CAIIB EXAMINATION

ADVANCED BANK MANAGEMENT

INDIAN INSTITUTE OF BANKING & FINANCE

macmillan education

ADVANCED BANK MANAGEMENT

INDIAN INSTITUTE OF BANKING & FINANCE

(ISO 9001:2015 Certified)
Kohinoor City, Commercial-II, Tower-1, 2nd & 3rd Floor,
Kirol Road, Off-L.B.S. Marg, Kurla-West,
Mumbai-400070

Established on 30th April 1928

MISSION

- To develop professionally qualified and competent bankers and finance professionals primarily through a process of education, training, examination, consultancy/counselling and continuing professional development programs.

VISION

- To be premier Institute for developing and nurturing competent professionals in banking and finance field.

OBJECTIVES

- To facilitate study of theory and practice of banking and finance.
- To test and certify attainment of competence in the profession of banking and finance.
- To collect, analyse and provide information needed by professionals in banking and finance.
- To promote continuous professional development.
- To promote and undertake research relating to Operations, Products, Instruments, Processes, etc., in banking and finance and to encourage innovation and creativity among finance professionals so that they could face competition and succeed.

COMMITTED TO PROFESSIONAL EXCELLENCE

Website: www.iibf.org.in

ADVANCED BANK MANAGEMENT

Indian Institute of Banking & Finance

macmillan
education

First published, 2023

MACMILLAN EDUCATION INDIA PRIVATE LIMITED
Bengaluru Delhi Chennai Kolkata Mumbai Ahmedabad Bhopal
Chandigarh Coimbatore Guwahati Hyderabad Lucknow Madurai
Nagpur Patna Pune Thiruvananthapuram Visakhapatnam Kochi
Bhubaneshwar Noida Sahibabad Hubli

ISBN: 978-93-5666-027-4

Published by Macmillan Education India Private Limited
(*formerly Macmillan Publishers India Private Limited*),
21, Patullos Road, Chennai 600002, India

ADVANCED BANK MANAGEMENT

First Edition (2023)
Authored, Revised & Updated by: Ms. Anuradha Hait (Basu), Head of Department of Statistics, Thakur College of Science & Commerce (Module A); Mr. Mrityunjay Kumar Gupta, Former Head-HR, BOI (Module B); Mr. P. D. Sankaranarayanan, Former AGM & Faculty, State Bank Academy (Module C); Mr. Shivkumar Sareen, Advisor, IPPB & Former GM & Chief Compliance Officer, BOI (Module D)

Vetted by: Mr. Butchi Babu, Former GM, BOI (Module A); Mr. G. S. Bhaskara Rao, Former DGM-Dev. HR, Central Bank of India (Module B); Mr. S. C. Bansal, Ex-faculty IIBF (Module C)

“This book is meant for educational and learning purposes. The author(s) of the book has/have taken all reasonable care to ensure that the contents of the book do not violate any copyright or other intellectual property rights of any person in any manner whatsoever. In the event the author(s) has/have been unable to track any source and if any copyright has been inadvertently infringed, please notify the publisher in writing for any corrective action.”

FOREWORD

Formal education will make you a living; self-education will make you a fortune.

–Jim Rohn

The banking sector, currently, is experiencing a transformation catalysed by digitalization and information explosion with the customer as the focal point. Besides, competition from NBFCs, FinTechs, changing business models, growing importance of risk and compliance, along with disruptive technologies, have contributed to this radical shift. Such an ever-evolving ecosystem requires strategic agility and constant upgradation of skill levels on the part of the Banking & Finance professionals to chart a clear pathway for their professional development.

The mission of the Indian Institute of Banking & Finance is to develop professionally qualified and competent bankers and finance executives primarily through a process of education, training, examination, counseling and continuing professional development programs. In line with the Mission, the Institute has been offering a bouquet of courses and certifications for capacity building of the banking personnel.

The flagship courses/examinations offered by the Institute are the JAIIB, CAIIB and the Diploma in Banking & Finance (DB&F) which have gained wide recognition among banks and financial institutions. With banking witnessing tectonic shifts, there was an imperative need to revisit the existing syllabi for the flagship courses.

The pivotal point for revising the syllabi was to ensure that, in addition to acquiring basic knowledge, the candidates develop concept-based skills in line with the developments happening in the financial ecosystem and to ensure greater value addition to the flagship courses and to make them more practical and contemporary. This will culminate in creating a rich pool of knowledgeable and competent banking & finance professionals who are capable of contributing to the sustainable growth of their organizations.

Keeping in view the above objectives, the Institute had constituted a high-level Syllabi Revision Committee comprising of members from public sector banks, private sector banks, co-operative banks and academicians. On the basis of the feedback received from various banks and changes suggested by the Committee, the syllabi of JAIIB & CAIIB have since been finalized.

The revised CAIIB syllabi will now have four compulsory subjects and one elective subject to be chosen from the five elective subjects. The subjects under the revised CAIIB Syllabi are:

Compulsory

1. Advanced Bank Management
2. Bank Financial Management
3. Advanced Business & Financial Management
4. Banking Regulations and Business Laws

Elective

1. Risk Management
2. Information Technology & Digital Banking
3. Central Banking
4. Human Resources Management
5. Rural Banking

A new module on Compliance has been introduced in Advanced Bank Management with Compliance, Corporate Governance and Audit becoming the focal point for a resilient banking system. New units covering Risks in Foreign Trade, GIFT-city etc have been added in the Bank Financial Management.

The new subject on Advanced Business & Financial Management will cover the management principles, the advanced concepts of Financial Management and emerging business solutions including Green Finance and Sustainable Financing.

The subject Banking Regulations & Business Laws (BRBL) is designed to familiarise the professionals with various laws concerning banking and finance with increased focus on case laws, court judgements covering different areas of banking and finance.

The elective subjects on Risk Management, Information Technology & Digital Banking and Rural Banking have also been thoroughly revised and will include new units to make the courses more contemporary. Insofar as the electives on Central Banking and Human Resources Management are concerned, new modules on NBFCs and Emerging Scenarios in HRM have been introduced respectively.

As is the practice followed by the Institute, a dedicated courseware for every paper/subject is published. The present courseware on Advanced Bank Management has now been authored in line with the revised syllabus for the subject. The book follows the same modular approach adopted by the Institute in the earlier editions/publications.

While the Institute is committed to revise and update the courseware from time to time, the book should, however, not be considered as the only source of information / reading material while preparing for the examinations due to rapid changes being witnessed in all the areas concerning banking & finance. The students have to keep themselves abreast with the current developments by referring to economic newspapers/journals, articles, books and Government / Regulators' publications / websites etc. Questions will be based on the recent developments related to the syllabus.

Considering that the courseware cannot be published frequently, the Institute will continue the practice of keeping candidates informed about the latest developments by placing important updates/Master Circulars/ Master Directions on its website and through publications like IIBF Vision, Bank Quest, etc.

The courseware has been updated with the help of Subject Matter Experts (SMEs) drawn from respective fields and vetted by practitioners to ensure accuracy and correctness. The Institute acknowledges with gratitude the valuable contributions rendered by the SMEs in updating/vetting the courseware.

We welcome suggestions for improvement of the courseware.

Mumbai
2023

Biswa Ketan Das
Chief Executive Officer

RECOMMENDED READING

The Institute has prepared comprehensive courseware in the form of study kits to facilitate preparation for the examination without intervention of the teacher. An attempt has been made to cover fully the syllabus prescribed for each module/subject and the presentation of topics may not always be in the same sequence as given in the syllabus.

Candidates are also expected to take note of all the latest developments relating to the subject covered in the syllabus by referring to Financial Papers, Economic Journals, Latest Books and Publications in the subjects concerned.

ADVANCED BANK MANAGEMENT
SYLLABUS

MODULE A: STATISTICS

Definition of Statistics, Importance & Limitations & Data Collection, Classification & Tabulation

Importance of Statistics; Functions of Statistics; Limitation or Demerits of Statistics; Definitions; Collection of Data; Classification and Tabulation; Frequency Distribution

Sampling Techniques

Random Sampling; Sampling Distributions; Sampling from Normal Populations; Sampling from Non-Normal Populations; Central Limit Theorem; Finite Population Multiplier

Measures of Central Tendency & Dispersion, Skewness, Kurtosis

Arithmetic Mean; Combined Arithmetic Mean; Geometric Mean; Harmonic Mean; Median and Quartiles; Mode; Introduction to Measures of Dispersion; Range and Coefficient of Range; Quartile Deviation and Coefficient of Quartile Deviation; Standard Deviation and Coefficient of Variation; Skewness and Kurtosis

Correlation and Regression

Scatter Diagrams; Correlation; Regression; Standard Error of Estimate

Time Series

Variations in Time Series; Trend Analysis; Cyclical Variation; Seasonal Variation; Irregular Variation; Forecasting Techniques

Theory of Probability

Mathematical Definition of Probability; Conditional Probability; Random Variable; Probability Distribution of Random Variable; Expectation and Standard Deviation of Random Variable; Binomial Distribution; Poisson Distribution; Normal Distribution; Credit Risk; Value at Risk (VaR); Option Valuation

Estimation

Estimates; Estimator and Estimates; Point Estimates; Interval Estimates; Interval Estimates and Confidence Intervals; Interval Estimates of the Mean from Large Samples; Interval Estimates of the Proportion from Large Samples

Linear Programming

Graphic Approach; Simplex Method

Simulation

Simulation Exercise; Simulation Methodology

MODULE B: HUMAN RESOURCE MANAGEMENT

Fundamentals of Human Resource Management

The Perspective; Relationship between HRM & HRD and their Structure and Functions; Role of HR Professionals; Strategic HRM; Development of HR Functions in India

Development of Human Resources

HRD and its Subsystems; Learning and Development – Role and Impact of Learning; Attitude Development; Career Path Planning; Self-Development; Talent Management; Succession Planning

Human Implications of Organisations

Human Behaviour and Individual Differences; Employees Behaviour at Work; Diversity at Workplace and Gender Issues; Theories of Motivation and their Practical Implications; 'Role' : Its Concept & Analysis

