore such a measure would not indicate any variation. This problem can be solved in two ways:

(i) Ignore the signs of the deviation by taking absolute value.

(ii) Square the deviations because the square of a negative number is positive.

NOTES

The absolute difference between a value x_i of an observation from A.M. (or median) is always a positive number. The average value of these deviations from the A.M. (or median) is called the ***mean absolute deviation*** **(MAD)**. The MAD value is used to compare relative tendency of values in the distribution to scatter around a central value or to disperse throughout the range.

In general, the mean absolute deviation is given by

$$\text{MAD} = \frac{1}{\text{N}}\sum_{i=1}^{\text{N}} |\, x - \mu \,|\text{, for a population}$$

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n} |\, x - \bar{x} \,|\text{, for a sample}$$

where || is the sign of absolute value. That is, the plus or minus sign of deviations from the mean are ignored.

For a grouped frequency distribution, MAD is given by

$$\text{MAD} = \frac{\sum_{i=1}^{n} f_i \,|\, x_i - \bar{x} \,|}{\sum f_i}$$

While calculating MAD, the median is also considered for computing mean absolute deviation because sum of the absolute values of deviations from the median is smaller than that from any other value. However, in general, arithmetic mean is used for this purpose.

If a frequency distribution is symmetrical, then MAD taken from either mean or median is equal. Thus, the interval $\bar{x} \pm$ MAD provides a range in which 57.5 per cent of the observations are included. Even if the frequency distribution is moderately skewed, the interval $\bar{x} \pm$ MAD includes the same percentage of observations. This shows that more than half of the observations are scattered within one unit of the MAD around the arithmetic mean. The MAD is useful in situations where extreme deviations are likely to occur.

Coefficient of MAD The relative measure of MAD is called the *coefficient of MAD* and is obtained by dividing the MAD by a measure of central tendency (arithmetic mean or median) used for calculating the MAD. Thus,

$$\text{Coefficient of MAD} = \frac{\text{Mean absolute deviation}}{\bar{x} \text{ or Me}}$$

If the value of relative measure is desired in percentage, then

$$\text{Coefficient of MAD} = \frac{\text{MAD}}{\bar{x} \text{ or Me}} \times 100$$

Example 12.3: The number of patients seen in the emergency ward of a hospital for a sample of 5 days in the last month was 153, 147, 151, 156 and 153. Determine the mean absolute deviation and interpret.

NOTES

Solution: The mean number of patients is $\bar{x} = (153 + 147 + 151 + 156 + 153)/5 = 152$. The calculations of MAD using formula (4-6) are shown below.

Numer of Patients (x)	$x - \bar{x}$	*Absolute Deviation* $\lvert x - \bar{x} \rvert$
153	153 – 152 = 1	1
147	147 – 152 = – 5	5
151	151 – 152 = – 1	1
156	156 – 152 = 4	4
153	153 – 152 = 1	1
		12

$$\text{MAD} = \frac{1}{n}\sum|x - \bar{x}| = \frac{12}{5} = 2.4 \cong 3 \text{ patients (approx)}$$

The mean absolute deviation is 3 patients per day. The deviation in the number of patients falls in the interval 152 ± 3 patients per day.

Example 12.4: Calculate the mean absolute deviation and its coefficient from median for the following data

Year	*Sales (₹ thousand)*	
	Product A	*Product B*
2006	23	36
2007	41	39
2008	29	36
2009	53	31
2010	38	47

Solution: The median sales of the two products A and B is 38 and 36, respectively. The calculations of MAD in both the cases are shown in Table 12.2.

Table 12.2 *Calculations of MAD*

Product A		*Product B*	
Sales (x)	$\lvert x - Me \rvert = \lvert x - 38 \rvert$	*Sales (x)*	$\lvert x - Me \rvert = \lvert x - 36 \rvert$
23	15	31	5
29	9	36	0
38	0	36	0
41	3	39	3
53	15	47	11
$n = 5$	$\Sigma\lvert x - \text{Me} \rvert = 42$	$n = 5$	$\Sigma\lvert x - \text{Me} \rvert = 19$

Product A: $\text{MAD} = \frac{1}{n}\sum|x - \text{Me}| = \frac{42}{5} = 8.4$

Coefficient of $\text{MAD} = \frac{\text{MAD}}{\text{Me}} = \frac{8.4}{38} = 0.221$

NOTES

Product B: $\text{MAD} = \frac{1}{n}\sum |x - \text{Me}| = \frac{19}{5} = 3.8$

$$\text{Coefficient of MAD} = \frac{\text{MAD}}{\text{Me}} = \frac{3.8}{36} = 0.106$$

Example 12.5: Find the mean absolute deviation from mean for the following frequency distribution of sales (₹ in thousand) in a co-operative store.

Sales	:	50–100	100–150	150–200	200–250	250–300	300–350
Number of days	:	11	23	44	19	8	7

Solution: The mean absolute deviation can be calculated by using the formula (4-6) for A.M. ($\bar{x}$). The calculations for MAD are shown in Table 12.3. Let, assumed mean, A = 175.

Table 12.3 *Calculations for MAD*

Sales (₹)	*Mid-Value* (*m*)	*Frequency* (*f*)	$d = (m-175)/50$	*fd*	$\lvert x - \bar{x} \rvert = \lvert m - \bar{x} \rvert$	$f\lvert x - \bar{x} \rvert$
50 – 100	75	11	–2	–22	104.91	1154.01
100 – 150	125	23	–1	–23	54.91	1262.93
150 – 200	175 ← A	44	0	0	4.91	216.04
200 – 250	225	19	1	19	45.09	856.71
250 – 300	275	8	2	16	95.09	760.72
300 – 350	325	7	3	21	145.09	1015.63
		112		11		5266.04

$$\bar{x} = A + \frac{1}{n}\sum fd \times h = 175 + \frac{11}{112} \times 50 = ₹179.91 \text{ per day}$$

$$\text{MAD} = \frac{1}{n}\sum f\,|x - \bar{x}| = \frac{5266.04}{112} = ₹47.01$$

Thus, the average sales are ₹1,79,910 per day and the mean absolute deviation of sales is ₹47,010 per day.

Advantages and Disadvantages of MAD

Advantages

1. The calculation of MAD is based on all observations in the distribution and shows the dispersion of values around the measure of central tendency.
2. While calculating MAD, equal weightage is given to each observed value to indicate how far each observation lies from either the mean or median.
3. Average deviation from arithmetic mean is always zero in any data set. In MAD this problem is taken care by using absolute values to eliminate the negative signs.

NOTES

Disadvantages

1. While calculating MAD, the algebraic signs are ignored. If the signs are not ignored, then sum of the deviations taken from arithmetic mean will be zero and close to zero when deviations are taken from median.
2. The value of MAD is considered to be best when deviations are taken from median. But median does not provide a satisfactory result in case the amount of variation is more in a data set.

Variance and Standard Deviation

While computing absolute value of each deviation from arithmetic mean, another way to ignore sign of negative deviations from mean is to square such values. The sum of all such squared deviations is then divided by the number of observations in the data set. A value so obtained is called **population variance** denoted by σ^2 (a lower-case Greek letter sigma). It is usually referred to as 'sigma squared'. Symbolically, it is written as

Population variance,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - (\mu)^2 \quad \text{(Deviation is taken from actual population A.M.)}$$

$$= \frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2 \quad \text{(Deviation is taken from assumed mean, A)}$$

where $d = x - A$ and A is any constant (also called assumed A.M.)

Since σ^2 is the average (or mean) of squared deviations from arithmetic mean, it is also called the ***mean square average.***

The population variance is used to measure variation among the values of observations in a population. However, in almost all applications of statistics, the data being analysed is a sample data. Thus, sample variance is determined to estimate population variance, σ^2.

If the *sum of the squared deviations* about a sample mean $\bar{x}$ in previous Eq. is divided by n (sample size), then invariably the estimated value of σ^2 is lower than its actual value. Such a difference in two values is called *bias*. However, this *bias* in the estimation of population variance from a sample variance can be removed by dividing the sum of the squared deviations between the sample mean and each value in the population by $n - 1$ rather than by n. The *unbiased sample variance* denoted by s^2 is defined as follows:

Sample variance, $s^2 = \dfrac{\sum(x-\bar{x})^2}{n-1} = \dfrac{\sum x^2}{n-1} - \dfrac{n\bar{x}^2}{n-1} = \dfrac{\Sigma x^2}{n-1} - \dfrac{(\Sigma x)^2}{n(n-1)}$.

NOTES

The numerator $\sum(x-\bar{x})^2$ in Eq. (4-10) is called the *total sum of squares*. This quantity measures the total variation among values in a data set (whereas the variance measures only the *average variation*). The larger the value of $\Sigma(x-\bar{x})^2$, the greater the variation among the values in a data set.

12.3.3 Standard Deviation

The numerical value of population or a sample variance is difficult to interpret because it is expressed in square units. To reach an interpretable measure of variance expressed in the units of original data, we take a positive square root of the variance, called ***standard deviation or root-mean square deviation***. The standard deviation of population and sample is denoted by σ and s, respectively.

(a) *Ungrouped Data*

Population standard deviation, $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum(x-\mu)^2} = \sqrt{\frac{1}{N}\sum x^2 - (\mu)^2}$

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

Sample standard deviation, $s = \sqrt{\frac{\Sigma x^2}{n-1} - \frac{n\bar{x}^2}{n-1}}$; where n = sample size

(b) *Grouped Data*

Population standard deviation, $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h$

where f is the frequency of each class interval; N is the total number of observations (or elements) in the population; h is the width of class interval; m is mid-value of each class interval and $d = (m - A) / h$, where A is any constant (also called assumed A.M.)

Sample standard deviation,

$$s = \sqrt{s^2} = \sqrt{\frac{\sum f(x-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum fx^2}{n-1} - \frac{(\sum fx)^2}{n(n-1)}}$$

Remarks: 1. For any data set, MAD is always less than the σ because MAD is less sensitive to the extreme observations. Thus, when a data contains few outliers, the MAD provides a more realistic measure of variation than σ. However, σ is often used in statistical applications because formula is capable of algebraic treatment.

2. When sample size (n) becomes very large, (n – 1) becomes irrelevant.

Advantages and Disadvantages of Standard Deviation

Advantages

1. The value of standard deviation is based on every observation in a set of data. The formula of standard deviation is capable of algebraic treatment and is less affected by fluctuations of sample size as compared to other measures of variation.
2. It is possible to calculate the combined standard deviation of two or more sets of data.
3. The area under the symmetric curve of a frequency distribution is expressed in terms of standard deviation and population mean.
4. Standard deviation is used for comparing skewness, correlation, and so on, and also widely used in sampling theory.

Disadvantages

1. Calculations of standard deviation are slightly difficult as compared to other measures of variation.
2. Since for calculating S.D., the deviations from the arithmetic mean are squared, therefore large deviations when squared are proportionately more than small deviations. For example, the deviations 2 and 10 are in the ratio of 1 : 5 but their squares 4 and 100 are in the ratio of 1 : 25.

Example 12.6: The wholesale prices of a commodity for seven consecutive days in a month are as follows:

Days	:	1	2	3	4	5	6	7
Commodity price/quintal	:	240	260	270	245	255	286	264

Calculate the variance and standard deviation.

Solution: The computations for variance and standard deviation from actual arithmetic mean, $\bar{x}$ are shown in Table 12.4.

Table 12.4 *Computations of Variance and Standard Deviation with Actual Mean*

Observation (x)	$x - \bar{x} = x - 260$	$(x - \bar{x})^2$
240	–20	400
260	0	0
270	10	100
245	–15	225
255	–5	25
286	26	676
264	4	16
1820		1442

NOTES

NOTES

$$\bar{x} = \frac{1}{n}\sum x = \frac{1}{7}(1820) = 260$$

$$\text{Variance } \sigma^2 = \frac{1}{n}\sum (x - \bar{x})^2 = \frac{1}{7}(1442) = 206$$

$$\text{Standard deviation } \sigma = \sqrt{\sigma^2} = \sqrt{206} = 14.352$$

If deviation is taken from an assumed A.M. = 255 instead of actual A.M. = 260, then calculations for standard deviation are shown in Table 12.5.

