SREENARAYANAGURU
OPEN UNIVERSITY

BUSINESS STATISTICS

Course Code : B21BB04DC

Discipline Core Course

Bachelor of Business Administration

SELF LEARNING MATERIAL

SREENARAYANAGURU OPEN UNIVERSITY

The State University for Education, Training and Research in Blended Format, Kerala

SREENARAYANAGURU OPEN UNIVERSITY

Vision

To increase access of potential learners of all categories to higher education, research and training, and ensure equity through delivery of high quality processes and outcomes fostering inclusive educational empowerment for social advancement.

Mission

To be benchmarked as a model for conservation and dissemination of knowledge and skill on blended and virtual mode in education, training and research for normal, continuing, and adult learners.

Pathway

Access and Quality define Equity.

Business Statistics

Course Code: B21BB04DC

Semester - II

Discipline Core Course
Bachelor of Business Administration
Self Learning Material

(With Model Question Paper Sets)

SREENARAYANAGURU
OPEN UNIVERSITY

SREENARAYANAGURU OPEN UNIVERSITY

The State University for Education, Training and Research in Blended Format, Kerala

Business Statistics
Course Code: B21BB04DC
Semester - II
Bachelor of Business Administration

SREENARAYANAGURU
OPEN UNIVERSITY

Printed and published on behalf of Sreenarayanaguru Open University by Registrar, SGOU, Kollam.

www.sgou.ac.in

ISBN 978-81-970547-8-5

9 788197 054785

DOCUMENTATION

Academic Committee

Dr. D. Chandrabose
Ajeesh V.
Dr. B. Chandrachoodan Nair
Dr. Regina Sibi Cleetus
Vishnu Gopan
Dr. Johney Johnson
Dr. V. S. Santhosh
Dr. S. Madhavan
Dr.Vineeth K. M.
Dr. S. Priya
Dr Raju G.

Development of the Content

Dr. Sanitha K.K, Dr. Ajina K.P., Dr. Midhun V., Vikraman Pillai R.

Review

Content : Dr. Anitha C. S.
Format : Dr. I. G. Shibi
Linguistics : Dr. Anitha C. S.

Edit

Dr. Anitha C. S.

Scrutiny

Dr. Bino Joy, Dr. Sanitha K. K., Amar Shariar

Co-ordination

Dr. I. G. Shibi and Team SLM

Design Control

Azeem Babu T. A.

Cover Design

Jobin J.

Production

June 2024

YouTube

MESSAGE FROM VICE CHANCELLOR

Dear learner,

I extend my heartfelt greetings and profound enthusiasm as I warmly welcome you to Sreenarayanaguru Open University. Established in September 2020 as a state-led endeavour to promote higher education through open and distance learning modes, our institution was shaped by the guiding principle that access and quality are the cornerstones of equity. We have firmly resolved to uphold the highest standards of education, setting the benchmark and charting the course.

The courses offered by the Sreenarayanaguru Open University aim to strike a quality balance, ensuring students are equipped for both personal growth and professional excellence. The University embraces the widely acclaimed "blended format," a practical framework that harmoniously integrates Self-Learning Materials, Classroom Counseling, and Virtual modes, fostering a dynamic and enriching experience for both learners and instructors.

The university aims to offer you an engaging and thought-provoking educational journey. The Bachelor of Business Administration programme is highly coveted due to the current demand for skilled professionals in the field. This factor was central to our approach while designing the curriculum for this course. It strikes a balanced combination, providing a profound understanding of theoretical concepts alongside a clear exposition of practical applications. We have been cautious in ensuring that the management modules are balanced, preserving the integrity and distinctiveness of the discipline. The Self-Learning Material has been meticulously crafted, incorporating relevant examples to facilitate better comprehension.

Rest assured, the university's student support services will be at your disposal throughout your academic journey, readily available to address any concerns or grievances you may encounter. We encourage you to reach out to us freely regarding any matter about your academic programme. It is our sincere wish that you achieve the utmost success.

Warm regards.
Dr. Jagathy Raj V. P.

01-05-2024

CONTENTS

BLOCK - 01
Introduction to Business Statistics

Unit - 1

Introduction to Business Statistics

Learning Outcomes

Upon the completion of this unit, the learner will be able to:

- ◊ define business statistics with its importance for data-driven decision making in organisations.
- ◊ explain the functions of business statistics
- ◊ familiarise with the leverage of statistical analysis to make informed decisions related to production, accounting, marketing, human resources, operations, and finance functions in a business.

Prerequisite

Our lives are full of numbers and data. Alex plays basketball for his school team. His coach keeps track of statistics like how many points, rebounds, and assists each player makes in each game. By looking at all this data, the coach can see who the best players are and decide which position they should play next season. Alex's mom works for a pharmaceutical company. She uses statistics too. She collects data on how well different medicines work, any side effects they cause, and more. This helps her advice doctors on which drugs are best for their patients. The weather forecast uses statistics to predict weather patterns. They measure temperature, rainfall, wind speed, and air pressure. Then they analyse all this data to figure out what the weather is likely to be in the coming days and weeks. Companies also use statistics to understand customers. They track how many people view videos or click ads online, how many likes their social media posts get, and how much people buy their products. All this data helps companies figure out what people like so that they can market better.

As you can see in the aforementioned instances, numbers and data help us make sense of information, it is called statistics. Statistics help us make sense of numbers and data all around us. By collecting and analysing statistics, we can make better decisions about sports, medicine, weather, and business. Statistics provide insights from data.

Keywords

Statistics, Business Statistics, Applications of Business Statistics, Importance of Business Statistics, Functions of Business Statistics

Discussion

1.1.1 Statistics

In modern times, numbers play a vital role in advancing human knowledge across all disciplines, including science and society as a whole. Typically, people gain understanding in nearly every area of life through quantitative methods and rely on numerical analysis to make many important decisions. It is primarily based on numbers that knowledge and information become definitive and meaningful. The word "statistics" is derived from the German word "Statistik" and the Latin word "Status," meaning state or government. "Statistics" was first used by Gottfried Achenwall, a renowned German mathematician known as the "Father of Statistics." This suggests that statistics originated as the "Science of Kings." In early eras, rulers periodically conducted surveys within their kingdoms, mostly related to populations, money, and military. Before 3050 BC, Egyptian kings gathered data to build the world-famous pyramids. Over time, statistics have been utilised in diverse domains and have become an integral part of human civilization.

In the past, statistics was viewed as the 'science of statecraft' and was limited to a narrow field. But today, it applies to almost every aspect of nature and human activity. So the old, restrictive definitions have been replaced by new, more comprehensive ones. The word 'statistics' conveys different meanings in its singular and plural forms. As plural form, 'statistics' refers to numerical data sets. But in its singular form, it refers to the scientific discipline involving methods for gathering, analysing, and making inferences from numerical data.