Employees' Feedback and Reward System

Employees' Feedback; Reward and Compensation System

Performance Management

Appraisal Systems; Performance Review and Feedback; Counselling; Competency Mapping and Assessment of Competencies; Assessment Centres; Behavioural Event Interview (BEI)

Conflict Management and Negotiation

Conflict: Concept & Definition; Characteristics of Conflict; Types of Conflicts; Reasons for Conflict; Different Phases of Conflict; Conflict Resolution; Conflict Management; Negotiation Skills for Resolution of Conflicts

HRM and Information Technology

Role of Information Technology in HRM; HR Information and Database Management; Human Resource Information System (HRIS); Human Resource Management System (HRMS); e–HRM; HR Research; Knowledge Management; Technology in Training; HR Analytics

MODULE C: CREDIT MANAGEMENT

Overview of Credit Management

Importance of Credit; Historical Background of Credit in India; Principles of Credit; Types of Borrowers; Types of Credit; Components of Credit Management; Role of RBI Guidelines in Bank's Credit Management

Analysis of Financial Statements

Which are the Financial Statements; Users of Financial Statements; Basic Concepts Used in Preparation of Financial Statements; Accounting Standards (AS); Legal Position Regarding Financial Statements; Balance Sheet; Profit and Loss Account; Cash Flow Statement; Funds Flow Statement; Projected Financial Statements; Purpose of Analysis of Financial Statements by Bankers; Rearranging the Financial Statements for Analysis; Techniques used in Analysis of Financial Statements; Creative Accounting; Related Party Transactions

Working Capital Finance

Concept of Working Capital; Working Capital Cycle; Importance of Liquidity Ratios; Methods of Assessment of Bank Finance; Working Capital Finance to Information Technology and Software Industry; Bills/Receivables Finance by the Banks; Guidelines of RBI for Discounting/Rediscounting of Bills by Banks; Trade Receivables Discounting System (TReDS); Non-Fund Based Working Capital Limits; Other Issues Related to Working Capital Finance

Term Loans

Important Points about Term Loans; Deferred Payment Guarantees (DPGs); Difference between Term Loan Appraisal and Project Appraisal; Project Appraisal; Appraisal and Financing of Infrastructure Projects

Credit Delivery and Straight Through Processing

Documentation; Third-Party Guarantees; Charge over Securities; Possession of Security; Disbursal of Loans; Lending under Consortium/Multiple Banking Arrangements; Syndication of Loans; Straight-Through Loan Processing or Credit Underwriting Engines

Credit Control and Monitoring

Importance and Purpose; Available Tools for Credit Monitoring/Loan Review Mechanism (LRM)

Risk Management and Credit Rating

Meaning of Credit Risk; Factors Affecting Credit Risk; Steps taken to Mitigate Credit Risks; Credit Ratings; Internal and External Ratings; Methodology of Credit Rating; Use of Credit Derivatives for Risk Management; RBI guidelines on Credit Risk Management; Credit Information System

Restructuring/Rehabilitation and Recovery

Credit Default/Stressed Assets/NPAs; Wilful Defaulters; Non-cooperative borrowers; Options Available to Banks for Stressed Assets; RBI Guidelines on Restructuring of Advances by Banks; Available Frameworks for Restructuring of Assets; Sale of Financial Assets

Resolution of Stressed Assets under Insolvency and Bankruptcy Code 2016

Definition of Insolvency and Bankruptcy; To Whom the Code is Applicable; Legal Elements of the Code; Paradigm Shift; Corporate Insolvency Resolution Process; Liquidation process; Pre-packed Insolvency Resolution Process for stressed MSMEs

MODULE D: COMPLIANCE IN BANKS & CORPORATE GOVERNANCE

Compliance Function in Banks

Compliance Policy; Compliance Principles, Process and Procedures; Compliance Programme; Scope of Compliance Function; Role & Responsibilities of Chief Compliance Officer (CCO)

Compliance Audit

Role of Risk Based Internal Audit and Inspection; Reporting Framework and Monitoring Compliance; Disclosure Requirements; Accounting Standards; Disclosures under Listing Regulations of SEBI

Compliance Governance Structure

Organisational Structure; Responsibility of the Board and Senior Management; Compliance Structure at the Corporate Office; Functional Departments; Compliance Structure at Field Levels; Internal Controls and its Importance

Framework for Identification of Compliance Issues and Compliance Risks

Compliance Issues; Compliance Risk; Inherent Risk and Control Risk; Independent Testing and Effective Audit Programme; Reporting Framework and Monitoring Compliance; Role of Inspection and Audit; Loan Review Mechanism/Credit Audit; What is Good Compliance

Compliance Culture and GRC Framework

How to Create Compliance Culture Across the Organisation; Governance, Risk and Compliance – GRC Framework; Benefits of an Integrated GRC Approach; Whistle-blower Policy; The Components of a Whistle-blower Policy; Reasons for Compliance Failures

Compliance Function and Role of Chief Compliance Officer in NBFCs

Framework for Scale Based Regulation for Non-Banking Financial Companies; Transition Path; Framework for Compliance Function and Role of Chief Compliance Officer in Non-Banking Financial Companies in Upper Layer and Middle Layer (NBFC-UL & NBFC-ML); Broad Contours of Compliance Framework in NBFCs

Fraud and Vigilance in Banks

Definition of Fraud; Definition of Forgery; Areas in which Frauds are Committed in Banks; Banking and Cyber Frauds; Fraud Reporting and Monitoring System; Vigilance Function in Banks; RBI Guidelines for Private Sector and Foreign Banks on Internal Vigilance

CONTENTS

MODULE A: STATISTICS

MODULE B: HUMAN RESOURCE MANAGEMENT

MODULE C: CREDIT MANAGEMENT

MODULE D: COMPLIANCE IN BANKS & CORPORATE GOVERNANCE

MODULE A

STATISTICS

Unit 1. Definition of Statistics, Importance & Limitations & Data Collection, Classification & Tabulation

Unit 2. Sampling Techniques

Unit 3. Measures of Central Tendency & Dispersion, Skewness, Kurtosis

Unit 4. Correlation and Regression

Unit 5. Time Series

Unit 6. Theory of Probability

Unit 7. Estimation

Unit 8. Linear Programming

Unit 9. Simulation

UNIT

1

Definition of Statistics, Importance & Limitations & Data Collection, Classification & Tabulation

STRUCTURE

1.0 OBJECTIVES

After studying this unit, you will be able to:

- Develop an understanding to use the proper methods to collect the data, employ the correct analyses, and effectively present the results.
- Familiarise with Data Management techniques by using Classification, Tabulation and Presentation.
- As Statistics is a crucial process behind how we make discoveries in science, make decisions based on data, and make predictions, collecting correct data and present it properly and making it ready for analysis and interpretation is very important. From this chapter, you will understand correctly getting the data ready to be analysed and interpreted.

1.1 INTRODUCTION

The word 'Statistics' has been derived from the Latin word 'statisticum', Italian word 'statistia' and German word 'statistik', each of which means a group of numbers or figures that represent some information of human interest. It was first used by professor Achenwell in 1749 to refer to the subject-matter as a whole. Achenwell defined statistics as the political science of many countries.

In the early years statistics is to be used only by the kings to collect facts about the state, revenue of the state or the people in the state of administrative or political purpose.

Gradually the use of statistics which means data or information has increased and widened. It is now used in almost in all the fields of human knowledge and skills like Business, Commerce, Economics, Social Sciences, Politics, Planning, Medicine and other sciences, physical as well as natural.

In many practical situations in life, we come across different types of data which are needed to be understood, analysed, compared and interpreted correctly. For example, in a college we need to analyse the data of marks obtained, in a hospital we need to analyse the data of number of patients having different diseases, rate of mortality, Different types of data need to be analysed in Economics, Government and Private organisations, Sports and in many other fields. Data mean information, which can be of two types – Qualitative and Quantitative. Statistics means quantitative or numerical data, which can be used for further calculations.

Statistical analysis of data can be comprised of four distinct phases:

1. **Collection of data:** In this first stage of investigation, numerical data is collected from different published or unpublished sources, primary or secondary.
2. **Classification and Tabulation of data:** The raw data collected is to be represented properly for further calculations. The raw data is divided into different groups or classes and represented in a form of a table.
3. **Analysis of data:** Classified and Tabulated data is analysed using different formulas and methods according to purpose of the study or investigation.
4. **Interpretation of data:** At the final stage, relevant conclusions are drawn after the data is thoroughly analysed

1.2 IMPORTANCE OF STATISTICS

Statistics is the subject that teaches how to deal with data, so statistical knowledge helps to use proper methods for collection of data, properly represent the data, use appropriate formula and methods to analyse correctly and effectively get the results and interpret the data. Applications of Statistics is important in every sphere of field – Business and economics, Medical, Sports, Weather forecast, Stock Market, Quality Testing, Government decisions and policies, Banks, Different educational and research organisations, etc.

1.2.1 Business and Economics

- In Business, the decision maker takes suitable policies and strategies based on information on production, sale, profit, purchase, finance, etc.
- By using the techniques of time series analysis, the businessman can predict the effect of a large number of variables with a fair degree of accuracy.
- By using 'Bayesian Decision Theory', the businessmen can select the optimal decisions to directly evaluate the payoff for each alternative course of action.
- In Economics, Statistics is used to analyse demand, cost, price, quantity, different laws of demand like elasticity of demand and consumer's maximum satisfaction which is determined on the basis of data pertaining to income and expenditure.

1.2.2 Medical

Statistics have extensive application in clinical research and medical field. Clinical research involves investigating proposed medical treatments, assessing the relative benefits of competing therapies, and establishing optimal treatment combinations.

1.2.3 Weather Forecast

Statistical methods, like Regression techniques and Time series analysis, are used in weather forecasting.

1.2.4 Stock Market

Statistical methods, like Correlation and Regression techniques, Time series analysis are used in forecasting stock prices. Return and Risk Analysis is used in calculation of Market and Personal Portfolios and Mutual Funds.

1.2.5 Bank

In banking industry, credit policies are decided based on statistical analysis of profitability, demand deposits, time deposits, credit ratio, number of customers and many other ratios. The credit policies are based on the application of probability theory.

1.2.6 Sports

Players use **statistics to identify or rectify their mistakes**.