Table 12.5 *Computations of Standard Deviation with Assumed Mean*

Observation (x)	$d = x - A = x - 255$	d^2
240	–15	225
260	5	25
270	15	225
245	–10	100
255 ← A	0	0
286	31	961
264	9	81
	35	1617

$$\text{Standard deviation } \sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{1617}{7} - \left(\frac{35}{7}\right)^2}$$

$$= \sqrt{231 - 25} = \sqrt{206} = 14.352$$

This result is same as shown in Table 12.4.

Remark: When actual A.M. is not a whole number, assumed A.M. method should be used to reduce the computation time.

Example 12.7: A study of 100 engineering companies gives the following information

Profit (₹ in crore)	:	0–10	10–20	20–30	30–40	40–50	50–60
Number of companies	:	8	12	20	30	20	10

Calculate the standard deviation of the profit earned.

Solution: Let assumed mean, A be 35. Calculations for standard deviation are shown in Table 12.6.

Table 12.6 *Calculations of Standard Deviation*

Profit (₹ *in crore*)	*Mid-value* (m)	$d = \frac{m-A}{h} = \frac{m-35}{10}$	*Number of Companies* (f)	fd	fd^2
0–10	5	–3	8	–24	72
10–20	15	–2	12	–24	48
20–30	25	–1	20	–20	20
30–40	35 ← A	0	30	0	0
40–50	45	1	20	20	20
50–60	55	2	10	20	40
			100	–28	200

NOTES

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times h$$

$$= \sqrt{\frac{200}{100} - \left(\frac{-28}{100}\right)^2} \times 10 = \sqrt{2 - 0.078} \times 10 = 13.863$$

Example 12.8: Mr Gupta, a retired government servant, is considering investing his money in two proposals. He wants to choose the one that has higher average net present value and lower standard deviation. The relevant data are given below. Can you help him in choosing the proposal?

Proposal A:	Net Present Value (NPV)	Chance of the Possible Outcome of NPV
	1559	0.30
	5662	0.40
	9175	0.30

Proposal B:	Net Present Value (NPV)	Chance of the Possible Outcome of NPV
	–10,050	0.30
	5,812	0.40
	20,584	0.30

Solution: The expected (average) net present value for both the proposals is:

Proposal A: Expected NPV = 1559 × 0.30 + 5662 × 0.40 + 9175 × 0.30

= 467.7 + 2264.8 + 2752.5 = ₹5485

Proposal B: Expected NPV = – 10,050 × 0.30 + 5812 × 0.40 + 20,584 × 0.30

= – 3015 + 2324.8 + 6175.2 = ₹5485

Since the expected NPV in both the cases is same, Mr. Gupta would like to choose less risky proposal. For this, we have to calculate the standard deviation in both the cases.

Standard deviation for proposal A:

NPV (x_i)	*Expected NPV* ($\bar{x}$)	$x - \bar{x}$	*Probability of NPV* (f)	$f(x - \bar{x})^2$
1559	5485	–3926	0.30	46,24,042.8
5662	5485	177	0.40	12,531.6
9175	5485	3690	0.30	40,84,830.0
			1.00	87,21,404.4

$$S_A = \sqrt{\frac{\sum f(x-x)^2}{N}} = \sqrt{87,21,404.4} = ₹2953.20$$

NOTES

Standard deviation for proposal B:

$NPV(x_i)$	Expected $NPV(\bar{x})$	$x - \bar{x}$	Probability of $NPV(f)$	$f(x - \bar{x})^2$
–10,050	5485	–15,535	0.30	7,24,00,867.5
5812	5485	327	0.40	42,771.6
20,584	5485	15,099	0.30	6,83,93,940
			1.00	14,08,37,579

$$s_B = \sqrt{\frac{\sum f(x-\bar{x})^2}{N}} = \sqrt{14,08,37,579} = ₹11,867.50$$

Since $s_A < s_B$, therefore proposal A indicates uniform net profit and hence may be chosen.

Mathematical Properties of Standard Deviation

1. ***Combined standard deviation*:** The combined standard deviation, σ_{12} of two sets of data containing n_1 and n_2 observations with means $\bar{x}_1$ and $\bar{x}_2$ and standard deviations σ_1 and σ_2, respectively, is given by

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

where $d_1 = \bar{x}_{12} - \bar{x}_1$; $d_2 = \bar{x}_{12} - \bar{x}_2$

and $\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$ (combined arithmetic mean)

This formula can also be extended to compute the standard deviation of more than two sets of data.

2. ***Standard deviation of natural numbers*:** The standard deviation of the first n natural numbers is given by

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

For example, the standard deviation of the first 100 (i.e., from 1 to 100) natural numbers will be

$$\sigma = \sqrt{\frac{1}{12}(100^2 - 1)} = \sqrt{\frac{1}{12}(9999)} = \sqrt{833.25} = 28.86$$

Example 12.9: For a group of 50 male workers, the mean and standard deviation of their monthly wages are ₹6300 and ₹900, respectively. For a group of 40 female workers, these are ₹5400 and ₹600, respectively. Find the standard deviation of monthly wages for the combined group of workers.

NOTES

Solution: Given that, Male workers : $n_1 = 50,\ \bar{x}_1 = 6300,\ \sigma_1 = 900$

Female workers : $n_2 = 40,\ \bar{x}_2 = 5400,\ \sigma_2 = 600$

Then, Combined mean, $\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{50 \times 6300 + 40 \times 5400}{50 + 40} = 5{,}900$

and Combined Standard Deviation

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

$$= \sqrt{\frac{50(8{,}10{,}000 + 1{,}60{,}000) + 40(3{,}60{,}000 + 2{,}50{,}000)}{50 + 40}} = ₹900$$

where $d_1 = \bar{x}_{12} - \bar{x}_1 = 5900 - 6300 = -400$ and $d_2 = \bar{x}_{12} - \bar{x}_2 = 5900 - 5400 = 500$.

Example 12.10: A study of the age of 100 persons grouped into intervals 20–22, 22–24, 24–26, ... revealed the mean age and standard deviation to be 32.02 and 13.18, respectively. While checking, it was discovered that the observation 57 was misread as 27. Calculate the correct mean age and standard deviation.

Solution: From the data given in the problem, we have $\bar{x} = 32.02$, $\sigma = 13.18$ and N = 100. We know that

$$\bar{x} = \frac{1}{n}\sum fx \quad \text{or} \quad \Sigma fx = n \times \bar{x} = 100 \times 32.02 = 3202$$

and $\sigma^2 = \frac{1}{n}\sum fx^2 - (\bar{x})^2$ or $\Sigma fx^2 = n[\sigma^2 + (\bar{x})^2] = 100[(13.18)^2 + (32.02)^2]$

$= 100[173.71 + 1025.28] = 100 \times 1198.99 = 1{,}19{,}899$

On substituting the correct observation, we get

$\Sigma fx = 3202 - 27 + 57 = 3232.$

Also $\Sigma fx^2 = 1{,}19{,}899 - (27)^2 + (57)^2 = 1{,}19{,}899 - 729 + 3248 = 1{,}22{,}419$

Thus, Correct A.M., $\bar{x} = \frac{1}{n}\sum fx = \frac{1}{100}(3232) = 32.32.$

and Correct variance, $\sigma^2 = \frac{1}{n}\sum fx^2 - (\bar{x})^2 = \frac{1}{100}(1{,}22{,}419) - (32.32)^2$

$= 1224.19 - 1044.58 = 179.61$

or Correct standard deviation, $\sigma = \sqrt{\sigma^2} = \sqrt{179.61} = 13.402.$

NOTES

Example 12.11: The mean of 5 observations is 15 and the variance is 9. If two more observations having values –3 and 10 are combined with these 5 observations, what will be the new mean and variance of 7 observations?

Solution: From the data of the problem, we have $\bar{x} = 15$, $s^2 = 9$ and $n = 5$. We know that

$$\bar{x} = \frac{1}{n}\sum x \quad \text{or} \quad \sum x = n \times \bar{x} = 5 \times 15 = 75$$

If two more observations having values –3 and 10 are added to the existing 5 observations, then after adding these 6th and 7th observations, we get

$$\sum x = 75 - 3 + 10 = 82$$

Thus, New A.M., $\bar{x} = \frac{1}{n}\sum x = \frac{1}{7}(82) = 11.71$

$$\text{Variance, } s^2 = \frac{1}{n}\sum x^2 - (\bar{x})^2$$

$$9 = \frac{1}{n}\sum x^2 - (15)^2$$

or $$\sum x^2 = 1170$$

On adding two more observations: –3 and 10, we get

$$\sum x^2 = 1170 + (-3)^2 + (10)^2 = 1279$$

$$\text{Variance, } s^2 = \frac{1}{n}\sum x^2 - (\bar{x})^2 = \frac{1}{7}(1279) - (11.71)^2 = 45.59$$

Hence, the new mean and variance of 7 observations is 11.71 and 45.59, respectively.

Chebyshev's Theorem

Standard deviation measures the variation among observations in a data set. If the standard deviation value is small, then values in the data set cluster close to the arithmetic mean. Conversely, a large standard deviation value indicates that the values are scattered more widely around arithmetic mean. The theorem states that:

For any set of data (population or sample) and any constant z greater than 1 (but need not be an integer), the proportion of the values that lie within z standard deviations on either side of the mean is at least $\{1 - (1/z^2)\}$. *That is*

$$\text{RF}\,[|x - \mu| \le z\sigma] \ge 1 - \frac{1}{z^2}$$

where RF is the relative frequency of a distribution.

$z = \frac{x - \mu}{\sigma}$ ← population standardized score, i.e., number of standard deviations a value, x is away from the mean μ (sample or population)

$$= \frac{x - \bar{x}}{s} \leftarrow \text{sample standardized score}$$

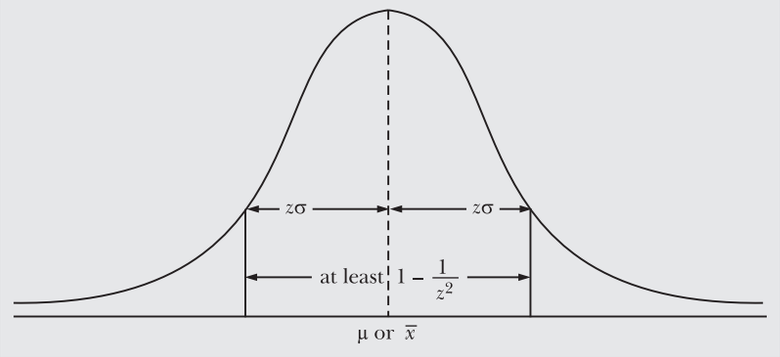


Fig. 12.5 *Chebyshev's Theorem*

For a symmetrical, bell-shaped distribution as shown in Fig. 12.5. Chebyshev's theorem indicates percentage of values that *approximately* fall within *z* standard deviations. The relationship among mean, standard deviation and the set of values is called ***empirical (*or *normal) rule.***

Illustration The theorem is applicable to any data set regardless of the shape of the frequency distribution of values. For example, assume that the marks obtained by 100 students in business statistics had an A.M., $\bar{x}$ = 70 per cent and standard deviation, σ = 10 per cent. Then number of students who obtained marks between 50 and 85 will be determined as follows:

(a) Since, $z = (50 - 70)/10 = -2$, 50 marks fall 2 standard deviations below the mean,
(b) Since, $z = (85 - 70)/10 = 1.5$, 85 marks fall 1.5 standard deviations above the mean.

Applying the Chebyshev's theorem with $z = 2.0$, we have

$$\left(1 - \frac{1}{z^2}\right) = \left[1 - \frac{1}{(2.0)^2}\right] = 0.75$$

This indicates that at least 75 per cent of the students must have obtained marks between 50 and 85.

Empirical Rule For symmetrical, bell-shaped frequency distribution (also called normal curve), the range within which a given percentage of values of the distribution are likely to fall within a specified number of standard deviations of the population mean, μ is determined as follows:

- $\mu \pm \sigma$ covers approximately 68.27 per cent of values in the data set.
- $\mu \pm 2\sigma$ covers approximately 95.45 per cent of values in the data set.