In general sense, 'statistics' refers to numerical information expressed in quantitative terms. This information can be about anything such as objects, subjects, activities, or phenomena. At a broad level, statistics includes data on economic factors like gross domestic product and industry shares. At a detailed level, individual companies produce extensive statistics on their operations, like sales, production, spending, inventory, capital, and more. This company data is often collected systematically using scientific survey methods. Unless regularly updated, such data is a one-time effort with limited ongoing usefulness. To statisticians, statistics is a field of study like economics or math. It is a discipline that scientifically handles data by collecting, summarising, analysing, and presenting it. So, the field of statistics consists of appropriate methods for working with data.

Statistics is about working with data in the real world. All those numbers and graphs are just tools to make sense of information. Statistics was born as a set of tools to describe data, make predictions, and support decision-making. It's first applications were in census data, insurance, and astronomy. But soon its use spread to business, health, politics, and more. Statistics is really the science of learning from data. It helps make the complex world a little easier to navigate.

Statistics may be defined as "the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other."- Prof. Horace Secrist.

1.1.2 Business Statistics

Meet Priya, a small business entrepreneur who just opened a boutique clothing store. At first, Priya took decisions on intuition and some rough estimates. She simply ordered whatever cute styles she liked from wholesalers and hoped customers would buy them. But the first few months were shaky. Some items didn't sell well, while others sold out immediately. She was constantly worried about having too little or too much inventory. Her purchasing, marketing and staffing budgets were mostly guesswork. This started to affect her business negatively. So, she decided to solve this issue and found statistical tools that could help organise and analyse her sales data to spot trends and make better decisions. Here comes the importance of statistics in the smooth running of business organisations.

Business operations today are undergoing a fundamental shift compared to earlier times. In most organisations, traditional intuitive thinking has been replaced by more scientific, data-driven decision making. This means assessing various data sources and making business decisions based on meaningful conclusions drawn from analysing the data. However, deriving meaningful insights from raw data is not easy, it requires a thorough knowledge of statistics.

For a business decision maker, probably a manager, being able to extract useful information from data is a crucial skill. The Manager needs to have a basic understanding of statistics in order to leverage data for effective decision making. Knowledge of statistics helps managers make sense of data and enables them to take sound, evidence-based actions. Companies like Hindustan Unilever Limited rely heavily on statistical analysis to drive various aspects of their business.

Statistics in Action: Hindustan Unilever Ltd

Hindustan Unilever Ltd (HUL) is an Indian subsidiary of Fortune 500 multinational company, 'Unilever'. HUL is the undisputed leader in home care, personal care, food and beverage products in India. The company's 35 major

brands help people with their nutrition, hygiene, and personal care needs. Unilever has been in India since the 1800s, starting by selling Sunlight soap bars. Over the decades it launched many well-known brands like Lifebuoy, Lux, Vim, and Dalda. India's economic reforms in 1991 allowed HUL to grow further through partnerships and acquisitions.

Hindustan Unilever Limited

Today, HUL's brands like Lifebuoy, Lux, Surf Excel, Rin, Wheel, Fair & Lovely, Pond's, Sunsilk, Clinic Plus, Close-Up, Lakme, Brooke Bond, Kissan, Knorr, and Wall's are household names across India in many different product categories.

Statistics are very important for decision making at HUL. Researchers spend a lot of time gathering and analysing data to help the company decide which segments to enter, how to design ad campaign, whether to re-launch or re-brand products, etc. Statistics and analysis are critical for these types of decisions.

1.1.2.1 Meaning and Definition of Business Statistics

Business statistics refers to the application of statistical analysis to business-related data. It encompasses a wide variety of quantitative tools and techniques that allow organisations to make data-driven decisions, identify trends and patterns, optimise processes, and gain valuable insights. At its core, business statistics involves the collection, description, analysis and presentation of business data. It relies on statistical techniques such as descriptive statistics, probability distributions, hypothesis testing, regression analysis, forecasting models, time series analysis, correlation analysis, sampling methods, and more. These tools help businesses organise, summarise, analyse, interpret and visualise data in order to discover actionable information that can drive competitive advantage. For example, business statistics enables identification of customer buying patterns, guiding pricing decisions, predicting future sales, gauging manufacturing and operational efficiencies, optimising budgets, developing growth strategies, assessing risks, and much more.

Business statistics is an essential application of statistical principles, methodologies and tools to real-world business data and problems. It delivers the quantitative analytical capabilities to drive fact-based decision making and strategic thinking in organi-

sations. Business statistics turns data into an invaluable asset that provides a competitive edge. The insightful information extracted through statistical analysis provides an evidence-based foundation for making critical business decisions at both strategic and operational levels. It minimises uncertainty and risk while maximising productivity, quality, and profitability. In today's highly competitive data-driven business environment, the ability to collect, comprehend, visualise and leverage business statistics is a crucial skill for managers, analysts and leaders. It empowers them to take decisions based on facts and quantitative modeling rather than gut feelings or intuition alone.

Business statistics studies numerical data relating to the operations of business enterprises. This field focuses on identification, collection, analysis, interpretation and presentation of data to support business decision making and performance evaluation (Berenson M. et al. 2022).

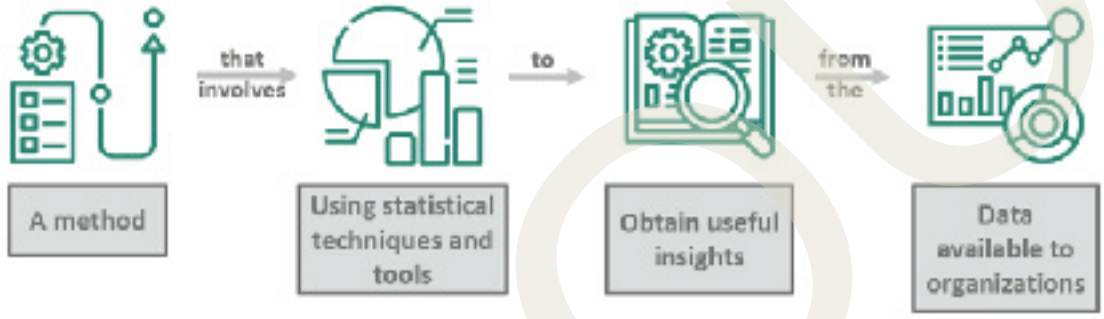


Fig 1.1.1 Business Statistics

1.1.2.2 Importance of Business Statistics

Following are the importance of statistics in business.

1. Make informed decisions - Statistics allow businesses to analyse data and identify trends, patterns and relationships. This information can be used to guide strategic decisions and planning. For example, analysing sales data might reveal insights for developing new products, expanding to new markets, adjusting pricing, etc.

2. Manage risks - Statistical analysis helps quantify uncertainty and potential risks. Businesses can use statistics to forecast demand, detect quality related problems, optimise inventory levels, and take measures to mitigate risks.