A proper understanding of the statistics determines the success of a team or a single athlete.

1.3 FUNCTION OF STATISTICS

1. Statistics present the facts in definite form.
2. Statistics simplify complex data.
3. It provides a techniques of comparison.
4. Statistics study the relationship between two or more variables.
5. It helps in formulating policies.
6. It helps in forecasting outcomes.

1.4 LIMITATIONS OR DEMERITS OF STATISTICS

1. **Statistics do not deal with Individuals**
   Statistical methods can't be applied for individual values of the observations as for individual observation, there is no point of comparing anything or analysing anything. Statistics is the study of mass data or a group of observations and deals with aggregates of facts.
2. **Statistics does not study Qualitative Data**
   Statistical methods can't be applied for qualitative or non-numerical data. Statistics is the study of only of those facts which are capable of being stated in number or quantity.
3. **Statistics give Result only on an Average**
   Statistical methods are not exact. Generally, when we have large number of observations, it becomes difficult to handle it. A part of the data (sample) is collected for study and draw conclusion from, as a representative for the whole. As a result, the result obtained are not exactly same, had we analysed the whole data. The results are true only on an average in the long run.
4. **The results can be biased:** The data collection may sometime be biased which will make the whole investigation useless. Generally, this situation arises when data is handled by inexperienced or dishonest person.

1.5 DEFINITIONS

Population

It is the entire collection of observations (person, animal, plant or things which is actually studied by a researcher) from which we may collect data. It is the entire group we are interested in and from which we need to draw conclusions.

Example

1. If we are studying the weight of adult men in India, the population is the set of weights of all men in India.

2. If we are studying the grade point average of students of Mumbai University, the population is the set of GPA's of all students of Mumbai University.

Sample: Sometimes the population from which we need to draw conclusion is too large to study. At times collecting data from too large a population becomes time-consuming and expensive. To save time and money, generally a part of population is selected for study. A sample is a part (a group of units) of population which is representative of the actual population. By studying the sample, it is expected that valid conclusions are drawn about the whole group.

Example: The population for a study of infant health might be all children born in India in one particular year. The sample might be all babies born on one particular day in that year.

Data can be classified into two types, based on their characteristics. They are:

1. **Variates**
2. **Attributes**

A characteristic that varies from one individual to another and can be expressed in numerical terms is called **variate**.

Example: Prices of a given commodity, wages of workers, heights and weights of students in a class, marks of students, etc.

A characteristic that varies from one individual to another but can't be expressed in numerical terms is called an **attribute**.

Example: Colour of the ball (black, blue, green, etc.), religion of human, etc.

Quantitative or Numerical variables can be further classified as discrete and continuous. A variate which takes discrete or distinct value or in other words can take only a countable and usually finite number of values is called Discrete Variable.

Example: Number of members in a family, Number of accidents, Age in years.

A variate that can take any value within a range (integral/fractional) is called Continuous Variable.

Example: Percentage of marks, Height, Weight.

A **parameter** is a numerical value or function of the observations of the entire population being studied. A parameter is usually an unknown value that is fixed.

Example: Population mean, population median, population standard deviation, etc.

Since parameter is unknown, it has to be calculated or estimated from a sample. **Statistic** is used to estimate parameter. A statistic is a quantity or function of the observations of the sample of data. It is used to give information about unknown values in the corresponding population. For example, the sample mean is used to estimate the parameter population mean. Statistic is also called Estimator.

1.6 COLLECTION OF DATA

Researchers or investigators need to collect data from respondents. There are two types of data.

1.6.1 Primary Data

Primary data is the data which is collected directly or first time by the investigator or researcher from the respondents. Primary data is collected by using the following methods:

Direct Interview Method: A face to face contact is made with the informants or respondents (persons from whom the information is to be obtained) under this method of collecting data. The interviewer asks them questions pertaining to the survey and collects the desired information.

Questionnaires: Questionnaires are survey instruments containing short closed-ended questions (multiple choice) or broad open-ended questions. Questionnaires are used to collect data from a large group of subjects on a specific topic. Currently, many questionnaires are developed and administered online.

Census and sample survey

In a **census**, data about all individual units (e.g., people or households) are collected in the population. In a **survey**, data are only collected for a sub-part of the population; this part is called a **sample**. These data are then used to estimate the characteristics of the whole population. In this case, it has to be ensured that the sample is representative of the population in question. For example, the proportion of people below the age of 18 or the proportion of women and men in the selected sample of households has to reflect the reality in the total population.

1.6.2 Secondary Data

Secondary data are the Second hand information. The data which have already been collected and processed by some agency or persons and is collected for the second time are termed as secondary data. According to M. M. Blair, “Secondary data are those already in existence and which have been collected for some other purpose.” Secondary data may be collected from

existing records, different published or unpublished sources, like WHO, UNESCO, LIC, etc., various research and educational organisations, banks and financial places, magazines, internet, etc.

Distinction between primary and secondary data

1. The data collected for the first time is called Primary data and data collected through some published or unpublished sources is called Secondary data.
2. The primary data in the hands of one person can become secondary for all others.
   For example, the population census report is primary for the Registrar General of India and the information from the report is secondary for others.
3. Primary data are original as they are collected first time from the respondents directly or by preparing questionnaires. So they are more accurate than the secondary data. But the collection of primary data requires more money, time and energy than the secondary data. A proper choice between the two forms of information should be made in an enquiry.

1.7 CLASSIFICATION AND TABULATION

So, we learned about the different methods of collecting primary and secondary data. The raw data, collected in real situations are arranged randomly, haphazardly and sometimes the data size is very large. Thus, the raw data do not give any clear picture and interpreting and drawing any conclusion becomes very difficult. To make the data understandable, comparable and to locate similarities, the next step is classification of data. The method of arranging data into homogeneous group or classes according to some common characteristics present in the data is called **Classification**.

Example: The process of sorting letters in a post office, the letters are classified according to the cities and further arranged according to the streets.

Classification condenses the data by removing unimportant details. It enables us to accommodate large number of observations into few classes and study the relationship between several characteristics. Classified data is presented in a more organised way so it is easier to interpret and compare them, which is known as Tabulation.

There are four important bases of classifications:

1. **Qualitative Base:** Here the data is classified according to some quality or attribute such as sex, religion, literacy, intelligence, etc.
2. **Quantitative Base:** Here the data is classified according to some quantitative characteristic like height, weight, age, income, marks, etc.
3. **Geographical Base:** Here the data is classified by geographical regions or location, like states, cities, countries, etc. like population in different states of India.
4. **Chronological or Temporal Base:** Here the data is classified or arranged by their time of occurrence, such as years, months, weeks, days, etc. This classification is also called Time Series data.

Example: Sales of a company for different years.

Types of Classification

1. If we classify observed data for a single characteristic, it is known as One-way Classification. Ex: Population can be classified by Religion – Hindu, Muslim, Christians, etc.
2. If we consider two characteristics at a time to classify the observed data, it is known as a Two-way classification. Ex: Population can be classified according to Religion and sex.
3. If we consider more than two characteristics at a time in order to classify the observed data, it is known as Multi-way Classification. Ex: Population can be classified by Religion, sex and literacy.

1.8 FREQUENCY DISTRIBUTION

Frequency: If the value of a variable (discrete or continuous) e.g., height, weight, income, etc. occurs twice or more in a given series of observations, then the number of occurrences of the value is termed as the “frequency” of that value.

The way of representing a data in a form of a table consisting of the values of the variable with the corresponding frequencies is called “frequency distribution”. So, in other words, Frequency distribution is a table used to organise the data. The left column (called classes or groups) includes numerical intervals on a variable under study. The right column contains the list of frequencies, or number of occurrences of each class/group. Croxton and Cowden defined frequency distribution as a statistical table which shows the sets of all distinct values of the variable arranged in order of magnitude, either individually or in groups with their corresponding frequencies side by side

Intervals are normally of equal size covering the sample observations range.

1. **Class-limits or Class Intervals**
   A class is formed within the two values, class-limits or class-intervals. The lower value is called lower class limit or lower-class interval and the upper value is called upper class limit or class interval.
2. **Class Length or Class Width**
   The difference between the class' upper and lower class limit is called the length or the width of class.

   Class Length = Class Width = Upper Class Interval – Lower Class Interval

3. **Mid-Value or Class Mark**
   The mid-point of the class is called mid-value or class mark.

   Class Mark = (Lower class-limit + Upper Class limit)/2

4. **Types of Class Intervals**
   There are two types of class-interval. (i) Exclusive type, (ii) Inclusive type
   Class intervals like 0–10, 10–20; 500–1000, 1000–1500 are called exclusive types. Here the upper limits of the classes are excluded from the respective classes and put in the next class while considering the frequency of the respective class.
   For example, the value 15 is excluded from the class 10–15 and put in the class 15–20.
   Class intervals like 60–69, 70–79, 80–89, etc. are inclusive type. Here both the lower and upper class limits are included in the class-intervals while considering the frequency of the respective class, e.g., 60 and 69 are both included in the class 60–69.

Example 1: Inclusive Type of Class Intervals

Class-intervals	*Frequency*
0–4	5
5–9	7
10–14	12
15–19	8

Example 2: Exclusive Type of Class Intervals

Class-intervals	*Frequency*
0–10	5
10–20	7
20–30	12
30–40	8

5. **Class Boundaries**

   Inclusive classes can be converted to exclusive classes and the new class intervals are called class boundaries.

   Example 3: The classes 5–9, 10–14 can be converted to exclusive type of classes using the formula → New UCI = Old UCI + (10 – 9)/2 = 9 + 0.5 = 9.5. New LCI = Old LCI – (10 – 9)/2 = 5 – 0.5 = 4.5. So the class-boundaries are 4.5–9.5, 9.5–14.5, etc.

6. **Open-end Class Interval**

   In open-end class interval either the lower limit of the first class or upper limit of the last class or both are missing.

   Example 4: Below 10
   10–20
   20–30
   30–40
   Above 40

7. **Relative Frequency** $= \dfrac{\textit{frequency}}{\textit{Total frequency}}$

   Example 5: Relative frequency of the class interval = 20–30 in Example 2 is 12/32 = 0.375

8. **Percentage Frequency**

   Percentage Frequency = (Class frequency/Total Frequency) × 100

   Example 6: Percentage frequency of the class interval = 20–30 in Example 2 is (12/32) 100 = 37.5.