NOTES

NOTES

- $\mu \pm 3\sigma$ covers approximately 99.73 per cent of values in the data set.

These ranges are illustrated in Fig. 12.5.

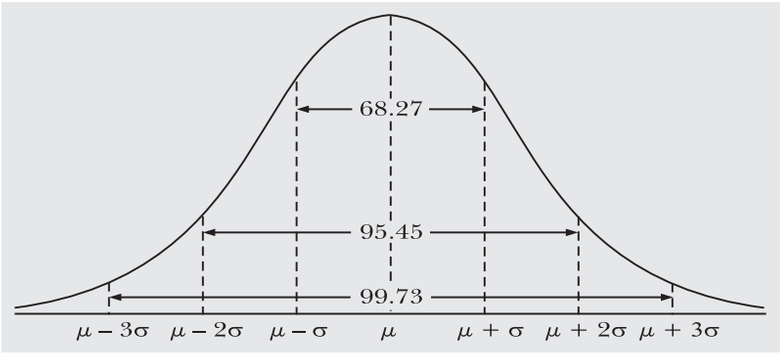


Fig. 12.6 *Area Under Normal Curve*

Relationship Between Different Measures of Variation

(a) Quartile deviation (Q.D.) = $\frac{2}{3}\sigma$

Mean absolute deviation (MAD) = $\frac{4}{5}\sigma$

(b) Quartile deviation = $\frac{5}{6}$MAD

Standard deviation = $\frac{5}{4}$MAD or $\frac{3}{2}$Q.D.

(c) Mean absolute deviation = $\frac{6}{5}$Q.D.

These relationships are applicable only to symmetrical distributions.

Example 12.12: Suppose you are in charge of rationing in a state affected by food shortage. The following reports arrive from a local investigator:

Daily caloric value of food available per adult during current period:

Area	*Mean*	*Standard Deviation*
A	2500	400
B	2000	200

The estimated requirement of an adult is taken as 2800 calories daily and the absolute minimum is 1350. Comment on the reported figures and determine which area in your opinion, need more urgent attention.

Solution: Taking into consideration the entire population of the two areas, we have

Area A: $\mu + 3\sigma = 2500 + 3 \times 400 = 3700$ calories

$\mu - 3\sigma = 2500 - 3 \times 400 = 1300$ calories

NOTES

This calculation shows that adults are taking only 1300 calories, which is much less than the absolute minimum requirement of 1350 calories.

Area B: $\mu + 3\sigma = 2000 + 3 \times 200 = 2600$ calories

$\mu - 3\sigma = 2000 - 3 \times 200 = 1400$ calories

This calculation shows that adults are taking sufficient amount of calories as per requirement of daily calorific need. Hence, area A needs more urgent attention.

Example 12.13: The following data give the number of passengers travelling by airplane from one city to another in one week.

115 122 129 113 119 124 132 120 110 116

Calculate the mean and standard deviation and determine the percentage of class that lie between (i) $\mu \pm \sigma$, (ii) $\mu \pm 2\sigma$ and (iii) $\mu \pm 3\sigma$. What percentage of cases lies outside these limits?

Solution: The calculations for mean and standard deviation are shown in Table 12.7.

Table 12.7 *Calculations of Mean and Standard Deviation*

x	$x - \bar{x}$	$(x - \bar{x})^2$
115	−5	25
122	2	4
129	9	81
113	−7	49
119	−1	1
124	4	16
132	12	144
120	0	0
110	−10	100
116	−4	16
1200	0	436

$$\mu = \frac{1}{n}\sum x = \frac{1200}{10} = 120 \text{ and } \sigma^2 = \frac{1}{n}\sum (x - \bar{x})^2 = \frac{436}{10} = 43.6$$

Therefore, $\sigma = \sqrt{\sigma^2} = \sqrt{43.6} = 6.60$

The percentage of cases that lies between a given limit is as follows:

Interval	*Values within Interval*	*Percentage of Population*	*Percentage Falling Outside*
$\mu \pm \sigma = 120 \pm 6.60$	113, 115, 116, 119	70%	30%
$= 113.4$ and 126.6	120, 122, 124		
$\mu \pm 2\sigma = 120 \pm 2(6.60)$	110, 113, 115, 116, 119	100%	nil
$= 106.80$ and 133.20	120, 122, 124, 129, 132		

NOTES

Example 12.14: A collar manufacturer is considering the production of a new collar to attract young men. Thus, following statistics of neck circumference are available based on measurement of a typical group of the college students:

Mid value (in inches) :	12.0	12.5	13.0	13.5	14.0	14.5	15.0	15.5	16.0
Number of students :	2	16	36	60	76	37	18	3	2

Compute the standard deviation and use the criterion $\bar{x} \pm 3\sigma$, where σ is the standard deviation and $\bar{x}$ is the arithmetic mean to determine the largest and smallest size of the collar he should make in order to meet the needs of practically all the customers bearing in mind that collar are worn on average half inch longer than neck size.

Solution: Calculations for mean and standard deviation in order to determine the range of collar size to meet the needs of customers are shown in Table 12.8.

Table 12.8 *Calculations for Mean and Standard Deviation*

Mid-value (in inches)	*Number of students*	$\frac{x-A}{h}=\frac{x-14}{0.5}$	*fd*	fd^2
12.0	2	–4	–8	32
12.5	16	–3	–48	144
13.0	36	–2	–72	144
13.5	60	–1	–60	60
14.0 ← A	76	0	0	0
14.5	37	1	37	37
15.0	18	2	36	72
15.5	3	3	9	27
16.0	2	4	8	32
	N = 250		– 98	548

$$\text{Mean, } \bar{x} = A + \frac{\sum fd}{n} \times h = 14.0 - \frac{98}{250} \times 0.5 = 14.0 - 0.195 = 13.805$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times h = \sqrt{\frac{548}{250} - \left(\frac{-98}{250}\right)^2} \times 0.5$$

$$= \sqrt{2.192 - 0.153} \times 0.5 = 1.427 \times 0.5 = 0.7135$$

Largest and smallest neck size = $\bar{x} \pm 3\sigma = 13.805 \pm 3 \times 0.173 = 11.666$ and 15.944.

Since all the customers are to wear collar half inch longer than their neck size, 0.5 is to be added to the neck size range given above. The new range then becomes (11.666 + 0.5) and (15.944 + 0.5) or 12.2 and 16.4 inches (approx).

Coefficient of Variation

A relative measure called the **coefficient of variation** (CV) developed by Karl Pearson is very useful measure for (i) comparing two or more data sets

NOTES

expressed in different units of measurement, and (ii) comparing data sets that are in same unit of measurement but the mean values of data sets are not same.

The coefficient of variation (CV) that measures the standard deviation relative to the mean in percentages is computed as follows:

$$\text{Coefficient of variation (CV)} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

Multiplying by 100 converts the decimal to a percent. The lower value of CV indicates uniformity (or consistency) among values in any data set.

Example 12.15: The weekly sales of two products A and B were recorded as given below:

Product A	:	59	75	27	63	27	28	56
Product B	:	150	200	125	310	330	250	225

Find out which of the two shows greater fluctuation in sales.

Solution: Calculating coefficient of variation for both the products to compare fluctuation in their sales.

Product A: Let A = 56 be the assumed mean of sales for product A.

Calculations of the Mean and Standard Deviation

Sales (x)	*Frequency* (f)	$d = x - A = x - 56$	fd	fd^2
27	2	−29	−58	1682
28	1	−28	−28	784
56 ← A	1	0	0	0
59	1	3	3	9
63	1	7	7	49
75	1	19	19	361
	7		−57	2885

$$\bar{x} = A + \frac{1}{n}\sum fd = 56 - \frac{57}{7} = 47.86$$

$$s_A^2 = \frac{1}{n}\sum fd^2 - \left(\frac{1}{n}\sum fd\right)^2 = \frac{2885}{7} - \left(-\frac{57}{7}\right)^2$$

$$= 412.14 - 66.30 = 345.84$$

$$s_A = \sqrt{345.84} = 18.59$$

Then $\text{CV (A)} = \frac{s_A}{\bar{x}} \times 100 = \frac{18.59}{47.86} \times 100 = 38.84$ per cent

NOTES

Product B: Let A = 225 be the assumed mean of sales for product B.

Calculations of Mean and Standard Deviation

Sales (x)	*Frequency* (f)	$d = x - A$ $= x - 225$	fd	fd^2
125	1	–100	–100	10,000
150	1	–75	–75	5625
200	1	–25	–25	625
225 ← A	1	0	0	0
250	1	25	25	625
310	1	85	85	7225
330	1	105	105	11,025
	7		15	35,125

$$\bar{x} = A + \frac{1}{n}\sum fd = 225 + \frac{15}{7} = 227.14$$

$$s_B^2 = \frac{1}{n}\sum fd^2 - \left(\frac{1}{n}\sum fd\right)^2 = \frac{35,125}{7} - \left(\frac{15}{7}\right)^2 = 5017.85 - 4.59 = 5013.26$$

or $s_B = \sqrt{5013.26} = 70.80$

Then CV(B)= $\frac{s_B}{\bar{x}} \times 100 = \frac{70.80}{227.14} \times 100 = 31.17$ per cent

Since the coefficient variation for product A is more than that of product B, therefore the sales fluctuation in case of product A is higher.

Example 12.16: From the analysis of monthly wages paid to employees in two service organizations X and Y, the following results were obtained:

	Organization X	*Organization Y*
Number of wage-earners	550	650
Average monthly wages	5000	4500
Variance of the distribution of wages	900	1600

(a) Which organization pays a larger amount as monthly wages?

(b) In which organization is there greater variability in individual wages of all the wage earners taken together?

Solution: (a) Comparing the total wages to find out which organization X or Y pays larger amount of monthly wages:

Total wage bill paid monthly by X and Y is

$X : n_1 \times \bar{x}_1 = 550 \times 5000 = ₹27,50,000$

$Y : n_2 \times \bar{x}_2 = 650 \times 4500 = ₹29,25,000$

Organization Y pays a larger amount as monthly wages as compared to organization X.

(b) For calculating the combined variation, first calculating the combined mean as follows:

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{27,50,000 + 29,25,000}{1200} = ₹4729.166$$

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

$$= \sqrt{\frac{550(900 + 73,351.05) + 650(1600 + 52,517.05)}{550 + 650}}$$

$$= \sqrt{\frac{4,08,38,080.55 + 3,51,76,082.50}{1200}}$$

$$= \sqrt{63345.13} = 251.68$$

where $d_1 = \bar{x}_{12} - \bar{x}_1 = 4729.166 - 5000 = -270.834$

$d_2 = \bar{x}_{12} - \bar{x}_2 = 4729.166 - 4500 = 229.166$

Check Your Progress

3. What are absolute measures?
4. What does the word coefficient mean?
5. How can the limitations of range be overcome?

12.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Certain essential requisites that help in identifying the merits and demerits of individual measure of variation are as follows:
   (i) Should be based on all the values (elements) in the data set.
   (ii) Should be calculated easily, quickly and accurately.
   (iii) Should be unaffected by the fluctuations in sample size and also by outliers.
   (iv) Should be further mathematical or algebraic changes are possible.
2. A small dispersion among values in the data set indicates that values in the data set are clustered closely around the mean, implying that the mean is a reliable average.
3. Absolute measures are described by a number (or value) to represent the amount of variation (or difference) among values in a data set. Such

NOTES

NOTES

a value is expressed in the same unit of measurement such as rupee, inch, foot, kilogram, ton, etc., as the values in the data set.

4. The word 'coefficient' means a number that is independent of any unit of measurement.
5. The limitations or disadvantages of range can partially be overcome by using another measure of variation called Interquartile Range or Deviation (IQR).