3. Set performance goals - Statistical tools like regression analysis allow businesses to identify driving factors behind performance metrics. Businesses can leverage insights to set realistic, data-driven goals and key performance indicators.

4. Optimise operations - Statistics help streamline operations by identifying inefficiencies, waste, and problems. For instance, statistical process control helps to monitor production quality over time. Businesses can tweak processes to enhance productivity.

5. Understand customers - Analysing customer data with statistics provides useful customer insights. Businesses can identify buying behaviours, preferences, demographics, and other attributes to tailor marketing, products and services.

6. Benchmark progress - Statistical analysis of historical performance metrics allows a business to evaluate progress over time. Businesses can determine their improvement, their stand against competitors, etc.

In summary, business statistics enable data-driven decision making. They provide valuable insights that help guide actions to achieve business objectives. Statistics are a fundamental tool for understanding patterns in data, identifying opportunities, and gaining a competitive edge.

1.1.2.3 Functions and Uses of Business Statistics

Business statistics turns data into usable information for improved business performance. Following are some of the major functions and uses. Let us learn it.

1. Descriptive statistics - Used to summarise and describe quantitative data related to a business. This includes measures like mean, median, standard deviation, range, etc., that help visualise and understand data.

2. Forecasting - Using historical data to predict future outcomes. This includes techniques like time series and regression analysis. Allows businesses to anticipate future demand, sales, economic conditions, etc.

3. Quality control - Statistical quality control techniques allow businesses to monitor production and service processes. Control charts and sampling help detect problems and minimize defects.

4. Decision making - Statistics help businesses make informed data-driven decisions. Metrics guide choices in operations, marketing, finance, HR, etc.

5. Market research - Businesses use statistics to gather and analyse market data, surveys, questionnaires, data mining, etc. provide insights into customers, competitors, pricing, etc.

6. Optimisation - Businesses optimise processes and activities using statistical techniques like linear programming models, queuing models, simulation, etc. that helps maximise output, efficiency, productivity.

7. Risk analysis - Statistical tools like Monte Carlo simulation allow businesses to model different scenarios and quantify potential risks that guides risk management.

8. Correlation analysis - Techniques like regression analysis help businesses identify and quantify relationships between variables that allows prediction of impacts.

1.1.2.4 Applications of Business Statistics

Business statistics provide actionable insights across all core business functions - from production to marketing to finance. They enable data-driven decision making to gain competitive advantage. Here is some of the important applications of business statistics:

1. Production: Businesses use statistics to determine optimal production quantities and scheduling. Techniques like forecasting, inventory modeling, quality control charts, and process optimisation help maximise efficiency and minimize costs.

2. Accounting: Statistical analysis is common in accounting, especially for auditing. Sampling methods allow auditors to test subsets rather than entire populations. Quantitative anomaly detection helps uncover potential fraud.

3. Research and Development: R&D departments rely heavily on statistics to improve product quality and develop new products. Experimental design, multivariate testing, and regression analysis help optimise research efforts and analyse results.

4. Economics: Statistics are crucial for understanding economic trends, relationships between indicators, and policy impacts. Models like regression quantify relationships between variables like inflation, unemployment, interest rates etc.

5. Human Resources: HR utilises statistics for performance evaluation, compensation planning, training need assessment, retention modeling, and more. Metrics help to guide effective talent management.

6. Marketing: Market analysis and segmentation depend on statistical techniques. Forecasting, conjoint analysis, and survey analysis inform product, pricing, and promotional decisions.

7. Operations: Statistical tools optimise inventory levels, supply chain management, workflows, equipment maintenance, and other operations. Quality control, Process Capability Analysis, Six Sigma, and Lean methods improve efficiency.

8. Finance: Statistics are used for portfolio optimisation, risk modeling, fraud detection, credit scoring, and financial forecasting. Metrics like Earnings Per Share (EPS), Return on Investment (ROI), Liquidity Ratios inform business valuation and decisions.

Recap

- ◊ Business Statistics has a vital role in advancing knowledge and decision-making.
- ◊ Originated as “science of kings” for surveys.
- ◊ Applying analysis to business data and enables data-driven decisions.
- ◊ Key functions: descriptive statistics, forecasting, quality control, decision-making, market research, optimisation, risk analysis, correlation analysis.
- ◊ Applications across production, accounting, R&D, economics, HR, marketing, operations, finance.
- ◊ Provides insights to maximise efficiency, productivity, quality, profits.

Objective Questions

1. What is the origin of the word “statistics”?
2. Who is known as the “Father of Statistics”?
3. Name two key techniques used in business statistics.
4. What does descriptive statistics help for?
5. How can statistics help manage risks?
6. How can statistics help for marketing?
7. What competitive edge does business statistics provide?
8. Which term refers to the application of statistical methods to business data?
9. Which technique can be used to predict future outcomes based on historical data?

Answers

1. German and Latin
2. Gottfried Achenwall
3. Forecast and Regression
4. Summarising data
5. Through quantifying uncertainty
6. Market analysis
7. Data-driven decisions
8. Business statistics
9. Forecasting

Assignments

1. Explain the meaning and definition of business statistics. Discuss in detail its scope, importance and applications in the context of business organisations.
2. What are the key functions and uses of business statistics? Explain with relevant examples.
3. Discuss the role and relevance of business statistics for a company's marketing efforts.
4. Choose an automobile company, identify key production and operational areas where business statistics can provide useful insights to improve efficiency, quality and profitability?
5. How can business statistics help in financial planning, analysis and decision making for a corporate entity? Explain the significance in areas of financial forecasting, budgeting, valuation, risk modeling and investment analysis.

Suggested Reading

1. Hair, J. F., Babin, B. J., & Krey, N. (2019). Marketing research: Text and cases. McGraw-Hill Education.
2. Tilakaratna, G. M. K. (2019). Business statistics for competitive advantage with Excel 2016. Springer.
3. Vazsonyi, A. (2019). Business statistics using Excel. Cambridge University Press.

Reference

1. Berenson, M., Levine, D. M., & Krehbiel, T. C. (2022). Basic business statistics: Concepts and applications. Pearson.
2. Anderson, D.R., Sweeney, D.J. & Williams, T.A. (2014). Statistics for business and economics (12th ed.). Cengage Learning.

Unit - 2

Statistical Investigation

Learning Outcomes

Upon the completion of this unit, the learner will be able to:

- ◊ explore the purpose and importance of statistical investigation in various fields.
- ◊ explain the steps involved in the statistical investigation process.
- ◊ describe the census method and the sampling method of statistical investigation.
- ◊ recognise the various types of probability sampling

Prerequisite

You are the owner of an ice cream parlour, and you want to find out which flavour of ice cream is the most popular among your customers. This is where statistical investigation comes into play. Instead of asking every single customer who visits your parlour about their favourite flavour (which would be a time-consuming and impractical approach), you decide to conduct a statistical investigation by taking samples of your customers.