9. **Frequency Density**

   Frequency density of a class interval = Class frequency/Width of Class

Frequency Distribution is of two types.

1. **Discrete Frequency Distribution:** Variable takes distinct values.

Example

Problem 1: Assume that a survey has been made to know number of post-graduates in 10 families at random; the resulted raw data could be as follows.

0, 1, 3, 1, 0, 2, 2, 2, 2, 4

Solution: This data can be classified into a discrete frequency distribution.

Number of Post-graduates (x)	*Frequency*
0	2
1	2
2	4
3	1
4	1

2. **Continuous Frequency Distribution:** Variable takes values which are expressed in class intervals within certain limits.

Problem 2: Marks obtained by 20 students in an exam for 50 marks are given below–convert the data into continuous frequency distribution form.

18, 23, 28, 29, 44, 28, 48, 33, 32, 43, 24, 29, 32, 39, 49, 42, 27, 33, 28, 29.

Solution

Marks	*Frequency*
15–20	1
20–25	2
25–30	7
30–35	4
35–40	1
40–45	3
45–50	2

Problem 3: Following data reveals information about the number of children per family for 25 families. Prepare frequency distribution of number of children (say variable x, taking distinct values 0, 1, 2, 3, 4).

3	2	1	1	2
4	0	1	2	3
1	2	0	4	2
2	1	2	3	2
1	3	4	0	1

Solution: Frequency distribution of number of children in 25 families

Number of children	*Number of families*
0	3
1	7
2	8
3	4
4	3
Total	25

Problem 4: For the following frequency distribution, prepare cumulative frequency distribution of less than and greater than type

X	*1*	*2*	*3*	*4*	*5*	*6*
7	5	7	15	11	6	4

Solution

X	*Frequency*	*Less than cf*	*Greater than cf*
1	5	5	43 + 5 = 48
2	7	5 + 7 = 12	36 + 7 = 43
3	15	12 + 15 = 27	21 + 15 = 36
4	11	27 + 11 = 38	10 + 11 = 21
5	6	38 + 6 = 44	4 + 6 = 10
6	4	44 + 4 = 48	4
Total	48		

Problem 5: Following is the marks of 50 students. Prepare cumulative frequency distribution of both the types. Also find Relative Frequencies.

Marks	*No. of Students*
0–10	7
10–20	11
20–30	19
30–40	7
40–50	6
Total	50

Solution

Marks	*Cum Frequency*	*Cum Frequency more than*	*Relative Frequency*
0–10	7	43 + 7 = 50	7/50
10–20	7 + 11 = 18	32 + 11 = 43	11/50
20–30	18 + 19 = 37	13 + 19 = 32	19/50
30–40	37 + 7 = 44	6 + 7 = 13	7/50
40–50	44 + 6 = 50	6	6/50

Problem 6: For the following frequency distribution, obtain cumulative frequencies, relative frequencies and relative cumulative frequencies.

Class Interval	*30–50*	*50–70*	*70–90*	*90–110*	*110–130*	*130–150*
Frequency	8	15	25	16	7	4

Solution

Class interval	Freq	Cumu. Frequency		Relative Frequency	Relative Cumu. frequency	
		Less than	More than		Less than	More than
30–50	8	8	75	0.11	0.11	1.00
50–70	15	23	67	0.20	0.31	0.89
70–90	25	48	52	0.33	0.64	0.69
90–110	16	64	27	0.22	0.86	0.36
110–130	7	71	11	0.09	0.95	0.14
130–150	4	75	4	0.05	1,.00	0.05
Total	75			1.00		

Problem 7: Represent the information given below with a suitable table.

1400 candidate were medically examined in a fitness test, out of which 21% were girls. From the doctor's report, it was found that 396 males and 104 females were unfit. 40 % of the remaining males and 60% of the remaining females were in good health. The rest were declared as temporarily unfit.

Solution: Distribution of candidates according to Male/Female and Fitness

Fitness ↓ Sex →	Male	Female	Total
Good health	284	114	398
Temp unfit	426	76	502
Unfit	396	104	500
Total	1106	294	1400

Problem 8: Tabulate the information given below.

In 2020, out of total of 3000 workers in a factory, 2300 were skilled workers.

The number of woman employees were 300 out of which 250 were unskilled.

In 2021, the number of skilled workers was 2750 of which 2500 were men. The number of unskilled workers was 760 of which 300 were women.

Solution: Classification of workers according to sex and skill for the year 2020, 2021

Sex/ Skill Year	Men			Women			Total		
	Skilled	Unskilled	Total	Skilled	Unskilled	Total	Skilled	Unskilled	Total
2020	2250	450	2750	50	250	300	2300	700	3000
2021	2500	460	2960	250	300	550	2750	760	3510

Problem 9: Out of total number of 2000 candidates interviewed for employment in a company, 628 were from Pune and the rest from Nashik. Amongst the graduate from Pune, 350 were experienced and 80 were unexperienced. While the corresponding figures for undergraduates from Nashik were 615 and

52 respectively. The total number of in experienced candidates from Pune and Nashik were 175 and 192 respectively.

Present the above information in a suitable tabular form.

Solution: Distribution of candidate from Pune and Nashik according to Education and Experience

Education Experience City	Graduate			Undergraduate			Total		
	Exp	Inexp	Total	Exp	Inexp	Total	Exp	Inexp	Total
Pune	350	80	430	103	95	198	453	175	628
Nashik	565	140	705	615	52	667	1180	192	1372
Total	915	220	1135	718	147	865	1633	367	2000

Questions

1. The data of some worker's salary are given as 2300, 2400, 2500, 2100, 2000, 2000, 2300, 2800, 3000, 2300, 2700, 2400, 2500. If desired number of class intervals is 10, class width is

   (a) 100
   (b) 200
   (c) 300
   (d) 400

2. The largest and smallest values of a data are 60 and 40 respectively. If desired number of class intervals is 5, class width is

   (a) 25
   (b) 20
   (c) 4
   (d) 5

3. The class intervals where upper and lower limits are also in the class interval are called

   (a) Exclusive type
   (b) Inclusive type
   (c) Discrete type
   (d) Continuous type

4. The type of cumulative frequencies where the frequencies are added starting from the highest class to the lowest class are called _______________

   (a) Relative Frequency
   (b) Percentage Frequency
   (c) Less than Cumulative Frequency
   (d) Greater than Cumulative Frequency

5. The data classification, which is based on variables, like, demand, supply, height and weight is considered as

   (a) Qualitative data
   (b) Quantitative data
   (c) Time series data
   (d) Discrete data

6. The data which is classified or arranged by their time of occurrence, such as years, months, weeks, days, etc., is called

   (a) Time series data
   (b) Geographical data
   (c) Historical data
   (d) Both (a) and (c)

Answers

1. (a); 2. (c); 3. (b); 4. (d); 5. (b); 6. (d)

UNIT 2 Sampling Techniques

STRUCTURE

2.0 OBJECTIVES

The objectives of this unit are as follows:

- Learn to take a sample from an entire population and use it to describe the population
- Make sure the samples represent the population.
- Introduce the concepts of sampling distributions
- Understand the tradeoff between costs of larger samples and accuracy
- Introduce experimental design – sampling procedures.
- Estimation – data analysis and interpretation of sample data
- Testing of hypotheses – one-sample data
- Testing of hypotheses – two-sample data

2.1 INTRODUCTION

As you know, statistics is a tool used in business and finance. Statistics is an appreciated and maligned tool, depending on how it is used. We need statistical methods to reduce risk and uncertainty and improve our decision-making skills. Decision making in a situation involves collecting data and then using it for the future strategy. Usually. We cannot use the entire data because of the sheer size or numbers. Therefore, we take a sample and test it. For example, if a milk plant processes 1 lakh litres of milk every day, one cannot break open each packet and test the milk for quality. Here we take samples from each batch. If we want to do a market survey, we cannot interview each and every household. We take a representative sample. Thus, sampling becomes an integral tool of the quantitative methods we use. We take a sample, collect data from the sample, and attempt to generalise the whole data results.

Tea tasters at tea auctions are very highly paid employees of tea companies. They sample a small portion of the tea produced from the plantation before the auction. Food products are often tasted before being sold. Before buying Diwali sweets, you may take a small bite before buying it. Obviously, everything cannot be opened and tasted or tested as there would be nothing left to sell. We have to select a sample and test that only. If you want to write a report on why many people are migrating from India to Canada or Australia, contacting each Indian who migrated would be time consuming and expensive. So you would choose a sample and make a report accordingly. Thus, time and size are decisive factors which make it necessary to take business decisions on a sample. If your sample is properly chosen, your report would truly reflect the reasons of the entire population for migration. Of course, the best results would be available in any situation if we collected data from the entire population. Such complete enumeration or census is used in the population census carried out in our country every ten years.

Statisticians use the word ***population*** to refer not only to people but to all items that are to be studied. A sample is a part or subset of the population selected to represent the entire group. The word sample is used to describe a portion chosen from the population, in other words.

We can describe samples and populations using mean, median, mode, and standard deviation measures. When these terms describe a sample, they are called ***statistic*** and are not from the population but estimated from the sample. When these terms describe a population, they are called ***parameters***.

A statistic is a characteristic of a sample; a parameter is a population characteristic.

Conventionally, statisticians use lower case Roman letters to denote sample statistics and Greek or Capital letters to denote population parameters.

Table 2.1 lists these symbols.