12.5 SUMMARY

- Identifying the causes and then measuring the dispersion is useful to draw statistical inference (estimation of parameter, hypothesis testing, forecasting and so on). A small dispersion among values in the data set indicates that values in the data set are clustered closely around the mean, implying that the mean is a reliable average.
- Conversely, if values in the data set are widely clustered around the mean, then this implies that the mean is not a reliable average, i.e. mean is not representative of the data.
- Measures of variation help to understand the extent an average represents the characteristic of a data set. If the extent of dispersion of values is less on each side of an average value, then it indicates high uniformity among values in the distribution.
- Measuring variation helps to identify the nature and causes of variation. Such information is useful in controlling the variations.
- Measures of variation help in comparing variation in two or more sets of data with respect to their uniformity or consistency.
- Measures of dispersion (or variation) based on the purpose of measuring are classified into two categories: absolute and relative measures.
- The limitations or disadvantages of range can partially be overcome by using another measure of variation called Interquartile Range or Deviation (IQR). The IQR measures the spread within middle half of the values in the data set so as to minimize the influence of outliers (extreme values) in the calculation of range.
- Since average (mean) deviation of individual values in the data set from their actual arithmetic mean (A.M.) is always zero, therefore such a measure would not indicate any variation.
- While computing absolute value of each deviation from arithmetic mean, another way to ignore sign of negative deviations from mean is to square such values. The sum of all such squared deviations is then divided by the number of observations in the data set.

- The numerical value of population or a sample variance is difficult to interpret because it is expressed in square units. To reach an interpretable measure of variance expressed in the units of original data, we take a positive square root of the variance, called standard deviation or root-mean square deviation.
- Interquartile range = $Q_3 - Q_1$

  Quartile deviation, QD = $\frac{Q_3 - Q_1}{2}$

  Coefficient of QD = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$
- Mean average deviation

  For ungrouped data

  (i) MAD = $\frac{\sum |x - \bar{x}|}{n}$, for sample

  (ii) MAD = $\frac{\sum |x - \mu|}{N}$, for population

  (iii) MAD = $\frac{\sum |x - Me|}{n}$, from median

  For grouped data MAD = $\frac{\sum f|x - \bar{x}|}{\sum f}$
- Coefficient of MAD = $\frac{MAD}{\bar{x} \text{ or } Me} \times 100$
- Variance

  Ungrouped data

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

  $$= \frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2$$

  where $d = x - A$; A is any assumed A.M. value

  Grouped data, $\sigma^2 = \left[\frac{\sum f d^2}{N} - \left(\frac{\sum f d}{N}\right)^2\right] h$

  where $d = (m - A)/h$; h is the class interval and m is the mid-value of class intervals.
- Standard deviation

  Ungrouped data, $\sigma = \sqrt{\sigma^2}$

  Grouped data, $\sigma = \sqrt{\frac{\sum f d^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h$
- Coefficient of variation (CV) = $\frac{\sigma}{\bar{x}} \times 100$

NOTES

12.6 KEY WORDS

- **Inquartile range:** The IQR measures the spread within middle half of the values in the data set so as to minimize the influence of outliers (extreme values) in the calculation of range.
- **Quartile deviation:** The Quartile Deviation is a simple way to estimate the spread of a distribution about a measure of its central tendency (usually the mean).
- **Standard deviation:** In statistics, the standard deviation is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

12.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. A petrol filling station has recorded the following data for litres of petrol sold per automobile in a sample of 680 automobiles:

Petrol Sold (*Litres*)	*Frequency*
0 – 4	74
5 – 9	192
10 – 14	280
15 – 19	105
20 – 24	23
25 – 29	6

   Compute the mean and standard deviation for the data.

2. A frequency distribution for the duration of 20 long-distance telephone calls in minutes is as follows:

Call Duration (*Minutes*)	*Frequency*
4 – 7	4
8 – 11	5
12 – 15	7
16 – 19	2
20 – 23	1
24 – 27	1

   Compute the mean, variance, and standard deviation.

3. Automobiles travelling on a highway are checked for speed by the police. Following is a frequency distribution of speeds:

Speed (*km per hours*)	*Frequency*
45 – 49	10
50 – 54	40
55 – 59	150
60 – 64	175
65 – 69	75
70 – 74	15
75 – 79	10

What is the mean, variance, and standard deviation of speed for the automobiles travelling on the highway?

4. A work-standards expert observes the amount of time (in minutes) required to prepare a sample of 10 business letters in the office with observations in ascending order: 5, 5, 5, 7, 9, 14, 15, 15, 16, 18.
   (a) Determine the range and middle 70 per cent range for the sample.
   (b) If the sample mean of the data is 10.9, then calculate the mean absolute deviation and variance.

Long Answer Questions

1. ABC Stereos, a wholesaler, was contemplating becoming the supplier to three retailers, but inventory shortages have forced him to select only one. ABC's credit manager is evaluating the credit record of these three retailers. Over the past 5 years these retailers' accounts receivable have been outstanding for the following average number of days. The credit manager feels that consistency, in addition to lowest average, is important. Based on relative dispersion, which retailer would make the best customer?

Lee :	62.20	61.80	63.40	63.00	61.70
Forest :	62.50	61.90	63.80	63.00	61.70
Davis :	62.00	61.90	63.00	63.90	61.50

2. A purchasing agent obtained samples of 60 watt bulbs from two companies. He had the samples tested in his own laboratory for length of life with the following results:

Length of Life	*Samples from*	
(in hours)	*Company A*	*Company B*
1700 – 1900	10	3
1900 – 2100	16	40
2100 – 2300	20	12
2300 – 2500	8	3
2500 – 2700	6	2

NOTES

(a) Which company's bulbs do you think are better in terms of average life?

(b) If prices of both the companies are same, which company's bulbs would you buy and why?

3. The Chief Medical Officer of a hospital conducted a survey of the number of days 200 randomly chosen patients stayed in the hospital following an operation. The data are given below

Hospital stay (in days):	1–3	4–6	7–9	10–12	13–15	16–18	19–21	22–24
Number of patients:	18	90	44	21	9	9	4	5

(a) Calculate the mean number of days patients stay in the hospital along with standard deviation of the same.

(b) How many patients are expected to stay between 0 and 17 days.

4. A nursing home is well-known in effective use of pain killing drugs for seriously ill patients. In order to know approximately how many nursing staff to employ, the nursing home has begun keeping track of the number of patients that come every week for checkup. Each week the CMO records the number of seriously ill patients and the number of routine patients. The data for the last 5 weeks is as follows:

Seriously ill patients :	33	50	22	27	48
Routine patients :	34	31	37	36	27

(a) Find the limits within which the middle 75 per cent of seriously ill patients per week should fall.

(b) Find the limits within which the middle 68 per cent of routine patients per week should fall.

5. There are a number of possible measures of sales performance, including how consistent a sales person is, in meeting established sales goals. The following data represent the percentage of goal met by each of three sales persons over the last five years

Raman :	88	68	89	92	103
Sindhu :	76	88	90	86	79
Prasad :	104	88	118	88	123

Which salesman is most consistent. Suggest an alternative measure of consistency (if possible).

12.8 FURTHER READINGS

Creswell, John W. 2002. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* London: Sage Publications.

Booth, Wayne, Gregory G. Colomb and Joseph M. Williams. 1995. *The Craft of Research.* Chicago: University of Chicago Press.

Kumar, B. 2006. *Research Methodology.* New Delhi: Excel Books.

Paneerselvam, R. 2009. *Research Methodology.* New Delhi: Prentice Hall of India.

Gupta, D. 2011. *Research Methodology*. New Delhi: PHI Learning Private Limited.

NOTES

UNIT 13 MEASURES OF DISPERSION-II

NOTES

Structure

13.0 INTRODUCTION

The median of a distribution splits the data into two equally-sized groups. In the same way, the quartiles are the three values that split a data set into four equal parts. In a similar way, the deciles of a distribution are the nine values that split the data set into ten equal parts. The percentiles of a distribution are the 99 values that split the data set into a hundred equal parts.

In this unit, we will discuss range, quartiles, deciles and percentiles with the help of illustrations.

13.1 OBJECTIVES

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

- Define range and understand its method of calculation
- Discuss the advantages, disadvantages and applications and range
- Explain the definition and formula for quartiles, deciles and percentiles

13.2 RANGE

The calculation of range as a measure of dispersion is based on the location of the largest and the smallest values in the data. Thus, **range** is defined as the difference between the largest and lowest observed values in a data set. In other words, it is the length of an interval which covers the highest and

lowest observed values in a data set and measures the dispersion or spread within the interval.

Range (R)= Highest value of an observation – Lowest value of an observation

$$=H - L \quad (13.1)$$

For example, if the smallest value of an observation in the data set is 160 and largest value is 250, then the range is 250 – 160 = 90.

For grouped frequency distribution of values in the data set, the range is the difference between the upper class limit of the last class and the lower class limit of first class. In this case, the range obtained may be higher than as compared to ungrouped data because class limits are extended slightly beyond the extreme values in the data set.

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

$$\text{Coefficient of range} = \frac{H - L}{H + L} \quad (13.2)$$

Example 13.1: The following are the sales figures of a firm for the last 12 months

Months	1	2	3	4	5	6	7	8	9	10	11	12
Sales (₹ '000)	80	82	82	84	84	86	86	88	88	90	90	92

Calculate the range and coefficient of range of sales for the last 12 months.

Solution: Given that, H = 92 and L = 80. Therefore,

$$\text{Range} = H - L = 92 - 80 = ₹12$$

and $$\text{Coefficient of range} = \frac{H - L}{H + L} = \frac{92 - 80}{92 + 80} = \frac{12}{172} = 0.069$$

Example 13.2: The following data show the waiting time (to the nearest 100th of a minute) of telephone calls to be matured:

Waiting Time	*Frequency*	*Waiting Time* (*Minutes*)	*Frequency* (*Minutes*)
0.10 – 0.35	06	0.88 – 1.13	8
0.36 – 0.61	10	1.14 – 1.39	4
0.62 – 0.87	08		

Calculate the range and coefficient of range for telephone calls to be matured.

Solution: Given that, H = 1.39 and L = 0.10. Therefore,

$$\text{Range} = H - L = 1.39 - 0.10 = 1.29 \text{ min}$$

and $$\text{Coefficient of Range} = \frac{H - L}{H + L} = \frac{1.39 - 0.10}{1.39 + 0.10} = \frac{1.29}{1.49} = 0.865$$

NOTES

13.2.1 Advantages, Disadvantages and Applications of Range

Advantages

1. The measurement of range is independent of the measure of central tendency and easy to calculate and understand.
2. The knowledge of range is useful in cases where the purpose is only to know the extent of extreme variation, such as quality control limits, temperature, rainfall and so on.

Disadvantages

1. The calculation of range is based on only two values—largest and smallest in the data set. Thus, the value of range is influenced by two extreme values and completely independent of the other values. For example, range of two data sets {1, 2, 3, 7, 12} and {1, 1, 1, 12, 12} is 11 but the two data sets differ in terms of overall dispersion of values
2. The value of range is sensitive to changes in sample size, i.e., different samples of the same size from the same population may have different ranges.
3. Range cannot be computed for open-ended frequency distributions because no highest or lowest value exists in such cases.
4. The value of range does not describe variation among values between highest and lowest value in a given data set. For example, each of the following data set

Set 1	:	9	21	21	21	21	21	21	21
Set 2	:	9	9	9	9	21	21	21	21
Set 3	:	9	10	12	14	15	19	20	21

has a range of 21 – 9 = 12, but the variation of values between the highest and lowest values is different in each case.

Applications

1. The knowledge of range is useful in the study of small variations among values in a data set. Variation (fluctuation) in share prices and other commodities that are very sensitive to price changes from one period to another may easily be understood by calculating the range of such variations (fluctuations).
2. Quality control is exercised by preparing suitable *control charts.* The control charts are prepared on setting an upper control limit (range) and a lower control limit (range) within which quality of products is acceptable. The variation in the quality beyond these *ranges* requires necessary remedial actions.

3. For weather forecasts, the knowledge of range (difference between maximum and minimum temperature or rainfall) is important.

NOTES

13.3 QUARTILES, DECILES AND PERCENTILES: CHARACTERISTICS AND SIMPLE PROBLEMS

The basic purpose of median is to explore the characteristics of a data set. The analysis begins by arranging all the observations in either ascending or descending order of their magnitude and then dividing this ordered series into two equal parts. However, to have more knowledge about the data set, we may decompose it into more parts of equal size. The measures of central tendency which are used for dividing the data into several equal parts are called *partition values*.