Here is how you might go about it:

- ◊ Define the population: In this case, the population is all the customers who visit your ice cream parlour.
- ◊ Determine the sampling: You decide to survey a representative sample of, say, 100 customers over the course of a week.
- ◊ Collect data: You ask each customer in the sample to indicate their favourite ice cream flavour from the available options (e.g., chocolate, vanilla, strawberry, etc.).

- ◊ Analyse the data: After collecting responses from the 100 customers, you can analyse the data to determine which flavour received the most votes or mentions.
- ◊ Draw conclusions: Based on the analysis of your sample data, you can conclude that the most popular flavour among the customers you surveyed is likely to be the most popular flavour among your overall customer base (the population).

By conducting this statistical investigation, you have obtained valuable information about your customers' preferences without having to ask every single customer who visits your parlour. The key is to ensure that the sample you select is representative of the overall population, so that the conclusions drawn from the sample data can be generalised to the entire population with a reasonable degree of accuracy.

This simple illustration demonstrates the essence of statistical investigation. Using a carefully selected sample to make inferences and draw conclusions about a larger population, saving time and resources compared to studying the entire population.

Keywords

Methods and steps of Statistical investigation, Census Method, Sampling Method, Probability sampling, Non-probability sampling

Discussion

1.2.1 Statistical Investigation

A statistical investigation is a comprehensive process that helps us make data-driven decisions and learn about the world around us. It provides a structured methodology to collect, organise, analyse, interpret and communicate data effectively. It is a process of transforming raw data into useful information that can tell us more about a subject and allow us to make recommendations and possibly make predictions of future outcomes. More elaborately, statistical investigation is an information gathering and learning process that is undertaken to seek meaning and to learn more about any aspect of the real world. It helps make informed decisions and take informed actions. Statistical investigations should use the statistical enquiry cycle such as Problem, Plan, Data, Analysis and Conclusion.

Statistical investigation, also known as statistical analysis or data analysis, refers to the process of using statistical methods to analyse and interpret data in order to answer questions, test hypotheses, and make informed decisions. It involves collecting, organising, analysing, interpreting, and drawing conclusions from data. It involves using statistical methods and techniques to explore patterns, relationships, and trends in data in order to make informed decisions or draw meaningful conclusion. This process is used in various fields such as science, business, economics, and social sciences to gain insights and make evidence-based decisions.

Statistical investigation is widely used in various fields, including scientific research, business analytics, social sciences, healthcare, economics, and many others. It helps researchers and analysts make informed decisions, solve problems, identify patterns, and gain insights from data to support evidence-based decision-making. Statistical analysis is key to business functions, as it transforms raw data into actionable insights that drive strategy and planning. From market research to financial modeling to quality control and beyond, statistical investigations enable data-based decision making across the commercial landscape.

Consider the launch of a new consumer product. The marketing team will invariably conduct field surveys to key customer segments, aiming to quantify demand through statistical analysis before committing major capital. Careful sampling and questionnaire design lead to quality data on likely interest and intent to purchase at various price points. Regressions quantify price elasticity of the product while descriptive statistics reveal consumer preferences on the features of the product. This market research investigation informs critical product and promotional decisions. On the production side, statistical process control methods actively monitor manufacturing performance. Control charts track key output metrics over time to visibly detect significant process shifts. Capability ratio analysis then assesses whether the process consistently meets key specifications under normal operation. Together they enable data-driven quality control, minimizing costs associated with defects and waste. Forecasting future demand also leverages investigative analytics. Time series methods discern patterns in historical sales figures, weighing more recent data through exponential smoothing techniques. Regressions extend the methodology by layering related causal factors like pricing, promotions, GDP growth etc. The quantitative models generated provide objective sales forecasts.

Across functions ranging from finance to operations and beyond, statistical investigations unlock the value of business data. They bring the empirical rigour to transform raw information into informed strategy and plans. Understanding the versatility of statistical methods equips future business professionals with a profoundly useful decision-making skillset.

1.2.1.1 Purpose of Statistical Investigation

1. Describe and summarise data: Statistical investigations allow us to synthesize large, complex data sets. Descriptive statistics like mean, percentile, variance, etc. summarise important characteristics of the data distribution.

2. Compare groups or measurements: Statistical tests help compare differences between groups, treatments, measurement methods, etc. Hypothesis tests, for example t-tests, ANOVA, determine if apparent differences are statistically significant.

3. Establish relationships and correlations: Techniques like regression analyses the relationship between variables, for example how one variable may predict or explain the other. Correlation analysis measures the strength of linear relationships.

4. Model data and predict outcomes: Statistical models such as linear models and machine learning models are used to model complex processes to explain and predict outcomes based on a set of predictor variables.

5. Inform decisions: Analysing data provides an objective way to gain insights to guide evidence-based decision making in business, policy making, medical diagnose and more fields.

6. Design experiments and surveys: Statistics helps design good experiments and surveys so that meaningful data can be collected efficiently.

7. Test theories and hypotheses: Statistics tests whether a hypothesis or scientific theory is true based on analysing real data. This allows theories to improve over time.

8. Reduce uncertainty: Statistical tools quantify uncertainty and show how confident the results are and what the risks are. They give the probability of something happening instead of a definitive yes or no prediction.

1.2.1.2 Steps in Statistical Investigation

The purpose of a statistical investigation is to gain a better understanding of a particular phenomenon or problem by examining relevant data. This process typically involves the following steps:

1. Defining the research question or objective: Clearly stating what you want to investigate or achieve through the analysis.

2. Data collection: Gathering relevant data that are suitable for answering the research question. This can involve various methods such as surveys, experiments, observations, or existing datasets.

3. Data organisation and cleaning: Ensuring the collected data is properly organised and preparing it for analysis. This step involves removing any errors, inconsistencies, or outliers that may affect the results.

4. Data analysis: Applying appropriate statistical techniques and methods to explore and analyse the data. This can involve descriptive statistics (e.g., measures of central tendency, variability), inferential statistics (e.g., hypothesis testing, confidence intervals), regression analysis, or other advanced statistical

modeling techniques.

5. Interpretation of results: Examining the outcomes of the analysis and drawing meaningful conclusions or insights. This involves relating the findings back to the research question or objective.

6. Presentation and communication: Communicating the results of the statistical investigation in a clear and concise manner. This can involve visualizations (e.g., graphs, charts), summary statistics, and written reports or presentations.

1.2.1.3 Methods of Statistical Investigation

There are two popular methods of statistical investigation viz, Census Method and Sample Method.

1.2.1.3.1 Census Method

You have a small classroom with only 10 students. The teacher wants to know the average height of the students in the class. One way to find this would be to measure the height of every single student in the classroom and then calculate the average. In this approach, where the teacher collects data from each and every unit (student) in the population (the classroom), is known as the Census Method. By measuring the height of all 10 students, the teacher is conducting a complete enumeration or a census of the population. This method ensures that no individual is left out, and the resulting average height calculated from the data will be an accurate representation of the entire classroom.