TABLE 2.1 **Important Statistic Notation and Symbol**

	Population	*Sample*
Definition	Collection of all items	Part of the population
Characteristics	Parameters	Statistics
Symbols	Size – N	Size – n
	Mean – μ	Mean – $\overline{X}$
	Standard	Standard
	Deviation – σ	Deviation – σ

Types of sampling

The process of selecting respondents is known as 'sampling.' The units under study are called sampling units, and the number of units in a sample is called a sample size. There are two methods of selecting samples from populations: *non-random* or *judgement* sampling and *random* or *probability* sampling. In probability sampling, all the items in the population have a chance of being chosen in the sample. In judgement sampling, personal knowledge or opinions are used to identify the items from the population that are to be included in the sample. A sample selected by judgement sampling is based on someone's experience with the population. An oil drilling company would ask an experienced geologist to test different terrains or land beneath the sea before deciding where to explore for oil. Sometimes a judgement sample is used as a pilot or trial sample to decide how to take a random sample later. If we want to launch a new city newspaper, a pilot test of the paper can be launched at a judgement sample to see the response. But, the rigorous statistical analysis, which can be done with random probability samples, cannot be done with judgement samples. On the other hand, they are more convenient and can be used successfully even if we cannot test their validity. But, if a study uses judgement sampling and loses a significant degree of representativeness, it will have purchased convenience at a high price.

Biased samples

Suppose the Parliament is debating on the women's bill. You are asked to conduct an opinion survey. Because women are the most affected by the women's bill, you interviewed many women in different cities, towns and rural areas of India. Then you report that an overwhelming 95 per cent are in favour of reservation for women in Parliament.

Sometime later, the government has to take up the issue of Foreign Direct Investment (FDI) in print media. Since newspaper publishers are the most affected, you contact all of them, both national and regional, in India and report that the majority is not in favour of FDI in print media.

In both these cases, you picked a biased sample by choosing people who would have strong feelings on this issue. You have to have sound samples.

A report based on the data collected from such a biased sample would not truly reflect public opinion. If we follow random sampling, it is possible to statistically determine the reliability of the estimates obtained from the sample to avoid such errors.

2.2 RANDOM SAMPLING

There are four main types of random sampling.

1. Simple Random Sampling
2. Systematic Sampling
3. Stratified Sampling
4. Cluster Sampling

Simple Random Sampling

Simple Random Sampling selects samples by methods that allow each possible sample to have an equal probability of being picked and each item in the entire population to be included in the sample.

Suppose we have four teenagers participating in a talk show. We want a sample of two teenagers at a time to participate with the chat show host.

The following table illustrates the possible combinations of samples of teenagers, the probability of each sample being picked and the probability that each teenager will be in a sample.

Teenagers A, B, C, D

Possible samples of two teenagers: AB, AC, AD, BC, BD, CD

Probability of drawing these samples of two people is the same as below:

$$P(AB) = 1/6$$
$$P(AC) = 1/6$$
$$P(AD) = 1/6$$
$$P(BC) = 1/6$$
$$P(BD) = 1/6$$
$$P(CD) = 1/6$$

You would observe that any one student appears in 3 of the 6 possible samples. Therefore, probability of a particular student in the sample is:

$$P(A) = 1/2$$
$$P(B) = 1/2$$
$$P(C) = 1/2$$
$$P(D) = 1/2$$

The example illustrates a finite population of four teenagers. If we write A, B, C, and D on four identical slips of paper, fold the papers, and randomly pick any two, we get a sample. While picking up two paper slips, we may pick up one, keep it away, and pick another from the remaining three. This type is called **sampling without replacement**.

There is another way of doing it. Suppose after picking the first slip, we note the name on it and put the slip back in the lot, i.e replace the paper slip. Then we draw the second slip. There is a chance that we may draw the same student again. This is called **sampling with replacement.**

Theoretically, it is possible to have an infinite population. For example, the population of all prime numbers is infinite. Although many populations seem exceedingly large, no truly infinite population of physical objects actually exists. After all, given unlimited resources and time, one can enumerate any finite population. As a practical matter, we will use the term infinite population when we are talking about a population that could not be enumerated in a reasonable period of time. Thus, we use a theoretical concept of infinite population as an approximation of a large finite population.

How to do Random Sampling?

Suppose there are 100 employees in a company, and we wish to interview a randomly chosen sample of 10. We write the name of each employee on a slip of paper and deposit the slips in a box. After mixing them thoroughly, we draw 10 slips at random. The employees whose names are on these 10 slips, are our random sample.

This method of drawing a sample works well with small groups of people but presents problems with large populations. Also, add to this the problem of the slips of paper not being mixed well.

We can also select a random sample by using random numbers. These numbers can be generated either by a computer programmed to scramble numbers, or by a table of random digits.

The table below shows some random digits. These numbers have been generated by a completely random process. The probability that any one digit from 0 through 9 will appear is the same for each digit, and the probability of one sequence of digits occurring is the same as for any other sequence of the same length.

15,819	20,685	82,621	83,748	69,662
09,281	72,950	85,961	48,980	06,840
41,120	48,326	18,824	54,466	94,470
74,574	63,491	00,923	04,142	51,336
00,995	02,727	96,703	12,671	31,976
59,987	93,247	72,301	34,290	69,051
69,787	72,950	56,709	54,709	70,945
71,642	54,358	40,316	88,897	99,907
35,939	34,406	11,987	23,691	70,582
67,494	30,909	20,395	50,973	74,338

96,974	22,756	34,574	81,235	60,089
20,079	07,000	50,890	37,689	39,622
45,847	76,661	55,448	14,318	74,840
38,401	64,888	08,409	90,352	56,923
63,151	91,208	72,286	78,612	98,120
01,904	02,727	18,928	53,446	01,778
55,700	48,000	38,595	60,089	26,117
26,926	36,478	33,822	45,786	36,984
31,406	67,652	15,134	53,872	87,520
70,645	65,145	96,286	59,837	79,304
83,832	97,712	31,209	83,209	09,007
11,865	68,401	98,654	43,211	11,559
16,264	19,856	40,972	75,623	09,406
52,203	64,287	05,963	23,673	32,329
49,420	22,936	42,003	32,367	15,130
02,875	33,739	27,092	29,805	75,614
88,094	92,125	72,709	42,514	40,618
86,905	64,893	74,804	69,873	40,372
31,196	26,299	50,839	67,173	82,465

Let us see how to use this table. We assign each employee a number from 00 to 99, look up the table above and pick a systematic method of selecting two-digit numbers, like the first two digits. So, we have 15, 09, and so on till we get our 10 numbers.

Systematic Sampling

In systematic sampling, elements are selected from the population at a consistent level that is measured in time, order, or space. If we wanted to interview every twentieth student on a college campus, we would choose a random starting point in the first twenty names in the student directory and then pick every twentieth name after that.

Systematic sampling differs from simple random sampling in that each element has an equal chance of being selected, but each sample does not have an equal chance of being selected. This would have been the case if, in our earlier example, we had assigned numbers between 00 and 99 to our employees and then began to choose our sample of 10 by picking every tenth number beginning 1, 11, 21, 31, and so forth. Employees numbered 2, 3, 4 and 5 would have no chance of being selected together.

In systematic sampling, there is a probability of introducing an error into the sampling process. The system chosen may cause a problem. If we want to check the chances of people eating out on different days of the week, choose Friday. There is a higher likelihood of Friday as it is the beginning of the weekend, and we get a higher result.

Systematic sampling has some advantages. Even though systematic sampling may be inappropriate when the elements lie in a sequential pattern, this method may require less time and sometimes results in lower costs than the simple random sample method.

Stratified Sampling

To use stratified sampling, we divide the population into relatively homogenous groups, called strata. Then we use one of the following two approaches. Either we select at random from each stratum a specified number of elements corresponding to the proportion of that stratum in the population as a whole or, we draw an equal number of elements from each stratum and give weight to the results according to the stratum's proportion of the total population. With either approach, stratified sampling guarantees that every element in the population has a chance of being selected.

When to use Stratified Sampling

Stratified sampling is appropriate when the population is already divided into groups of different sizes, and we wish to acknowledge this fact. Example – middle class, upper class, lower middle class, etc. or according to age, race, sex or any other stratification. A jeans company may want to study which age group prefers to wear jeans the maximum. Thus the age groups may be 13 to 20, 20 to 30, 30 to 40, or 40 plus. The advantages of stratified samples are that when they are properly designed, they more accurately reflect the characteristics of the population from which they are chosen than do other kinds of samples.

Cluster Sampling

A well-designed cluster sampling procedure can produce a more precise sample at considerably less cost than simple random sampling. In cluster sampling, we divide the population into groups or clusters and then select a random sample of these clusters. We assume that these individual clusters are representative of the population as a whole. Suppose a market Research team is attempting to determine by sampling the average number of television sets per household in a large city. They could use a city map and divide the territory into blocks and then choose a certain number of blocks (clusters) for interviewing. Every household in each of these blocks would be interviewed.

Comparison of Stratified and Cluster Sampling

The population is divided into well-defined groups with both stratified and cluster sampling. We use stratified sampling when each group has a small variation within itself, but there is wide variation between the groups. We use cluster sampling in the opposite case – when there is considerable variation within each group, but the groups are essentially similar.

Basis of Statistical Inference: Simple Random Sampling

Systematic sampling, stratified sampling and cluster sampling attempt to approximate simple random sampling. All are methods that have been developed for their precision, economy or physical ease. However, as we do problems, we shall assume that the entire sample we are talking about are data based on simple random sampling. The process of making statistical inferences is based on the principles of random sampling. Once you understand the basics of random sampling, the same can be extended to other samples with some amendments which are best left to professional statisticians. It is important that you get a grasp of the concepts concerned.

2.3 SAMPLING DISTRIBUTIONS

In this section, we presume you are familiar with mathematical concepts such as mean, mode, median, standard deviation, etc. Each sample you draw from a population would have its own mean or measure of central tendency and standard deviation. Thus, the statistics we compute for each sample would vary and be different for each random sample taken.

Let us take an example.

We take a finite population of 5 young boys; A, B, C, D, and E, and collect data about their heights in centimetres. The data is shown in Table 2.2.

TABLE 2.2 Sampling Distribution Table

Boy	*A*	*B*	*C*	*D*	*E*
height (cm)	160	162	164	170	156

Now, if we take samples of size 3 (that is, select 3 boys in each sample), we will get 10 different samples. We list these samples, the corresponding data and their mean in Table 2.3.