In this section, we will discuss three methods for data analysis by dividing it into *four*, *ten* and *hundred* parts of equal sizes and the corresponding partition values are called *quartiles*, *deciles* and *percentiles*. All these values can be determined in the same way as median. The only difference is in their location.

Quartiles: The values of observations in a data set, when arranged in an ordered sequence, can be divided into four equal parts using three quartiles namely Q_1, Q_2, and Q_3. The first quartile Q_1 divides a distribution in such a way that 25 per cent (=n/4) of observations have a value less than Q_1 and 75 per cent (= 3n/4) have a value more than Q_1. Second quartile Q_2 has the same number of observations above and below it. It is therefore same as median value.

The quartile Q_3 divides the data set in such a way that 75 per cent of the observations have a value less than Q_3 and 25 per cent have a value more than Q_3.

The generalized formula for calculating quartiles in case of grouped data is:

$$Q_i = l + \left\{\frac{i\,(n/4) - cf.}{f}\right\} \times h; \quad i = 1, 2, 3 \qquad (13.3)$$

where *cf* is the cumulative frequency prior to the quartile class interval; *l* is the lower limit of the quartile class interval; *f* is the frequency of the quartile class interval and *h* is width of the class interval.

Deciles: The values of observations in a data set, when arranged in an ordered sequence, can be divided into 10 equal parts, using nine deciles, D_i (i = 1, 2, ..., 9). The generalized formula for calculating deciles in case of grouped data is

$$D_i = l + \left\{\frac{i\,(n/10) - cf.}{f}\right\} \times h; \quad i = 1, 2, ..., 9 \qquad (13.4)$$

where the symbols have their usual meaning and interpretation.

NOTES

Percentiles: The values of observations in a data, when arranged in an ordered sequence, can be divided into hundred equal parts using 99 percentiles, P_i (i = 1, 2,..., 99). In general, the ith percentile is a number that has i per cent of the data values at or below it and (100 – i) per cent of the data values at or above it. The lower quartile (Q_1), median and upper quartile (Q_3) is also the 25th percentile, 50th percentile and 75th percentile, respectively. For example, if you are told that you scored 90th percentile in a test (like the CAT), it indicates that 90 per cent of the scores were at or below your score while 10 per cent were at or above your score.

The generalized formula for calculating percentiles in case of grouped data is

$$P_i = l + \left\{\frac{i(n/100) - cf.}{f}\right\} \times h; \quad i = 1, 2, ..., 99 \qquad (13.5)$$

where the symbols have their usual meaning and interpretation.

13.3.1 Graphical Method for Calculating Partition Values

The graphical method of determining various partition values can be summarized into following steps:

1. Draw an Ogive (cumulative frequency curve) by 'less than' method.
2. Take the values of observations or class intervals along the horizontal scale (i.e. x-axis) and cumulative frequency along vertical scale (i.e., y-axis).
3. Determine the median value, i.e., value of (n/2)th observation, where n is the total number of observations in the data set.
4. Locate this value on the y-axis and from this point draw a line parallel to the x-axis meeting the ogive at a point, say P. Draw a perpendicular on x-axis from P and it meets the x-axis at a point, say M.

The other partition values such as quartiles, deciles and percentiles can also be obtained by drawing lines parallel to the x-axis to the distance $i(n/4)$(i=1,2,3); $i(n/10)$ (i = 1, 2,..., 9), and i (n/100) (i = 1, 2,..., 99), respectively.

Example 13.3: The following is the distribution of weekly wages of 600 workers in a factory:

Weekly Wages (*in* ₹)	*Number of Workers*	*Weekly Wages* (*in* ₹)	*Number of Workers*
Below 875	69	1100 – 1175	58
875 – 950	167	1175 – 1250	24
950 –1025	207	1250 – 1325	10
1025 –1100	65		600

(a) Draw an ogive for the above data and hence obtain the median value. Check it against the calculated value.

(b) Obtain the limits of weekly wages of central 50 per cent of the workers.

(c) Estimate graphically the percentage of workers who earned weekly wages between 950 and 1250

NOTES

Solution: (a) The calculations required for median value are shown in Table 13.1.

Table 13.1 *Calculations of Median Value*

Weekly Wages (in ₹)	*Number of Workers (f)*	*Cumulative Frequency (Less than type)*	*Percent Cumulative Frequency*
Less than 875	69	69	11.50
Less than 950	167	236 ← Q_1 class	39.33
Less than 1025	207	443 ← Median class	73.83
Less than 1100	65	508 ← Q_3 class	84.66
Less than 1175	58	566	94.33
Less than 1250	24	590	98.33
Less than 1325	10	600	100.00

Since a median observation in the data set is the $(n/2)$th observation = (600/2)th= 300th observation. This observation lies in the class interval 950–1025. Applying the formula (3-16) to calculate median wage value, we have

$$\text{Med} = l + \frac{(n/2) - \text{c}f.}{f} \times h$$

$$= 950 + \frac{300 - 236}{207} \times 75 = 950 + 23.2 = ₹973.2 \text{ per week}$$

The median wage value can also be obtained by applying the graphical method as shown in Fig. 13.1.

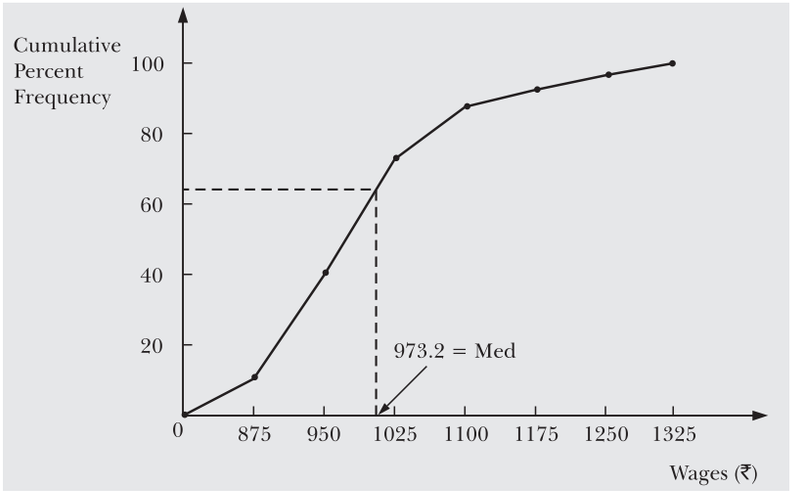


Fig. 13.1 *Cumulative Frequency Curve*

NOTES

Q_1 = value of $(n/4)$th observation
= value of (600/4)th = 150th observation

(b) The limits of weekly wages of central 50 per cent of the workers can be calculated by taking the difference of Q_1 and Q_3. This implies that Q_1 lies in the class interval 875–950. Thus,

$$Q_1 = \text{value of } (n/4)\text{th observation}$$
$$= l + \frac{(n/4) - \text{c.f.}}{f} \times h$$
$$= 875 + \frac{150 - 69}{167} \times 75 = 875 + 36.38 = ₹911.38 \text{ per week}$$

Similarly, Q_3 = Value of $(3n/4)$th observation
= Value of (3×600/4)th = 450th observation

This value of Q_3 lies in the class interval 1025–1100. Thus,

$$Q_3 = l + \frac{(3n/4) - \text{c.f.}}{f} \times h = 1025 + \frac{450 - 443}{65} \times 75$$
$$= 1025 + 8.08 = ₹1033.08 \text{ per week}$$

Hence, the limits of weekly wages of central 50 per cent workers are ₹411.38 and ₹533.08.

(c) The percentage of workers who earned weekly wages less than or equal to ₹950 is 39.33 and who earned weekly wages less than or equal to ₹1250 is 98.33. Thus, the percentage of workers who earned weekly wages between ₹950 and ₹1250 is (98.33 – 39.33) = 59.

Example 13.4: You are working for the transport manager of a 'call center' which hires cars for the staff. You are interested in the weekly distances covered by these cars. Kilometers recorded for a sample of hired cars during a given week yielded the following data:

Kilometers Covered	*Number of Cars*	*Kilometers Covered*	*Number of Cars*
100–110	4	150–160	8
110–120	0	160–170	5
120–130	3	170–180	0
130–140	7	180–190	2
140–150	11		40

(a) Form a cumulative frequency distribution and draw a cumulative frequency ogive.

(b) Estimate graphically the number of cars which covered less than 165 km in the week.

(c) Calculate Q_1, Q_2, Q_3 and P_{75}.

NOTES

Solution: (a) The calculations to obtain a cumulative frequency distribution and to draw ogive are shown in Table 13.2.

Table 13.2 *Calculations to Draw Ogive*

Kilometers Covered Less than	*Number of Cars*	*Cumulative Frequency*	*Per cent Cumulative Frequency*
110	4	4	10.0
120	0	4	10.0
130	3	7	17.5
140	7	14 ← Q_1	35.0
150	11	25← Me = Q_2	62.5
160	8	33 ← Q_3	82.5← P_{75}
170	5	38	95.0
180	0	38	95.0
190	2	40	100.0

Plotting cumulative frequency values on the graph paper, frequency polygon is as shown in Fig. 13.2.

(b) The number of cars which covered less than 165 km in the week is 35 as shown in the Fig. 13.2.

(c) Since there are 40 observations in the data set, taking 10th, 20th and 30th cumulative values that correspond to Q_1, Q_2 and Q_3, respectively. These values from the graph give $Q_1 = 134$, $Q_2 = 146$ and $Q_3 = 156$.

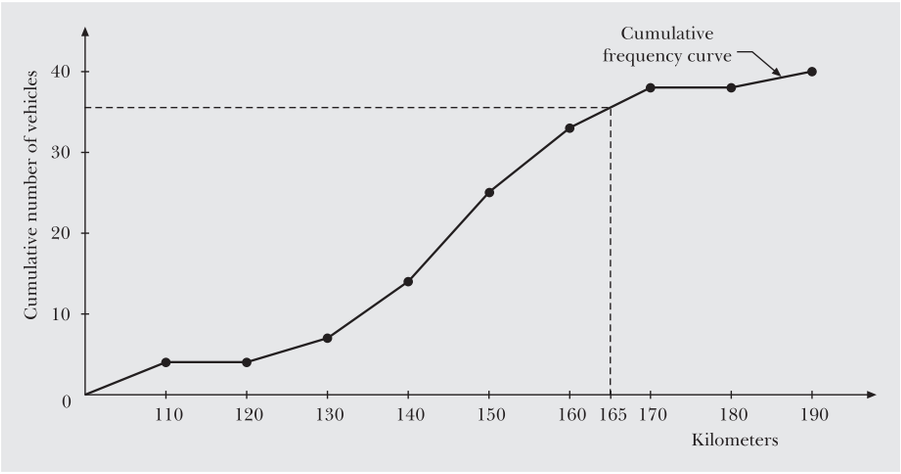


Fig. 13.2 *Cumulative Frequency Curve*

$$P_{75} = l + \frac{i(n/100) - c.f.}{f} \times h = 150 + \frac{75(40/100) - 25}{8} \times 10 = 156.25$$

NOTES

This implies that 75 per cent of cars covered less than or equal to 156.25 kilometers.

Example 13.5: The following distribution gives the pattern of overtime work per week done by 100 employees of a company. Calculate median, first quartile and seventh decile.

Overtime hours :	10–15	15–20	20–25	25–30	30–35	35–40
No. of employees :	11	20	35	20	8	6

Calculate Q_1, D_7 and P_{60}.

Solution: The calculations of median, first quartile (Q_1), and seventh decile (D_7) are shown in Table 13.3.