When the investigator collects information from all the units and elements, then it is called census method. Under this, information about each and every unit of the aggregate is collected. This method is also known as Complete Enumeration Method. It is a complete enumeration of the entire population, leaving no unit uncounted or unobserved. This method is often employed when the population size is relatively small or when the study demands a comprehensive understanding of the entire population. Population census is an example of census method. Under population census, information is obtained about every household and every person. This information is expensive. The results derived from this method are authentic and reliable. This method is costly because labour and time involved is high. To give you another example, if you record the marks of all students of B.Com of the Mumbai University for analysis, it is a census investigation.

Merits of a Census Method

Let's look at the merits of a census investigation.

1. Intensive Study - Under census investigation, you must obtain data from each and every unit of the population. Further, it enables the statistician to study more than one aspect of all items of the population. To give you an example, the Indian Government conducts a census investigation once every 10 years. The authorities collect the data regarding the population size, males, and females,

education levels, sources of income, religion, etc.

2. Reliable Data - The data that a statistician collects through a census investigation is more reliable, representative, and accurate. This is because, in a census, the statistician observes every item personally.

3. Suitable Choice - It is a great choice in situations where the different items of the population are not homogeneous.

4. The basis of various surveys - Data from a census investigation is used as a basis in various surveys.

Demerits of a Census Method

A census method of investigation also has certain demerits. Some of these demerits are:

1. Costs - Since the statistician closely observes each and every item of the population before collecting the data, it makes census investigation a very costly method of investigation. Usually, government organisations adopt this method to collect detailed data like the population census or agricultural census or the census of industrial protection, etc.

2. Time-consuming - A census investigation is time-consuming and also requires manpower to collect original data.

3. Possibilities of Errors - There are many possibilities of errors in the census investigation method due to non-response, measurement, lack of preciseness of the definition of statistical units or even the personal bias of the investigators.

1.2.1.3.2 Sampling Method

Priyanka was tasked with estimating the proportion of defective products in a large manufacturing batch. Instead of inspecting every single item, which would be extremely time-consuming and labour-intensive, she decided to employ the Sample Method. Priyanka began by defining the population as all the products in the manufacturing batch. Next, she had to determine an appropriate sample size and select a representative sample from the batch. Following established sampling techniques, Priyanka randomly selected 100 products from the batch. She carefully inspected each item in the sample, meticulously recording any defects found. Out of the 100 products in her sample, Priyanka identified 8 defective items. Based on this sample data, she calculated that the proportion of defective products in her sample was 8/100 or 8%. Assuming that her sample was truly representative of the entire manufacturing batch, Priyanka could reasonably infer that approximately 8% of the products in the overall population were likely to be defective. By using the Sampling Method, she could make an accurate estimate about the larger batch without having to inspect every single product, and can save valuable time and resources. Priyanka understood the importance of adhering to the principles of probability, statistical regularity, and the inertia of large numbers when employing the Sampling Method. She ensured that her sample was randomly selected,

representative of the population, and of an adequate size to minimise the influence of random fluctuations. By following these guidelines, Priyanka could confidently present her findings to her superiors, knowing that her conclusions were statistically sound and reliable.

Under this, some representative units are selected from the aggregate and conclusions are derived through study of those selected units. In daily life, while purchasing things for the household such as wheat, rice, etc. we do not inspect each and every piece of wheat or rice in the packet, rather we take sample of some pieces and decide to purchase wheat or rice on this basis. Time and money are saved through this method.

Sampling is the process of choosing a group of observations or samples to study. In majority types of research, whether quantitative or qualitative, sampling is essential. Why? Because no research can include everyone and everything. Imagine trying to study every single person, everywhere, doing everything – it's just not practical in all situations. So, sampling often involves selecting a group of people to study. The main ideas here are the "population," which is the entire group you ideally want to study, and the "sample," which is the specific group of people you actually study and collect data from. So, in simple terms, sampling is how researchers pick a smaller group to represent the bigger group they want to learn about. It is important to maintain extreme caution in this method, otherwise the possibility of deriving incorrect conclusions may arise.

The Sampling Method is a statistical technique used to study and make inferences about a larger population by selecting and analysing a representative subset or sample from that population. Instead of examining every single unit or element in the population, which can be costly, time-consuming, and sometimes impractical, the Sampling Method involves carefully selecting a sample and collecting data from that sample.

Principles of Sampling

1. Principle of Probability

A polling organisation wants to conduct a survey to estimate the voting preferences of the population in an upcoming election. They use random digit dialing to randomly select phone numbers from the entire population of registered voters, ensuring that every voter has an equal probability of being included in the sample. This is called the principles of probability. This principle states that every unit in the population must have a known, non-zero probability of being selected in the sample. This ensures that the sample is not biased towards any particular segment of the population.

For example, you want to estimate the average income of households in a city. If you only survey households in wealthy neighbourhoods, your sample will be biased and will not represent the true average income of all households in the city. Instead, you should use a probability sampling method, where every household has an equal chance of being selected.

2. Principle of Statistical Regularity

A market research firm wants to study the preferences of consumers for a particular product. They select a sample that includes respondents from different age groups, income levels, and geographic regions in the same proportions as they exist in the overall population of consumers. This ensures that the sample is representative and can provide reliable insights into the preferences of the entire consumer population. It is known as the principle of Statistical Regularity. This principle states that the sample should be representative of the population, meaning that it should exhibit the same characteristics and patterns as the population as a whole.

For instance, if you want to estimate the average height of students in a school, your sample should include students from different grades, genders, and ethnic backgrounds in the same proportions as they exist in the overall student population. If your sample consists mostly of students from a particular grade or ethnic group, it will not be representative of the entire school.

3. Principle of Inertia of Large Numbers

In a quality control process, a manufacturer selects a sample of 100 products to inspect for defects. The sample size of 100 is considered large enough to provide a reliable estimate of the defect rate in the entire production batch, as the influence of random fluctuations is minimized for a sample of this size. This is known as the principle of Inertia of large sample. This principle states that as the sample size increases, the sample becomes more representative of the population, and the influence of random fluctuations diminishes.

For example, you want to estimate the proportion of defective products in a batch of 10,000 items. If you inspect a sample of 10 items, random fluctuations can significantly affect your estimate. However, if you inspect a sample of 500 items, the influence of random fluctuations will be much smaller, and your estimate will be more accurate and closer to the true proportion in the population.

By adhering to these three principles such as Probability, Statistical Regularity, and Inertia of Large Numbers, researchers can ensure that their samples are representative, unbiased, and that the conclusions drawn from the sample data can be generalised to the larger population with a reasonable degree of accuracy and reliability.