TABLE 2.3 Samples, Their Data and Mean

No	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Sample	*ABC*	*ABD*	*ABE*	*BCD*	*BCE*	*ACD*	*ACE*	*ADE*	*BDE*	*CDE*
data	160	160	160	162	162	160	160	160	162	164
	162	162	162	164	164	164	164	170	170	170
	164	170	156	170	156	170	156	156	156	156
mean	162	164	159.33	165.33	160.66	164.66	160	162	162.66	163.33

From Table 2.3, you can see that sample mean for each sample is different. This collection of different values of the sample mean for samples of size 3, forms a distribution of sample means. This distribution has a mean. If we add all sample means in Table 2.3, and divide the sum by the number of samples, i.e., 10, we get 162.397 (say, 162.40)

Normally, we will be dealing with large populations. Hence the number of samples of a particular size is also very large.

Suppose we have to determine the proportion of sugar plants in a plantation affected by pest disease in samples of 100 plants taken from a very large plantation. We have taken a large number of these 100 item samples. If we plot a probability distribution of the proportions of infested plants in all these samples, we would see a distribution of the sample proportion. (The term proportion here refers to the proportion that is infected.) We could also have a sampling distribution of a proportion.

Sampling distribution is the distribution of all possible values of a statistic from all possible samples of a particular size drawn from the population.

Describing Sampling Distributions

Any probability distribution (and, therefore, any sampling distribution) can be partially described by its mean and standard deviation. Table 2.4 describes how different sampling distributions can be described.

TABLE 2.4 Different Sampling Distribution

S. No.	Population	Sample	Sample Statistic	Sampling Distribution
1.	Water in a River	10-one litre containers of water	Mean number of parts of mercury per million parts of water	Sampling distribution of the mean
2.	All professional basketball teams	Groups of 5 players	Median height	Sampling distribution of the median
3.	All parts produced in a manufacturing process	50 of each part	Proportion defective	Sampling distribution of the proportion

Each of the above sampling distributions can be partially described by its mean and standard deviation.

Concept of Standard Error

The standard deviation of the distribution of the sample means is called the *standard error of the mean. Similarly, standard error of the proportion is the standard deviation of the distribution of the sample proportions*.

The term standard error is used because it has a very specific connotation. For example, we take various samples to find the average heights of college girls across India and calculate the mean height for each sample. Obviously, there would be some variability in the observed mean. This variability in sampling statistics results from the *sampling error* due to chance. Thus, the difference between the sample and population means is due to the choice of samples.

Thus, the standard deviation of the sampling distribution of means measures the extent to which the means vary because of a chance error in the sampling process. Thus, *the standard deviation of the distribution of a sample statistic is known as the standard error of the statistic*.

Thus, a standard error indicates the size of the chance error and the accuracy we are likely to get if we use the sample statistic to estimate a population statistic. Thus, a mean with a smaller standard deviation is a better estimator than one with a higher standard deviation.

Understanding sampling distributions allows statisticians to take both meaningful and cost-effective samples. Because collecting large samples is expensive, decision-makers should always aim for the smallest sample which gives the most reliable results.

Sampling Distributions

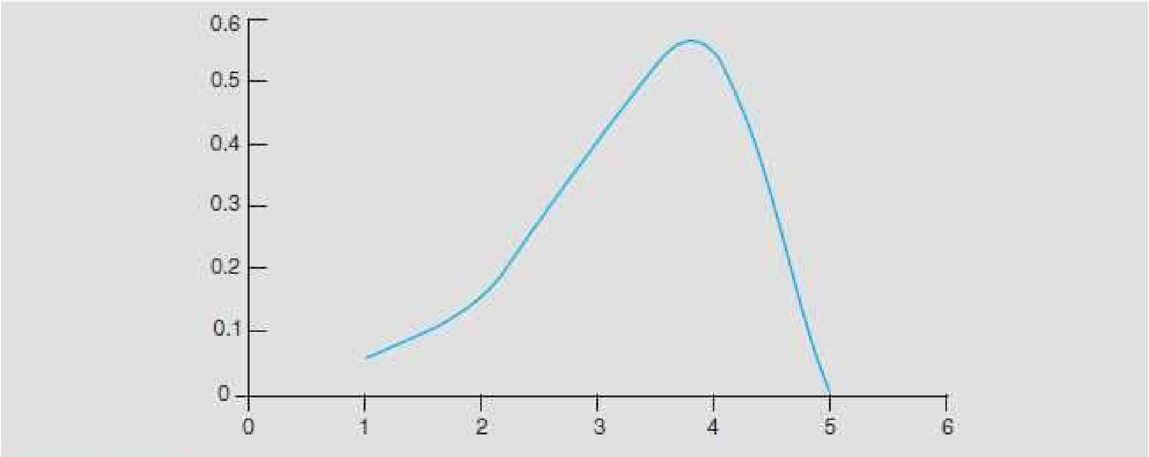

FIGURE 2.1(a) Population Distribution

Figures 2.1(a) and (b) will help you to understand sampling distributions. There are three parts to this illustration. Figure 2.1(a) illustrates a population distribution. Assume that this population is all the non-performing assets of a large bank, and this distribution is the number of years the assets have been classified as nonperforming in the balance sheet. This distribution of "number of years" has a mean **m** (or **µ** pronounced mu) and a standard deviation of **s** (or **σ** pronounced sigma).

Suppose we can take all possible samples of 10 non-performing assets (NPAs) from the population distribution. There would be too many, so we possibly cannot take all. Next, we would calculate the mean and standard deviation for each one of these samples, as represented in Fig. 2.1(b). As a result, each sample would have its own mean and its standard deviation. All the individual sample means would not be the same as the population mean. They would tend to be near the population mean, but only rarely would they be exactly that value.

As the last step, we would produce a distribution of all the means from every sample that could be taken. This distribution, called the sampling distribution of the mean, is illustrated in Fig 2.1(c).

This distribution of the sample means (the sampling distribution) would have its mean μx and standard deviation or error Sx.

In statistical terminology, the sampling distribution is obtained by taking all the possible samples of a given size is the theoretical sampling distribution.

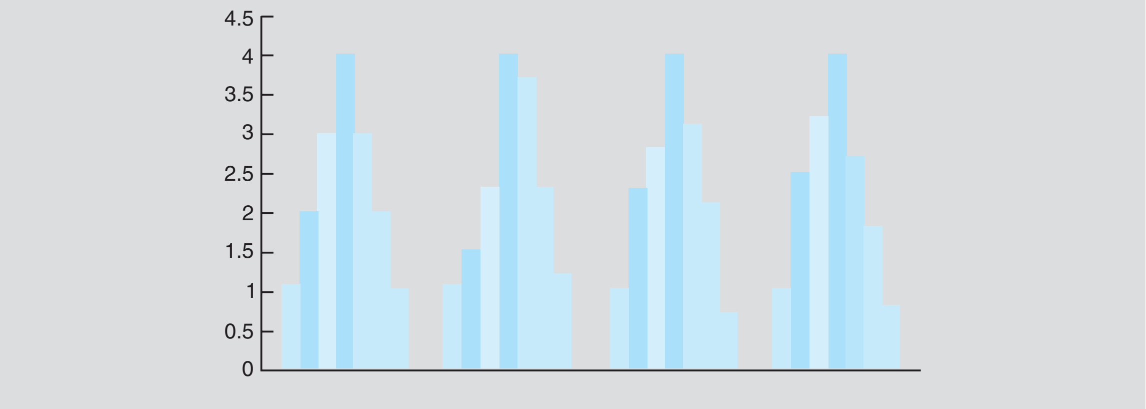

FIGURE 2.1(b) Sampling Frequency Distribution

Figure 2.1(c) shows an example of such a distribution. In practice, the size and character of most real-life populations prohibit us from taking all the possible samples from a population distribution. Fortunately, formulas are developed for estimating the characteristics of these theoretical sampling distributions, making it unnecessary to collect large numbers of samples. In most cases, decision-makers take only one sample from the population, calculate some statistics from the sample, and from those statistics estimate something about the parameters of the entire population. We shall show this shortly.

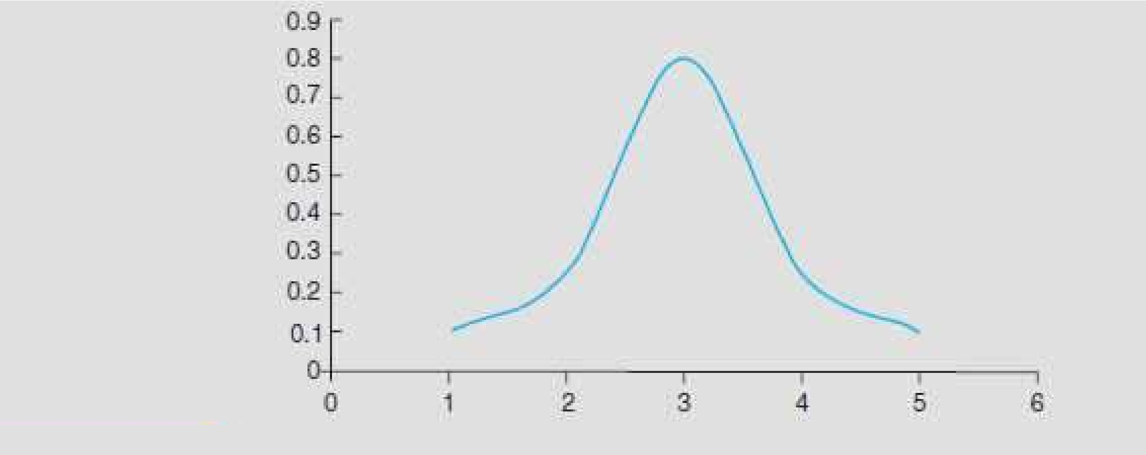

FIGURE 2.1(c) Sampling Distribution of the Mean

Now, we shall discuss the sampling distribution of the mean in the following examples. Once you understand the formulae and concepts here, the same can be applied to other sampling distributions as well.

2.4 SAMPLING FROM NORMAL POPULATIONS

Let us go back to the example of taking samples of size 3 from a population of size 5.

From Table 2.2, we find the mean μ of the population

$$\mu = 162.40$$

In Table 2.3, we have the sample means, x for the 10 samples.

The mean of these sample means μ_x = 162.40; which is the same as population mean. This is not a coincidence. Whenever we use simple random sampling, the mean of the sample means is the same as the population mean.