Table 13.3 *Calculations for Q_1, D_7 and P_{60}*

Overtime Hours	*Number of Employees*	*Cumulative Frequency (Less than type)*
10–15	11	11
15–20	20	31← Q_1 class
20–25	35	66← Median and P_{60} class
25–30	20	86← D_7 class
30–35	8	94
35–40	6	100
	100	

Since the number of observations in the data set are 100, the median value is $(n/2)$th = (100/ 2)th = 50th observation. This observation lies in the class interval 20–25. Applying the formula (13.4) to get median overtime hours value, we have

$$\text{Med} = l + \frac{(n/2) - cf}{f} \times h = 20 + \frac{50 - 31}{35} \times 5 = 20 + 2.714 = 22.714 \text{ hours}$$

Thus, Q_1 = value of $(n/4)$ th observation = value of (100/4)th = 25th observation

$$= l + \frac{(n/4) - cf}{f} \times h = 15 + \frac{25 - 11}{20} \times 5 = 15 + 3.5 = 18.5 \text{ hours}$$

D_7 = value of $(7n/10)$th observation = value of $(7 \times 100)/10$ =70th observation

$$= l + \frac{(7n/10) - cf}{f} \times h = 25 + \frac{70 - 66}{20} \times 5 = 25 + 1 = 26 \text{ hours}$$

P_{60} = Value of $(60n/100)$th observation = 60×(100/100) = 60th observation

$$= l + \frac{(60 \times n/100) - cf}{f} \times h = 20 + \frac{60 - 31}{35} \times 5 = 24.14 \text{ hours}$$

Check Your Progress

1. Define range and its calculation.
2. State the advantages of range.
3. What is the basic purpose of median?
4. What are partition values?

13.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The calculation of range as a measure of dispersion is based on the location of the largest and the smallest values in the data. Thus, range is defined as the difference between the largest and lowest observed values in a data set. In other words, it is the length of an interval which covers the highest and lowest observed values in a data set and measures the dispersion or spread within the interval.
2. The advantages of range are as follows:
   - The measurement of range is independent of the measure of central tendency and easy to calculate and understand.
   - The knowledge of range is useful in cases where the purpose is only to know the extent of extreme variation, such as quality control limits, temperature, rainfall and so on.
3. The basic purpose of median is to explore the characteristics of a data set. The analysis begins by arranging all the observations in either ascending or descending order of their magnitude and then dividing this ordered series into two equal parts.
4. The measures of central tendency which are used for dividing the data into several equal parts are called *partition values*.

13.5 SUMMARY

- The calculation of range as a measure of dispersion is based on the location of the largest and the smallest values in the data. Thus, range is defined as the difference between the largest and lowest observed values in a data set.
- Range (R)= Highest value of an observation – Lowest value of an observation

  =H – L

NOTES

NOTES

- The measurement of range is independent of the measure of central tendency and easy to calculate and understand.
- The knowledge of range is useful in cases where the purpose is only to know the extent of extreme variation, such as quality control limits, temperature, rainfall and so on.
- The calculation of range is based on only two values—largest and smallest in the data set. Thus, the value of range is influenced by two extreme values and completely independent of the other values.
- The knowledge of range is useful in the study of small variations among values in a data set. Variation (fluctuation) in share prices and other commodities that are very sensitive to price changes from one period to another may easily be understood by calculating the range of such variations (fluctuations).
- Quality control is exercised by preparing suitable *control charts.* The control charts are prepared on setting an upper control limit (range) and a lower control limit (range) within which quality of products is acceptable.
- The basic purpose of median is to explore the characteristics of a data set. The analysis begins by arranging all the observations in either ascending or descending order of their magnitude and then dividing this ordered series into two equal parts.
- The values of observations in a data set, when arranged in an ordered sequence, can be divided into four equal parts using three quartiles namely Q1, Q2, and Q3.
- The values of observations in a data set, when arranged in an ordered sequence, can be divided into 10 equal parts, using nine deciles.
- The values of observations in a data, when arranged in an ordered sequence, can be divided into hundred equal parts using 99 percentiles.

13.6 KEY WORDS

- **Quartile:** A quartile is a type of quantile. The first quartile (Q1) is defined as the middle number between the smallest number and the median of the data set. The second quartile (Q2) is the median of the data. The third quartile (Q3) is the middle value between the median and the highest value of the data set.
- **Decile:** In descriptive statistics, a decile is any of the nine values that divide the sorted data into ten equal parts, so that each part represents 1/10 of the sample or population. A decile is one possible form of a quantile; others include the quartile and percentile.

- **Percentile:** A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations falls.

NOTES

13.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. How is coefficient of range calculated?
2. State the disadvantages of range.
3. What is the generalised formula for calculating quartiles in case of grouped data?

Long Answer Questions

1. Describe the applications of range.
2. Analyse the graphical method for calculating partition values.
3. Differentiate between quartiles, deciles and percentiles.

13.8 FURTHER READINGS

Creswell, John W. 2002. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* London: Sage Publications.

Booth, Wayne, Gregory G. Colomb and Joseph M. Williams. 1995. *The Craft of Research.* Chicago: University of Chicago Press.

Kumar, B. 2006. *Research Methodology.* New Delhi: Excel Books.

Paneerselvam, R. 2009. *Research Methodology.* New Delhi: Prentice Hall of India.

Gupta, D. 2011. *Research Methodology*. New Delhi: PHI Learning Private Limited.

NOTES

UNIT 14 DIAGRAMMATIC AND GRAPHIC REPRESENTATION OF DATA

Structure

14.0 INTRODUCTION

This unit will introduce you to graphic representation of data. Graphical or pictorial representation of data helps in giving a visual indication of magnitudes, groupings, trends and patterns in the data. These also help facilitate comparisons between two or more sets of data. Diagrammatic representations include bar diagrams, pie charts and pictograms, whereas graphic representation includes histograms, frequency polygons and cumulative frequency curves or ogives.

14.1 OBJECTIVES

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

- Explain the diagrammatic representation of data
- Analyse pictogram as a sign language
- Discuss graphic representation of data
- Differentiate between histograms, frequency polygon and ogives

14.2 GRAPHIC PRESENTATION: BAR DIAGRAM, PIE CHART AND PICTOGRAM

NOTES

The data we collect can often be more easily understood for interpretation if it is presented graphically or pictorially. Diagrams and graphs give visual indications of magnitudes, groupings, trends and patterns in the data. These important features are more simply presented in the form of graphs. Also, diagrams facilitate comparisons between two or more sets of data.

The diagrams should be clear and easy to read and understand. Too much information should not be represented through the same diagram; otherwise, it may become cumbersome and confusing. Each diagram should include a brief and self-explanatory title dealing with the subject matter. The scale of the presentation should be chosen in such a way that the resulting diagram is of appropriate size. The intervals on the vertical as well as the horizontal axis should be of equal size; otherwise, distortions would occur.

Diagrams are more suitable to illustrate discrete data, while continuous data is better represented by graphs. The following are the diagrammatic and graphic representation methods that are commonly used.

Diagrammatic representation

(i) Bar diagram

(ii) Pie chart

(iii) Pictogram

(iv) Cartograms

(i) Bar diagram: Bars are simply vertical lines where the lengths of the bars are proportional to their corresponding numerical values. The width of the bar is unimportant but all bars should have the same width so as not to confuse the reader of the diagram. Additionally, the bars should be equally spaced.

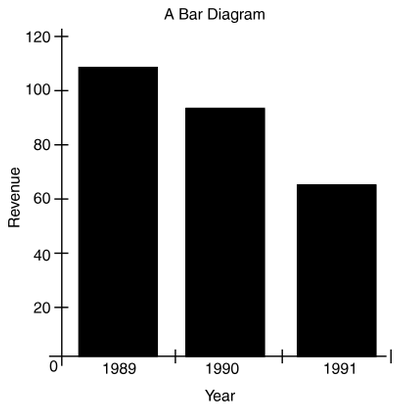
Example 14.1: Suppose that the following were the gross revenues (in $100,000.00) for a company *XYZ* for the years 1989, 1990 and 1991.

Year	Revenue
1989	110
1990	95
1991	65

Construct a bar diagram for this data.

Solution: The bar diagram for this data can be constructed as follows with the revenues represented on the vertical axis and the years represented on the horizontal axis.

NOTES

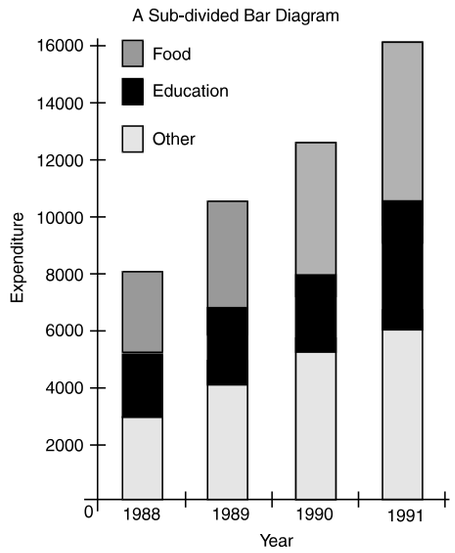


The bars drawn can be subdivided into components depending upon the type of information to be shown in the diagram. This will be clear by the following example in which we are presenting three components in a bar.

Example 14.2: Construct a subdivided bar chart for the three types of expenditures in dollars for a family of four for the years 1988, 1989, 1990 and 1991 as given as follows:

Year	Food	Education	Other	Total
1988	3000	2000	3000	8000
1989	3500	3000	4000	10500
1990	4000	3500	5000	12500
1991	5000	5000	6000	16000

Solution: The subdivided bar chart would be as follows:



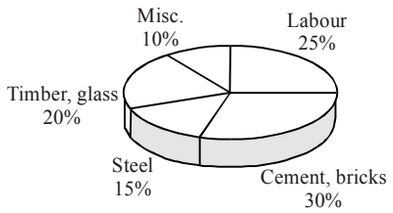
(ii) Pie chart: This type of diagram enables us to show the partitioning of a total into its component parts. The diagram is in the form of a circle and is also called a pie because the entire diagram looks like a pie and the components resemble slices cut from it. The size of the slice represents the proportion of the component out of the whole.

Example 14.3: The following figures relate to the cost of the construction of a house. The various components of cost that go into it are represented as percentages of the total cost.

Item	% Expenditure
Labour	25
Cement, Bricks	30
Steel	15
Timber, Glass	20
Miscellaneous	10

Construct a pie chart for the above data.

Solution: The pie chart for this data is presented as follows:



Pie charts are very useful for comparison purposes, especially when there are only a few components. If there are too many components, it may become confusing to differentiate the relative values in the pie.

(iii) Pictogram: Pictogram means presentation of data in the form of pictures. It is quite a popular method used by governments and other organizations for informational exhibitions. Its main advantage is its attractive value. Pictograms stimulate interest in the information being presented.

News magazines are very fond of presenting data in this form. For example, in comparing the strength of the armed forces of USA and Russia, they will simply make sketches of soldiers where each sketch may represent 100,000 soldiers. Similar comparison for missiles and tanks is also done.

Pictograms or pictographs are symbols of representation of the pictorial graphic system. Pictographs originated from prehistoric drawings on ancient rocks signifying an object or thing with its depiction. It is meant to convey, share or represent an idea or concept. A pictogram conveys meaning without

NOTES

words, with the help of its diagrammatic representation and are generally used in graphic systems and writings which have characters that appear in a pictorial form. It sometimes uses the representation of phonetic letters to form a base for cuneiform and even hieroglyphic writing.

Better known as 'icons', pictograms have been popularised with the use and familiarization of software's. Today the term is used widely and casually with the broad sweep of many icons representing things. The major role in getting familiarised with pictograms has been played by mobile devices and computers.

Pictograms are often used in writing, citing references, sign boards and as graphical systems where the characters illustrated are a representation of the natural self and to a considerable extent are pictorial in resemblance. These are used in various fields such as leisure, tourism and geography.

Herbert W. Kapitzki (Professor of Visual Communications, University of Arts, Berlin) defines the pictogram by its formal quality and abstractness. According to him, a pictogram is an iconic sign that depicts the character of what is being represented and through abstraction takes on its quality as a sign.

Otl Aicher (Ulm College of design) states that the pictogram must have the character of a sign and should not be an illustration.

Pictogram: Sign Language

The Pictogram is a friendly visual language that is developed for all classes of people and even those with no ability to speak, read or write.

- Pictograms can help one understand without help.
- With a pictogram representation, one can ask questions and get replies.