Methods/Techniques of Sampling

The sampling techniques are divided into two broad categories;

A. Probability / Random Sampling

B. Non-probability / Non-random sampling

A. Probability / Random Sampling

Priyanka wants to conduct a survey to understand the movie preferences of people in her town. Instead of randomly knocking on doors and hoping to get a representative sample, Priyanka decides to use probability sampling. This means she gives every

person in the town an equal chance of being selected for the survey, like putting all their names in a hat and randomly drawing out the participants. This way, Priyanka's sample is truly representative of the entire town's population, and she can be confident that the survey results reflect the overall movie preferences accurately. Similarly, the researchers have different ways of picking who or what they study. What they both aim for is to choose a smaller group to look at, which represents a bigger group or population. This smaller group helps them make conclusions about the larger group. They tend to use sampling based on theories of probability from mathematics called probability sampling. Researchers use probability or random sampling for two main reasons. First, it helps save a lot of time and money. Imagine if you had to collect data from 20 million people – it would take forever and cost a ton. But with sampling, you can pick a smaller group, like 2000 people, and get results that are almost as good as if you had surveyed everyone. Second, probability sampling makes your results more accurate. It's like trying to count every single person in a big country, which can be really hard. But with the right sampling method, you can still get very accurate results without counting every single person. For example, the U.S. Census Bureau in 2000 could have used this method to get more precise results.

Qualitative researchers don't worry as much about having a sample that perfectly represents a big group of people. Instead, they are more interested in how the small group who can help them understand important things about social life. They pick cases, events, or actions that can give them a deeper insight into what is happening in a specific situation.

Types of probability Sampling

1. ***Simple Random Sampling***: Imagine you want to select a group of students to represent your entire school in a debate competition. Instead of handpicking the students, you decide to use simple random sampling. You put all the students' names in a hat and randomly draw out the required number of names. This ensures that every student in the school has an equal chance of being selected, and the selected group is truly representative of the entire school population. Simple random sampling gives you an unbiased sample, allowing you to make reliable inferences about the school as a whole. In simple random sampling, a researcher follows these steps: First, they make a list of all the things or people they want to study (this is called a sampling frame). Then, they use a random method, like picking numbers out of a table with random numbers, to choose which things or people to include in their sample. They go and find the exact things or people they picked from the list. The idea is to make sure that everything on the list has an equal chance of being chosen, so it's fair. You can find these random number tables in books about statistics and research methods.

2. ***Systematic Sampling****:* Priyanka needs to select a sample of 50 customers from a list of 1,000 customers for a market research survey. Instead of randomly picking names, she decides to use systematic sampling. She starts by dividing the total number of customers (1,000) by the desired sample size (50) to get a sampling interval of 20. Then, Priyanka randomly selects the first customer from the list, and after that, she chooses every 20th customer until she has 50 participants. This

systematic approach ensures that the sample is spread evenly across the entire customer list, making it representative of the whole population. Priyanka can now confidently use the data from her systematically selected sample to gain insights about the preferences of all 1,000 customers. Systematic sampling is a method of selecting a sample from a larger population in a systematic and organised way. It involves choosing every nth item or individual from a list or sequence after a random starting point has been determined. For example, if you want to select a systematic sample of 100 students from a school with 1,000 students, you could start by randomly selecting a number between 1 and 10. Let's say you choose 3. Then, you would select every 3rd student from the list of students, starting with the randomly chosen student. So, if the randomly selected student is the 7th student on the list, you would select the 7th, 10th, 13th, and so on until you have your sample of 100 students. Systematic sampling is a more efficient way to create a representative sample compared to simple random sampling, especially when dealing with large population.

3. ***Stratified Sampling***: Priyanka wants to understand the reading habits of students in her school, which has both primary and secondary grades. Instead of randomly selecting students, Priyanka decides to use stratified sampling. First, she divides the students into two distinct groups, or "strata" - primary and secondary. Then, Priyanka randomly selects a proportionate number of students from each stratum to include in her sample. This way, her sample accurately represents the different grade levels in the school, ensuring that Priyanka gets a comprehensive understanding of the reading habits across all students, not just one group. Stratified sampling helps Priyanka capture the diversity within the student population and make more accurate inferences about the reading preferences of students in her school.

   Stratified sampling is a method of selecting a sample from a larger population by dividing the population into subgroups or strata based on certain characteristics that are important to the research. Each subgroup or stratum represents a homogeneous group with respect to those characteristics. Then, researchers randomly select samples from each stratum in proportion to its size or importance in the population. The goal of stratified sampling is to ensure that the sample is representative of the entire population, especially when there are significant variations or differences within the population. By dividing the population into strata and then randomly selecting samples from each stratum, researchers can capture the diversity of the population's characteristics in their sample. For example, if you were conducting a survey about people's income levels in a city, you might first divide the population into strata based on income brackets (e.g., low-income, middle-income, high-income). Then, you would randomly select samples from each stratum to ensure that your overall sample reflects the distribution of income levels in the city accurately. Stratified sampling is particularly useful when there is a need to analyse subgroups within the population separately, or when certain characteristics are known to significantly affect the research outcomes. It helps improve the accuracy and representativeness of the sample for the purposes of research and analysis.

4. ***Cluster Sampling***: Priyanka wants to survey the shopping habits of households in a large city. Instead of randomly selecting individual households across the entire city, which could be time-consuming and costly, she decides to use cluster sampling. Priyanka first divides the city into several smaller, geographically distinct areas or "clusters," such as neighborhoods. Then, she randomly selects a few of these clusters and surveys all the households within those chosen clusters. This method allows Priyanka to efficiently collect data from a representative sample of the city's population, without having to travel to every corner of the city. By focusing on the selected clusters, Priyanka can gather reliable insights about the shopping habits of households across the larger urban area.

Cluster sampling is a sampling method used in research where the larger population is divided into clusters or groups, and then a random sample of these clusters is selected for further study. Instead of selecting individual members from the population directly, researchers randomly choose entire clusters and then study all the members within those selected clusters. Cluster sampling is particularly useful when it's difficult or impractical to collect a simple random sample of individuals from the entire population. It can save time and resources, especially when the population is large and widely dispersed. Here is an example: Suppose you want to conduct a health survey of people in a large country. Instead of trying to randomly select individuals from every part of the country, which could be very costly and logistically challenging, you could randomly choose a few cities or regions (clusters) from different parts of the country. Then, within each selected city or region, you would study all the individuals you can find or select a sample from within that specific cluster. Cluster sampling is especially useful when the clusters themselves are similar in some way, but they differ from each other. This method helps capture the variation between clusters and can provide a reasonable representation of the entire population without the need to study every single individual. However, it's important to ensure that the clusters are chosen randomly to avoid bias.