Now, from Table 2.2, we see that the values range from 156 to 170. From Table 2.3, we see that the sample means range from 159.33 to 165. Thus, sample means have a smaller spread than the population.

Thus, the sampling distribution is also normal if our population is normally distributed.

Further, (i) the mean of the sampling distribution is the same as the population mean μ x= μ and (ii) the standard deviation of the sampling distribution is equal to the population standard deviation divided by the square root of the sample size.

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

Now, suppose we increase our sample size from 3 to 4. This would not change the standard deviation of the items in the original population. But with samples of size 4, we have decreased the standard deviation of a sampling distribution. The properties are also explained below.

	Property	Equation
Properties of Sampling Distribution of Mean, when Population is normally Distributed	Sampling Distribution has Mean equal to Population Mean.	$\mu_{\bar{x}} = \mu$
	Sampling distribution has Standard Deviation equal to Population Standard Deviation divided by Square root of Sample size	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

An example will further illustrate these properties. A bank calculates that its individual savings accounts are normally distributed with a mean of Rs. 2,000 and a standard deviation of Rs. 600. If the bank takes a random sample of 100 accounts, what is the probability that the sample mean will lie between Rs. 1,900 and Rs. 2,050? This is a question about the sampling distribution of the mean; therefore, we must first calculate the standard error of the mean. In this case, we shall use the equation for the standard error of the mean designed for a situation in which the population is infinite (later, we shall introduce an equation for finite populations):

Since the sampling distribution is a normal distribution, let us first see how to find the probabilities associated with a normal distribution.

The probabilities associated with a standard normal variable, with mean 0, and standard deviation 1, are available in the form of a Table. See Annexure at the end of this unit. We first convert a value in our sample to a standard normal value by using the equation

$$Z = \frac{1}{\sigma_{\bar{x}}}[\bar{x} - \mu]$$

The Annexure gives us the probability that the variable z is between 0 and the given value. For example, if we want to know the probability that the variable z is between 0 and 1.54, we get the answer from the Probability Table in the Annexure as 0.4382.

The probability that the variable is between –0.96 and 0 is given as 0.3315. (Note that we ignore the minus sign.)

The probability that the variable is between –0.96 and 1.54 is then 0.4382 + 0.3315 = 0.7697.

The probability that the variable is between –0.96 and –1.54 is 0.4382 – 0.3315 = 0.1067.

The probability that the variable is between 0.96 and 1.54 is also 0.4382 – 0.3315 = 0.1067.

So if the values have the same sign (plus or minus), we subtract the smaller value in the Table from the bigger one.

On the other hand, if the values have opposite signs, we add the values obtained from the Table.

Standard Error of the Mean for Infinite Populations

Standard error of the mean = $\frac{\sigma}{\sqrt{n}}$

where σ = population standard deviation

n = sample size

Applying this to our example, we get

Standard error of the mean = $\frac{600}{10} = 60$

Using the equation
we get 2z values.

$$z = \frac{(\bar{x} - \mu)}{\sigma_{\bar{x}}}$$

For $\bar{x}$ = Rs. 1,900; $z = \frac{(1900 - 2000)}{60} = -1.67$

For $\bar{x}$ = Rs. 2,050; $z = \frac{(2050 - 2000)}{60} = 0.83$

The annexure table gives us an area or probability of 0.4525 corresponding to a z value of –1.67. It gives the probability of 0.2967 for a z value of 0.83. If we add these two together, we get 0.7492 as the total probability that the sample mean will lie between Rs. 1,900 and Rs. 2,050.

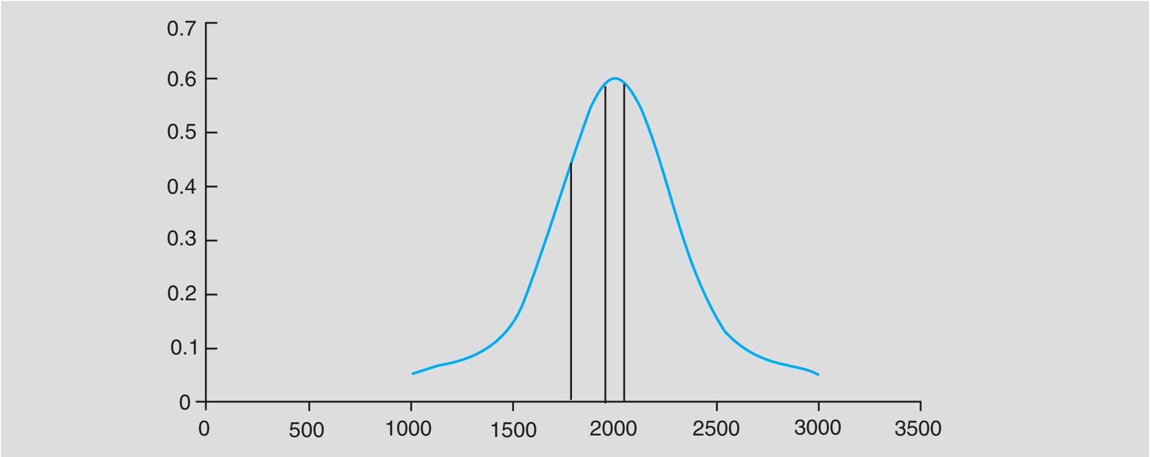

FIGURE 2.2 Probability of Sample Mean Lying between Rs 1900 and Rs 2050

The areas also show the probabilities under the probability curve shown in Fig. 2.2.

2.5 SAMPLING FROM NON-NORMAL POPULATIONS

In the preceding section, we stated that when the population is normally distributed, the sampling distribution of the mean is also normal. But, we come across many populations that are not normally distributed. How does the sampling distribution of the mean behave when the population from which the samples are drawn is not normal? An illustration will help us answer this question.

Consider the data in Table 2.5, concerning five motorcycle owners and the life of the Tyres. Because only five people are involved, the population is too small to be approximated by a normal distribution. Let us take all of the possible samples of the owners in groups of three, compute the sample means ($\bar{x}$) list them, and compute the mean of the sampling distribution ($\mu_{\bar{x}}$). We have done this in Table 2.5. These calculations show that even in a case in which the population is not normally distributed, ($\mu_{\bar{x}}$), the mean of the sampling distribution is still equal to the population mean, μ.

TABLE 2.5 Experience of Five Motorcycle Owners with Life of Tyres

Owner	*Chetan (C)*	*Dinesh (D)*	*Eswar (E)*	*Feroz (F)*	*George (G)*
Tyre Life in months	3	3	6	9	15

Total life = 36 months
Mean = 36/5 = 7.2 months

TABLE 2.6 Calculation of Sample Mean Tyre Life with $n = 3$

Samples of Three	*Sample Data*	*Sample Mean*
EFG	6 + 9 + 15	10
DFG	3 + 9 + 15	9
DEG	3 + 6 + 15	8
DEF	3 + 6 + 9	6
CFG	6 + 6 + 9	7
CEG	3 + 6 + 15	8
CEF	3 + 6 + 9	6
CDF	3 + 3 + 9	5
CDE	3 + 3 + 9	5
CDG	3 + 3 + 15	8

Total = 72 months
Mean = 7.2 months

Now, look at Fig. 2.3(a), which shows the population distribution of tyre lives for the five motorcycles owners, a distribution that is anything but normal in shape. In Fig. 2.3(b), we show the sampling distribution of the mean for a sample size of three, taking the information from Table 2.6. Notice the difference between the probability distributions in Figs. 2.3(a) and 2.3(b). In Figure 2.3(b), the distribution looks a little more like the bell shape of the normal distribution.

If we repeat this exercise and enlarge the population size to 40, we could take samples of different sizes. Then plot the sampling distributions of the mean that would occur for the different sizes. This will show quite dramatically how quickly the sampling distribution of the mean approaches normality, regardless of the shape of the population distribution.

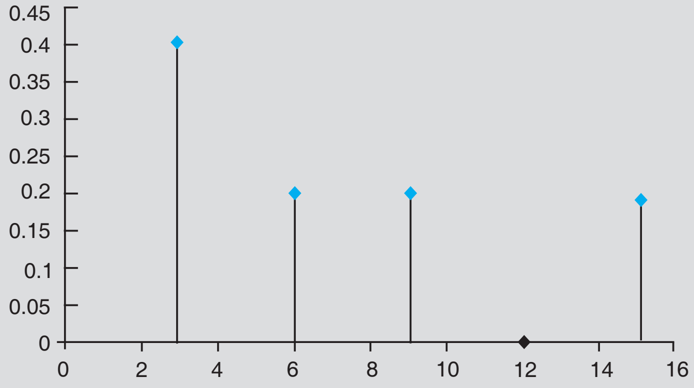

FIGURE 2.3 (a) Problem Distribution

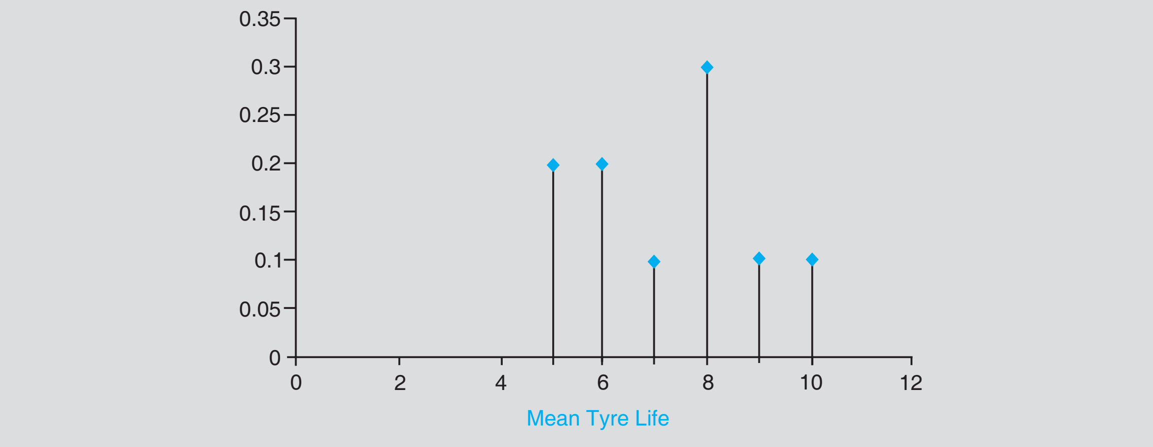

FIGURE 2.3 (b) Distribution of Mean Tyre Life

2.6 CENTRAL LIMIT THEOREM

The example in the above Table and the probability distribution in the above graphs tell us many things. First, the mean of the sampling distribution of the mean will equal the population mean regardless of the sample size, even if the population is not normal. As the sample size increases, the sampling distribution of the mean will approach normality, regardless of the shape of the population distribution.