Here below are a few diagrammatic representations of pictograms:

Fig. 14.1 *A Common Utility Pictogram Chart*

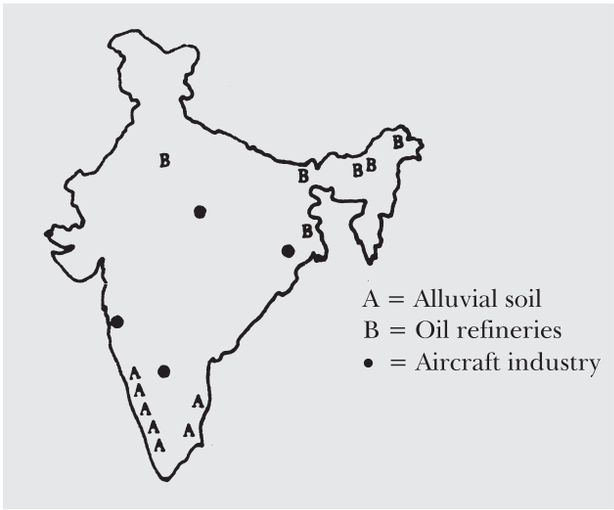
Source: *http://www.scratchinginfo.net/wp-content/uploads/2013/04/Modern-Pictograms.png*

Fig. 14.2 *A Pictogram Chart of Daily Use Signs*

Source: *http://kudesign.co.nz/studio/wp-content/uploads/pictograms.jpg*

(iv) Cartograms or Statiscal Maps

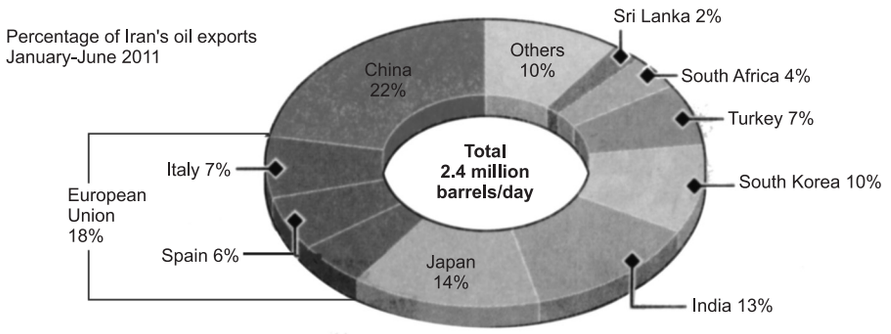
Cartograms are used to represent graphical distribution of data on maps. The various figures in different regions on maps are shown either by (i) shades or colours, (ii) dots or bars, (iii) diagrams or pictures, or (iv) by putting numerical figures in each geographical area.



The following maps show the location of a particular type of soil, refineries and aircraft industry in the country.

NOTES

Illustration



Buyers of Iranian Oil

Sources: *International Energy Agency, Wire Agencies*

Check Your Progress

1. What is the need for graphical or pictorial presentation of data?
2. What are bar diagrams?
3. Name the chart that shows the partitioning of a total into component parts.

14.3 GRAPHICAL PRESENTATION: HISTOGRAM, FREQUENCY POLYGON (GRAPHS) AND OGIVE

Graphic representation can be classified into the following:

(i) Histogram

(ii) Frequency polygon

(iii) Cumulative frequency curve (Ogive)

Each of these is briefly explained and illustrated.

(i) **Histogram:** A histogram is the graphical description of data and is constructed from a frequency table. It displays the distribution method of a data set and is used for statistical as well as mathematical calculations.

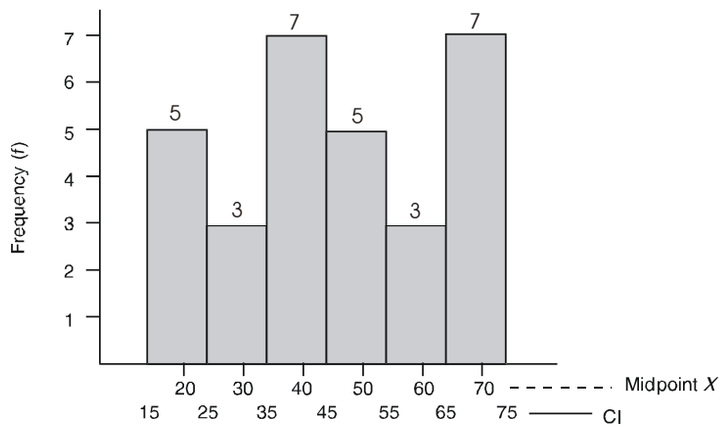
The word histogram is derived from the Greek word *histos* which means 'anything set upright' and *gramma* which means 'drawing, record, and 'writing'. It is considered as the most important basic tool of statistical quality control process.

In this type of representation, the given data are plotted in the form of a series of rectangles. Class intervals are marked along the *X*-axis and the frequencies along the *Y*-axis according to a suitable scale. Unlike the bar chart, which is one-dimensional, meaning that only the length of the bar is important and not the width, a histogram is two-dimensional in which both the length and the width are important. A histogram is constructed from a frequency distribution of a grouped data where the height of the rectangle is proportional to the respective frequency and the width represents the class interval. Each rectangle is joined with the other and any blank spaces between the rectangles would mean that the category is empty and there are no values in that class interval.

As an example, let us construct a histogram for our example of ages of 30 workers. For convenience sake, we will present the frequency distribution along with the mid-point of each interval, where the mid-point is simply the average of the values of the lower and upper boundary of each class interval. The frequency distribution table is shown as follows:

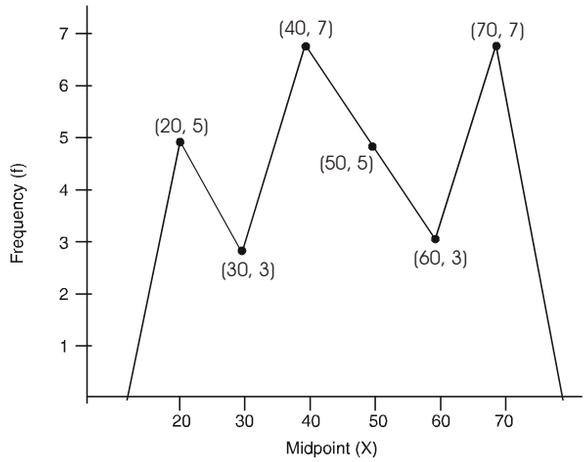
Class Interval (Years)	**Mid-point**	**(*f*)**
15 and upto 25	20	5
25 and upto 35	30	3
35 and upto 45	40	7
45 and upto 55	50	5
55 and upto 65	60	3
65 and upto 75	70	7

The histogram of this data would be shown as follows:



(ii) **Frequency polygon:** A frequency polygon is a line chart of frequency distribution in which either the values of discrete variables or mid-points of class intervals are plotted against the frequencies. These plotted points are joined together by straight lines. Since the frequencies generally do not start at zero or end at zero, this diagram as such would not touch the horizontal axis. However, since the area under the entire curve is the same as that of a histogram which is 100 per cent of the data presented, the curve can be enclosed so that the starting point is joined with a fictitious preceding point whose value is zero, so that the start of the curve is at horizontal axis and the last point is joined with a fictitious succeeding point whose value is also zero, so that the curve ends at the horizontal axis. This enclosed diagram is known as the frequency polygon.

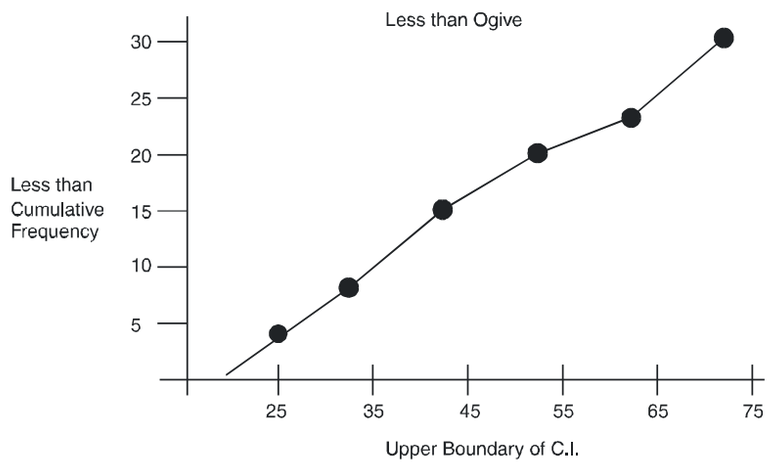
We can construct the frequency polygon from the preceding table as follows:



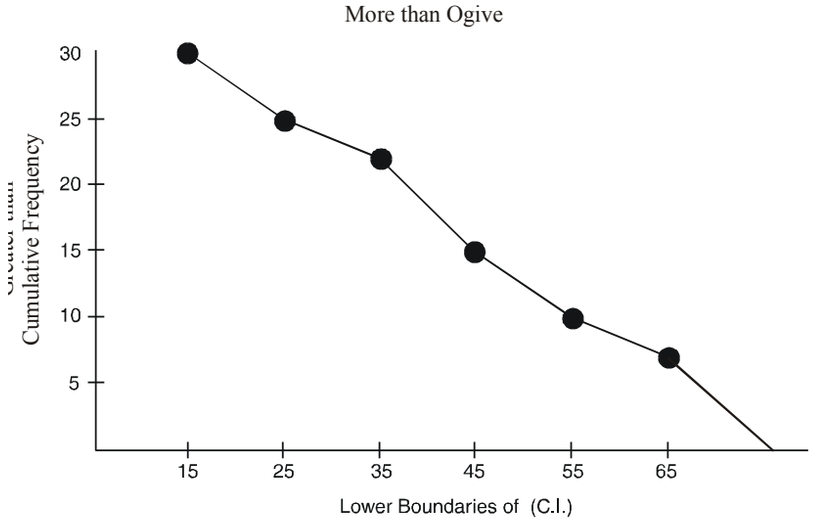
(iii) **Cumulative frequency curve (Ogive):** The cumulative frequency curve or ogive is the graphic representation of a cumulative frequency distribution. Ogives are of two types. One of these is less than and the other one is greater than ogive. Both these ogives are constructed based upon the following table of our example of 30 workers.

Class Interval (Years)	Mid-point	(*f*)	Cum. Freq. (Less Than)	Cum. Freq. (Greater Than)
15 and upto 25	20	5	5 (less than 25)	30 (more than 15)
25 and upto 35	30	3	8 (less than 35)	25 (more than 25)
35 and upto 45	40	7	15 (less than 45)	22 (more than 35)
45 and upto 55	50	5	20 (less than 55)	15 (more than 45)
55 and upto 65	60	3	23 (less than 65)	10 (more than 55)
65 and upto 75	70	7	30 (less than 75)	7 (more than 65)

(a) **Less than ogive:** In this case, less than cumulative frequencies are plotted against upper boundaries of their respective class intervals.



(b) **Greater than ogive:** In this case, greater than cumulative frequencies are plotted against the lower boundaries of their respective class intervals.



These ogives can be used for comparison purposes. Several ogives can be drawn on the same grid, preferably with different colours for easier visualization and differentiation.

NOTES

Although, diagrams and graphs are a powerful and effective media for presenting statistical data, they can only represent a limited amount of information and they are not of much help when intensive analysis of data is required.

Solved Problems

Example 14.4: Standard tests were administered to 30 students to determine their IQ scores. These scores are recorded in the following table.

120 115 118 132 135 125 122 140 137 127

129 130 116 119 132 127 133 126 120 125

130 134 135 127 116 115 125 130 142 140

(a) Arrange this data into an ordered array.

(b) Construct a grouped frequency distribution with suitable class intervals.

(c) Compute for this data:

- Cumulative frequency (<)
- Cumulative frequency (>)

(d) Compute:

- Relative frequency
- Cumulative relative frequency (<)
- Cumulative relative frequency (>)

(e) Construct for this data:

- A histogram
- A frequency polygon
- Cumulative relative ogive (<)
- Cumulative relative ogive (>)

Solution:

(a) The ordered array for this data is as follows:

115 115 116 116 118 119 120 120 122 125

125 125 126 127 127 127 129 130 130 132

132 132 133 134 135 135 137 140 140 142

NOTES

(b) Let there be six groupings, so that the size of the class interval be five. The frequency distribution is shown as follows:

Class Interval (CI)	*Frequency (f)*
115 to less than 120	6
120 ” ” ” 125	3
125 ” ” ” 130	8
130 ” ” ” 135	7
135 ” ” ” 140	3
140 ” ” ” 145	3

(c) The required elements are computed in the following table.