B. Non-Probability Sampling

Priyanka wants to understand the preferences of diners at a popular local restaurant. Instead of randomly selecting customers, she decides to use a non-probability sampling method. Priyanka chooses to approach customers as they enter the restaurant and ask them to participate in her survey. This convenience sampling technique allows Priyanka to quickly gather responses, but it also means that her sample may not be fully representative of the restaurant's entire customer base. Some customers may be more willing to participate than others, leading to potential biases in Priyanka's data. While this non-probability approach is faster and easier, Priyanka knows she must be cautious when drawing conclusions, as the sample may not accurately reflect the true preferences of all the restaurant's diners. Non-probability sampling is a method of selecting a sample from a larger population in a way that does not provide each member of the population with a known or equal chance of being included in the sample. Unlike probability sampling, which relies on random selection, non-probability sampling methods involve subjective judgment or convenience.

Types of Non-Probability Sampling

1. ***Convenience Sampling****:* Priyanka, a marketing researcher, wants to gather feedback on a new product her company is launching. Instead of spending time and resources to randomly select participants, she decides to use convenience sampling. Priyanka sets up a booth at the local mall and approaches shoppers passing by, asking them to take a quick survey about the product. This allows Priyanka to easily and quickly collect responses from people who are readily available and willing to participate. While this method is efficient, Priyanka knows that her sample may not be truly representative of the target market, as the people she surveys at the mall may have different preferences and characteristics than the general population. Priyanka will need to be cautious when generalizing the results from her convenience sample to the broader market. This method involves selecting individuals or items that are most convenient or readily available for the study. It's often used for its simplicity and low cost, but it may introduce bias because it doesn't ensure that the sample is representative of the population.

2. ***Quota Sampling****:* Priyanka, a market researcher, wants to understand the preferences of different age groups for a new product launch. Instead of randomly selecting participants, she decides to use quota sampling. Priyanka first identifies the key demographic segments she wants to represent, such as younger adults, middle-aged, and older adults. She then sets a target number of participants, or quota, for each age group. Priyanka then approaches people in public places and selects participants until she has filled her quota for each age group. This ensures that her final sample reflects the proportions of the different age segments in the overall population, allowing Priyanka to make more reliable comparisons and draw insights about how the product may appeal to diverse age groups. While not as random as probability sampling, quota sampling helps Priyanka efficiently gather representative data for her market research. Researchers divide the population into subgroups (strata) based on certain characteristics and then select participants from each subgroup until a predetermined quota is reached. Quota sampling aims to match the proportions of subgroups in the population but doesn't use random selection.

3. ***Purposive or Judgement Sampling***: Priyanka, a market researcher, wants to evaluate the effectiveness of a new advertising campaign targeting young urban professionals. Instead of randomly selecting participants, she decides to use purposive or judgmental sampling. Priyanka carefully selects individuals who she believes are representative of the target audience - young, college-educated, and working in professional fields. She approaches these individuals directly, knowing that their experiences and opinions will provide valuable insights about the campaign's impact. While this method is not random, Priyanka's expert judgment allows her to deliberately choose participants who can offer the most relevant and meaningful feedback. By using purposive sampling, Priyanka can gather in-depth qualitative data to understand how the advertising resonates with her target market, even though her sample may not be statistically representative

of the broader population. Researchers intentionally choose specific individuals or items based on their expertise or judgment. This method is useful when the researcher wants to study a particular subgroup or when they believe certain cases are most relevant to the research.

4. ***Snowball Sampling***: Priyanka, a social researcher, wants to study the challenges faced by homeless individuals in her city. Since this is a difficult-to-reach population, she decides to use snowball sampling. Priyanka starts by identifying a few homeless individuals and interviews them. She then asks these initial participants to refer her to other homeless people they know who might be willing to share their experiences. As Priyanka follows these referrals, her sample begins to grow, much like a snowball rolling downhill and picking up more snow. By leveraging the existing connections and networks of her initial participants, Priyanka is able to gradually build a larger, more diverse sample of homeless individuals to interview. While this non-probability method may not yield a statistically representative sample, it allows Priyanka to effectively reach and gather data from a hard-to-access population that would be difficult to locate through other sampling techniques. This technique is often used in studies where the target population is hard to reach. Researchers start with a small number of participants and ask them to refer others they know who meet the criteria. This method is common in social network or hidden population studies. It is also called network; chain referral or reputational sampling is a method for identifying and sampling the cases in a network. Snowball sampling is a multistage technique. It begins with one or few people or cases and spreads out on the basis of links to the initial cases. On use of snowball sampling is to sample a network.

5. ***Deviant Case Sampling***: Priyanka, a social science researcher, is studying the factors that contribute to academic success among high school students. Instead of focusing only on the typical or average students, she decides to use deviant case sampling. Priyanka identifies students who have achieved exceptionally high grades despite facing significant challenges or disadvantages, such as economic hardship or learning disabilities. By closely examining these outliers or "deviant cases," Priyanka hopes to uncover unique insights and patterns that may not be apparent in a more general sample of students. While this non-probability sampling method does not aim to be representative of the entire student population, it allows Priyanka to gain a deeper understanding of the exceptional factors and strategies that enable some students to thrive, even in the face of adversity. Priyanka believes that by studying these deviant cases, she can better inform interventions and support systems to help more students achieve academic success. A researcher uses deviant case sampling also called extreme case sampling, when he or she seek cases that differ from the dominant pattern or that differ from the dominant pattern or that differ from the predominant characteristics of other cases. Deviant case sampling differs from purposive sampling in that the goal is to locate a collection of unusual, different or peculiar cases that are not representative of the whole. The deviant cases are selected because they are unusual and a researcher hopes to learn more about the social life by considering cases that fall outside the general pattern or including what is beyond the main flow of events.

6. ***Sequential Sampling***: Priyanka, a market researcher, is studying consumer preferences for a new product line. Instead of collecting data all at once, she decides to use sequential sampling. Priyanka starts by gathering data from an initial set of participants. She then analyses the results and determines if she has enough information to answer her research questions. If not, Priyanka collects data from additional participants, repeating this process until she is satisfied that her sample size is sufficient. This sequential approach allows Priyanka to be efficient with her time and resources, as she can stop data collection once she has reached a point of saturation or diminishing returns. By continuously evaluating the data and making informed decisions about when to gather more information, Priyanka can ensure that her final sample provides the necessary insights to understand consumer preferences, without over-sampling or wasting effort. It is similar to purposive sampling with one difference. In purposive sampling, the researcher tries to find as many relevant cases as possible, where in sequential sampling, a researcher continues to gather cases until the amount of new information or diversity of cases is filled.

Advantages of Sampling Method

1. In the sampling method, the number of units utilised is significantly less. It helps to arrive at results much quicker.
2. Sample method incurs substantially lower cost than the census method as these tests are done over a limited sample.
3. It has greater scope than the census method as it acts as a substitute in such cases where the latter becomes impracticable. For instance, if a manufacturer wants to test its range of toasters and other kitchen appliances, it will apply a sample method and not a census method.
4. The nature of the sample method is such that it can be employed to check the results from the census method. Also, due to the small size of its sample, the method is useful for cross-checking the reliability of its own results. A small sample can be taken out of generated results, and that sample will have to be investigated.
5. Accuracy And Reliability: Census method confirms a higher degree of accuracy than other techniques. The Census method provides complete information because each and every item is investigated carefully. Therefore, it is a very reliable method of data collection.
6. Suitability: Census method is effective if the universe is small.
7. Intensive Study: Census method examines each unit completely and gathers important data for intensive study.
8. Indispensable: Census method is most reliable in certain cases where other methods cannot provide reliable and accurate results.
9. Heterogeneous Units: This method is also applicable to examine heterogeneous units.