The Central Limit Theorem is the relationship between the shape of the population distribution and the shape of the sampling distribution of the mean. The central limit theorem is perhaps the most important in all statistical inference. It assures us that the sampling distribution of the mean approaches normal as the sample size increases.

1. Actually, a sample does not have to be very large for the sampling distribution of the mean to approach normal.
2. Statisticians use the normal distribution as an approximation to the sampling distribution whenever the sample size is at least 30, but the sampling distribution of the mean can be nearly normal with samples of even half the size.
3. The significance of the central limit theorem is that it permits us to use sample statistics to make inferences about population parameters without knowing anything about the shape of the frequency distribution of that population

Let's illustrate the use of the central limit theorem. The distribution of annual earnings of all bank tellers with five years' experience is as shown below in Fig. 2.4. This distribution has a mean of Rs. 19,000 and a standard deviation of Rs. 2,000. If we draw a random sample of 30 tellers, what is the probability that their earnings will average more than Rs. 19,750 annually? In Fig. 2.4 we show the sampling distribution of the means that would result and highlight the area representing 'earnings over Rs. 19,750.'

Our first task is to calculate the standard error of the mean from the population standard deviation as follows:

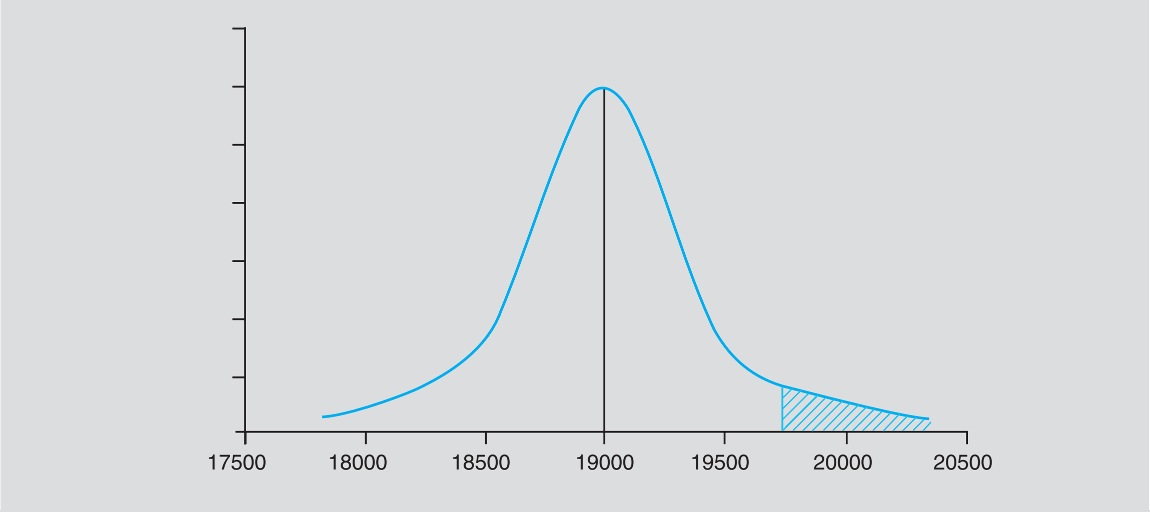

FIGURE 2.4

Standard error of the mean $\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$

$$= \frac{2000}{\sqrt{30}}$$

$$= \frac{2000}{5.477} = \text{Rs. } 365.16$$

Because we are dealing with a sampling distribution, we must now use the equation for z value and the Standard Normal Probability Distribution (App. Table $z = (\bar{x} - \mu)/\sigma_{\bar{x}}$

For $\bar{x} = \text{Rs. } 19750;$

$$z = \frac{(19750 - 19{,}000)}{365.16}$$

$$= \frac{750}{365.16}$$

$$= 2.05$$

Annexure Table gives us the probability of 0.4798 for a *z* value of 2.05. We show the corresponding area in Fig. 2.4 as the area between the mean and Rs. 19,750. Since half or 0.5000 of the area under the curve lies between the mean and the right-hand tail, the shaded area must be

0.5000	(Area between the mean and the right-hand tail)
– 0.4798	(Area between the mean and 19,750)
0.0202	(Area between the right-hand tail and 19,750)

Thus, we have determined that there is slightly more than a 2 per cent chance of average earnings being more than Rs. 19,750 annually in a group of 30 tellers.

The central limit theorem is one of the most powerful concepts in statistics, which states that the distribution of sample means tends to be a normal distribution. This is true regardless of the shape of the population distribution from which the samples were taken.

Result 1: If X ~ Bin(n, p), then $Z = \frac{X-np}{\sqrt{npq}}$ tends to standard Normal Deviation as $n \to \infty$

Result 2: If X ~ Pλ), then $Z = \frac{X-\lambda}{\sqrt{\lambda}}$ tends to standard Normal Deviation as *sample size* $\to \infty$

Examples

1. A sample of 25 observations from a normal distribution has a mean of 98.6 and a standard deviation of 17.2.

   (a) What is $P(92 < \bar{x} < 102)$?

   (b) Find the corresponding probability given a sample of 36.

Solution:

(a) $N = 25, \mu = 98.6, \sigma = 17.2,$

$\sigma x = \sigma / \sqrt{n} = 17.2 / \sqrt{25} = 3.44$

$P\,(92 < \bar{x} < 102) = p[(92 - 98.6)/3.44 < (\bar{x} - \text{m})/\text{s}_x < (102 - 98.6)/3.44]$

$= P\,(-1.72 < z < 0.99) = 0.4573 + 0.3389 = 0.7962$

(b) $n = 36, \quad \sigma_{\bar{x}} = \sigma / \sqrt{n} = 17.2 / \sqrt{36} = 2.87$

$P\,(92 < \bar{x} < 102) = p[(92 - 98.6)/2.87 < (\bar{x} - \mu)/\sigma_x < (102 - 98.6)/2.87]$

$= P(-2.30 < z < 1.18) = .4893 + 0.3810 = 0.8703$

2. Kamala, an auditor for a large credit card company, knows that, on average, the monthly balance of any given customer is Rs. 112, and the standard deviation is Rs. 56. If Kamala audits 50 randomly selected accounts, what is the probability that the sample average monthly balance is

   (a) Below Rs. 100?

   (b) Between Rs. 100 and Rs. 130?

Solution: The sample size of 50 is large enough to use the central limit theorem

$\mu = 112, \ \sigma = 56, \ n = 50, \ \sigma_{\bar{x}} = 56 / \sqrt{50} = 7.920$

(a) $P(\bar{x} < 100) = P[(\bar{x} - \mu)/\sigma_{\bar{x}} < (100 - 112)/7.920]$

$= P(z < -1.52) = 0.5 - 0.4357 = 0.0643$

(b) $P(100 < x < 130) = P[(100 - 112)/7.920 < (\bar{x} - \mu)/\sigma_{\bar{x}} < (130 - 112)/7.920]$

$= P(-1.52 < z < 2.27) = 0.4357 + 0.4884 = 0.9241$

3. It has been found that 2% of the tools produced by a certain machine are defective. What is the probability that in a shipment of 400 tools, 3% or more defective?

   The sample size of 400 is large enough to use the central limit theorem.

Solution: X ~ P$\left(\lambda = np = 0.02 * 400 = 8\right)$, then Z = $\dfrac{X-\lambda}{\sqrt{\lambda}}$ tends to standard Normal Deviation

3% of 400 = 12

P(X > 12) = P[(x – 8)/√8 > (12 – 8))/√8 = P([(x – 8)/2.82 > 1.43) = 0.5 – 0.4236 = 0.0764

4. A coin is tossed 700 times. Using Normal approximation find the probability of getting number of Heads between 280 and 375.

   The sample size of 700 is large enough to use the central limit theorem.

Solution: X~ Bin$\left(n = 700, p = 0.5\right)$, then Z = $\dfrac{X-np}{\sqrt{npq}} = \dfrac{X-350}{13.22}$ tends to standard Normal Deviation

P(280 < X < 375) = P[(280 – 350)/13.22 < (X – 350)/13.22 < (375 – 350)/13.22)]

= P[–5.29 < Z < 1.89] = 0.5 + 0.4706 = 0.9706

An Important Consideration in Sampling: The Relationship between Sample Size and Standard Error

We saw earlier in this chapter that the standard error, σ_x is a measure of the dispersion of the sample means around the population mean. If the dispersion decreases (if σ_x becomes smaller), then the values taken by the sample mean tend to cluster more closely around m. Conversely, if the dispersion increases (if σ_x becomes larger), the values taken by the sample mean tend to cluster less closely around m. We can think of this relationship this way: *As the standard error decreases, the value of any sample mean will probably be closer to the value of the population mean. As the standard error decreases, the precision with which the sample mean can be used to estimate the population mean, increases.*

If we refer to Equation, we can see that as *n* increases, s_x decreases. This happens because, in an Equation, a larger denominator on the right side would produce smaller s_x on the left side. Two examples will show this relationship; both assume the same population standard deviation s of 100.

When *n* = 10, $\sigma_{\bar{x}} = \sigma / \sqrt{n}$

= 100/3.162 = 31.63, standard error of the mean.

And when *n* = 100:

Standard error of the mean; $\sigma_{\bar{x}} = 100 / \sqrt{100} = 10$

What have we shown? As we increased our sample size from 10 to 100 (a tenfold increase), the standard error dropped from 31.63 to 10, which is only about one-third of its former value. Our examples show that, because s*x* varies inversely with the square root of *n*, there is diminishing return in sampling.

It is true that sampling more items will decrease the standard error, but this benefit may not be worth the cost. It seldom pays to take excessively large samples. Managers should always assess both the worth