Class Interval	(*f*)	Cum. Freq.(<)	Cum. Freq. (>)
115–120	6	6 (less than 120)	30 (more than 115)
120–125	3	9 (less than 125)	24 (more than 120)
125–130	8	17 (less than 130)	21 (more than 125)
130–135	7	24 (less than 135)	13 (more than 130)
135–140	3	27 (less than 140)	6 (more than 135)
140–145	3	30 (less than 145)	3 (more than 140)

(d) The computed values of relative frequency, cumulative relative frequency (<) and cumulative relative frequency (>) are shown in the following table:

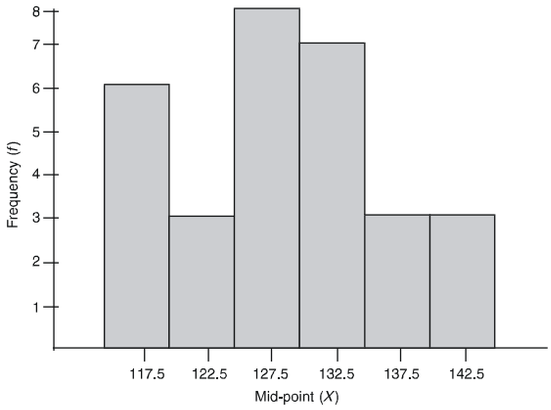
Class Interval	(*f*)	Rel. Freq.	Cum. Rel. Freq. (<)	Cum. Rel. Freq. (>)
115 and upto 120	6	6/30 or 20%	6/30 or 20% (<120)	30/30 or 100% >115)
120 and upto 125	3	3/30 or 10%	9/30 or 30% <125)	24/30 or 80% (>120)
125 and upto 130	8	8/30 or 26.7%	17/30 or 56.7% (<130)	21/30 or 70% (>125)
130 and upto 135	7	7/30 or 23.3%	24/30 or 80% (<135)	13/30 or 43.3% (>130)
135 and upto 140	3	3/30 or 10%	27/30 or 90% (<140)	6/30 or 20% (>135)
140 and upto 145	3	13/30 or 10%	30/30 or 100% (<145)	3/ 30 or 10% (>140)
Total = 30				

(e) Before we construct the histogram and other diagrams, let us first determine the midpoint (*X*) of each class interval.

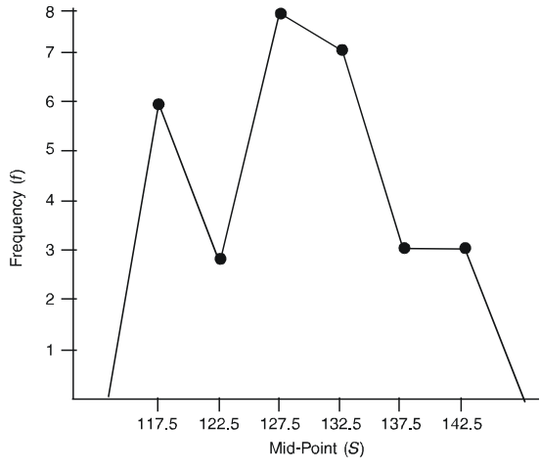
Class Interval	(*f*)	Mid-point (*X*)
115–120	6	117.5
120–125	3	122.5
125–130	8	127.5
130–135	7	132.5
135–140	3	137.5
140–145	3	142.5

NOTES

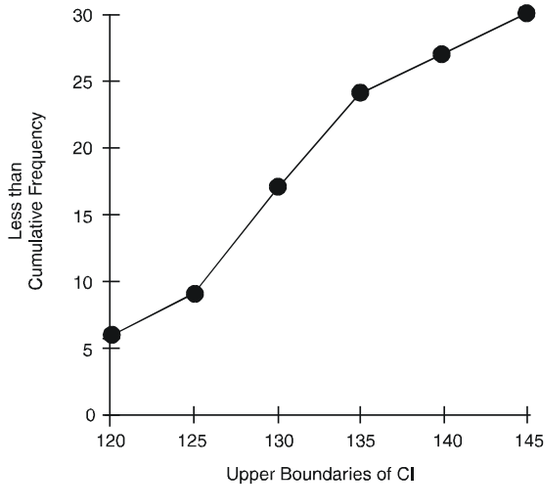
A histogram



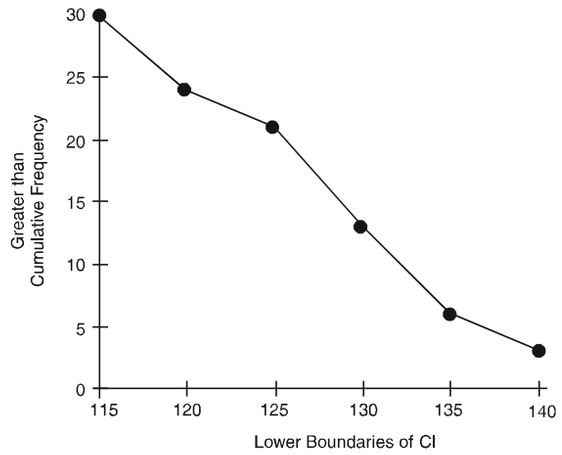
A frequency polygon



A cumulative frequency ogive (<)



A cumulative frequency ogive (>)



Example 14.5: Construct a stem and leaf display for the data of IQ scores presented in the preceding example.

Solution: The IQ scores of the given thirty students are presented in an ordered array, as follows:

115 115 116 116 118 119 120 120 122 125
125 125 126 127 127 127 129 130 130 132
132 132 133 134 135 135 137 140 140 142

The stem would consist of the first two digits and the leaf would consist of the last digit.

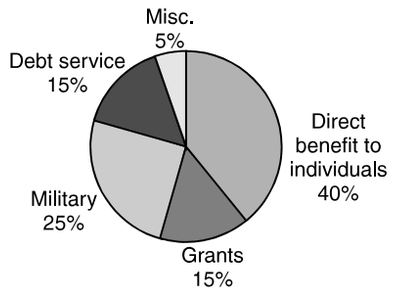
Stem	Leaves
11	5 5 6 6 8 9
12	0 0 2 5 5 5 6 7 7 7 9
13	0 0 2 2 2 3 4 5 5 7
14	0 0 2

Example 14.6: Suppose the Office of the Management and Budget (OMB) has determined that the Federal Budget for 2008 would be utilized for proportionate spending in the following categories. Construct a pie chart to represent this data.

Category	Per cent Allocation
Direct benefit to individuals	40
State, local grants	15
Military spending	25
Debt service	15
Misc. operations	5
	Total 100%

NOTES

Solution: The pie chart is presented as follows. Care must be taken so that the percentage allocation of budget is represented by the appropriate proportion of the pie.



14.3.1 Lorenz Curve

The Lorenz curve is a graphic method of measuring deviations from the average. It was devised by Dr Lorenz for measuring the inequalities in the distribution of wealth. But it can be applied with equal advantage for comparing the distribution of profits amongst different groups of business and such other things. It is a *cumulative percentage curve*. In it the percentages of items are combined with the percentage of such other things as wealth, profits or turn-over, etc.

In drawing a Lorenz curve the following steps are necessary:

1. The various groups of each variable should be reduced to percentage. Thus, if it is desired to show the distribution of income amongst the various groups of population of a country the various groups of population should be reduced in the form of percentages of total population; so also the incomes derived by these groups in terms of the total income of the country.
2. The two sets of the percentages obtained by step 1 should then the cumulated and cumulative percentages thus determined.
3. The cumulative percentages of these two variables should then be plotted along the axis of *Y* and axis of *X*. The scale along the axis of *Y* begins from zero at the point of intersection and goes upward up to 100, while the scale along the axis of *X* begins with 100 at the point of intersection and goes up to zero towards the right.
4. The points 100, 100 along the axis of *Y* and the points 0, 0 along the axis of *X* should be joined by a straight line. The line so obtained is

called the line of equal distribution, and serves as the basis for the determination of the extent to which the actual distribution deviates from the ideal distribution given by this line.

5. The actual data map now be plotted on this graph in the ordinary manner and the plotted points may be connected by means of a curve.

The farther the curve obtained under step 5 is from the line of equal distribution, the greater is the deviation.

Check Your Progress

4. What is a histogram?
5. What is the graphic representation of a cumulative frequency distribution called?

14.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Graphical or pictorial presentation of data makes the data easy to understand and interpret.
2. Bar diagrams are simply vertical lines where the lengths of the bars are proportional to their corresponding numerical values.
3. A pie chart shows the partitioning of a total into component parts.
4. A histogram is the graphical description of data and is constructed from a frequency table.
5. The graphic representation of a cumulative frequency distribution is called an ogive.

14.5 SUMMARY

- The data we collect can often be more easily understood for interpretation if it is presented graphically or pictorially. Diagrams and graphs give visual indications of magnitudes, groupings, trends and patterns in the data.
- The diagrams should be clear and easy to read and understand. Too much information should not be represented through the same diagram; otherwise, it may become cumbersome and confusing.

- Bars are simply vertical lines where the lengths of the bars are proportional to their corresponding numerical values. The width of the bar is unimportant but all bars should have the same width so as not to confuse the reader of the diagram.
- This type of diagram enables us to show the partitioning of a total into its component parts. The diagram is in the form of a circle and is also called a pie because the entire diagram looks like a pie and the components resemble slices cut from it.
- Pictogram means presentation of data in the form of pictures. It is quite a popular method used by governments and other organizations for informational exhibitions. Its main advantage is its attractive value. Pictograms stimulate interest in the information being presented.
- Pictograms or pictographs are symbols of representation of the pictorial graphic system. Pictographs originated from prehistoric drawings on ancient rocks signifying an object or thing with its depiction. It is meant to convey, share or represent an idea or concept.
- Better known as 'icons', pictograms have been popularised with the use and familiarization of softwares. Today the term is used widely and casually with the broad sweep of many icons representing things.
- The Pictogram is a friendly visual language that is developed for all classes of people and even those with no ability to speak, read or write.
- A histogram is the graphical description of data and is constructed from a frequency table. It displays the distribution method of a data set and is used for statistical as well as mathematical calculations.
- The word histogram is derived from the Greek word histos which means 'anything set upright' and gramma which means 'drawing, record, and 'writing'. It is considered as the most important basic tool of statistical quality control process.
- A frequency polygon is a line chart of frequency distribution in which either the values of discrete variables or mid-points of class intervals are plotted against the frequencies. These plotted points are joined together by straight lines.
- The cumulative frequency curve or ogive is the graphic representation of a cumulative frequency distribution. Ogives are of two types. One of these is less than and the other one is greater than ogive.

14.6 KEY WORDS

- **Pie charts:** They are basically circle charts, which are usually drawn for component-wise per cent data.
- **Component charts:** These charts are meant for exhibiting the changes in the components or parts of a given total in relative terms.
- **Pictogram:** These are symbols of representation of the pictorial graphic system.
- **Frequency polygon:** It is a line chart of frequency distribution in which either the values of discrete variables or mid-points of class intervals are plotted against the frequencies.

NOTES

14.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. What is graphical presentation of data? What should be taken care of while presenting data graphically?
2. How general is usage of the sign language? Give a few examples of pictograms from your daily life.
3. Discuss graphic representation of data in detail. List the forms of graphic representation.
4. State how 'less than ogive' is different from 'greater than ogive'?

Long Answer Questions

1. Discuss the diagrammatic representation of data.
2. Differentiate between a bar chart, pie chart and a pictogram. Explain the primary differences between them and their utility.
3. What is a frequency polygon? When plotted on the horizontal and vertical axis, why does the polygon not touch the horizontal axis? Explain with the help of an example.

14.8 FURTHER READINGS

Creswell, John W. 2002. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* London: Sage Publications.

Booth, Wayne, Gregory G. Colomb and Joseph M. Williams. 1995. *The Craft of Research.* Chicago: University of Chicago Press.

Kumar, B. 2006. *Research Methodology.* New Delhi: Excel Books.

Paneerselvam, R. 2009. *Research Methodology.* New Delhi: Prentice Hall of India.

Gupta, D. 2011. *Research Methodology*. New Delhi: PHI Learning Private Limited.

NOTES