Limitations of Census Method

1. The expenditure incurred during the census is much higher because of the sheer size of the population. Also, data is collected from each unit of a sample population, which requires additional costing.
2. Owing to the huge volume of data that is collated, a greater number of the workforce (as well as man-hours) is required for completion.
3. Costly Method: Census method is a very costly method of data collection.
4. Time Consuming: Census method consumes more time and labor to complete data collecting tasks.
5. Unsuitability: Census method is not applicable or suitable if the universe is large. This method is suitable only for a small universe.
6. Chance of Errors: There is a comparatively higher chance of statistical errors in this method.

Recap

- Statistical Investigation:
  - A comprehensive process to make data-driven decisions and learn about the world
  - Involves collecting, organising, analysing, interpreting, and drawing conclusions from data
  - Used in various fields to gain insights and make evidence-based decisions
- Purpose of Statistical Investigation:
  - Describe and summarise data
  - Compare groups or measurements
  - Establish relationships and correlations
  - Model data and predict outcomes
  - Inform decisions
  - Design experiments and surveys
  - Test theories and hypotheses

- Reduce uncertainty

◊ Steps in Statistical Investigation:

- Defining the research question or objective
- Data collection
- Data organisation and cleaning
- Data analysis
- Interpretation of results
- Presentation and communication

◊ Methods of Statistical Investigation:

- Census Method
  - * Collecting data from every unit in the population
  - * Merits: Intensive study, reliable data, suitable for heterogeneous populations
  - * Demerits: Costly, time-consuming, possibility of errors
- Sampling Method
  - * Collecting data from a representative sample of the population
  - * Principles: Probability, statistical regularity, inertia of large numbers
  - * Methods/Techniques of Sampling: Probability and non-probability sampling
  - * Advantages: Lower cost, faster, greater scope, accuracy, and reliability

Objective Questions

1. What is the primary purpose of statistical investigation?
2. What is the term used for collecting data from every unit in the population?
3. Which principle states that every unit in the population must have a known, non-zero probability of being selected in the sample?

4. Which principle states that the sample should exhibit the same characteristics and patterns as the population as a whole?
5. As the sample size increases, what principle states that the influence of random fluctuations diminishes?
6. What is the advantage of the sampling method in terms of cost?
7. What is the advantage of the census method in terms of data reliability?
8. What is the limitation of the census method in terms of suitability for large populations?
9. Which method is suitable for examining heterogeneous units?
10. What is the advantage of the sampling method in terms of scope?
11. What is the advantage of the census method in terms of the possibility of cross-checking the results?
12. What is the disadvantage of the census method in terms of time consumption?
13. What are the two broad categories of sampling techniques?
14. What is the purpose of using probability or random sampling in research?
15. Which sampling technique involves selecting every n^{th} item or individual from a list or sequence after a random starting point?

Answers

1. Decision-making
2. Census method
3. Probability
4. Statistical regularity
5. Inertia of large numbers
6. Lower cost
7. Reliable data
8. Unsuitable

9. Census method
10. Greater scope
11. Reliability
12. Time-consuming
13. Probability/Random sampling and Non-probability/Non-random sampling.
14. To save time and money, and to make the results more accurate.
15. Systematic sampling.

Assignments

1. What are the steps involved in a statistical investigation?
2. Explain the purpose of statistical investigation.
3. Differentiate between the census method and the sampling method.
4. Discuss the principles of sampling in the sampling method.
5. Analyse the advantages and limitations of the census method and the sampling method.
6. Explain the concept of stratified sampling and its advantages.
7. Describe the different types of non-probability sampling techniques and their applications.
8. Design a statistical investigation to study the impact of a new marketing campaign on customer satisfaction.
9. Develop a sampling plan to estimate the average income of households in a city.
10. Discuss the role of statistical investigation in evidence-based decision-making in a business context.
11. Critically evaluate the use of the census method and the sampling method in different scenarios.
12. Apply the principles of sampling to select a representative sampling for a survey on consumer preferences.

Suggested Reading

1. Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE Publications.
2. Babbie, E. (2016). The practice of social research (14th ed.). Cengage Learning.
3. Gravetter, F. J., & Wallnau, L. B. (2017). Statistics for the behavioral sciences (10th ed.). Cengage Learning.

Reference

1. Howell, D. C. (2016). Fundamental statistics for the behavioral sciences (9th ed.). Cengage Learning.
2. Agresti, A., & Finlay, B. (2014). Statistical methods for the social sciences (4th ed.). Pearson.
3. Moore, D. S., Notz, W. I., & Fligner, M. A. (2013). The basic practice of statistics (6th ed.). W. H. Freeman.
4. Triola, M. F. (2018). Elementary statistics (13th ed.). Pearson.
5. McClave, J. T., Benson, P. G., & Sincich, T. (2018). Statistics for business and economics (13th ed.). Pearson.

Unit - 3

Data and Collection of Data

Learning Outcomes

Upon the completion of this unit, the learner will be able to:

- ◊ familiarise the different types and formats of data.
- ◊ describe the primary and secondary data collection methods and their applications.
- ◊ identify and address common challenges in the data collection process.

Prerequisite

When you go to the grocery store, you collect details about the items you need to buy, the prices of those items, and the quantities you purchase. These details can be used to create a shopping list, budget your spending, and track your consumption patterns over time. For example, as you stroll through the aisles, you make a mental note of the items you need to restock in your pantry, such as eggs, rice, bread, chicken and so on. You also observe the prices displayed on the shelves, which allows you to estimate the total cost of your purchases. Finally, you carefully select the quantities of each item, ensuring you have enough to last until your next grocery trip. Here you have done the collection of data. The specific items, their prices, and the quantities you buy are the data that can then be utilised to plan your next shopping trip, manage your household budget, and even identify trends in your eating habits over the weeks and months ahead. By recognising the power of the data you gather during a simple trip to the grocery store, you can unlock valuable insights and make more informed decisions. In our daily lives, we encounter data all the time, even in the most mundane activities like grocery shopping. By understanding the concept of data and its applications, we can harness its power to enhance our decision-making processes, solve complex problems, and gain a deeper understanding of the world we live in.

Keywords

Data, Collection of data, Primary Data, Secondary Data, Qualitative data, Quantitative data, Categorical data, Nominal data, Ordinal data, Discrete data, Continous data

Discussion

1.3.1 Data

A group of friends who are passionate gardeners, wanted to understand how different soil types and watering patterns affect the growth of their favourite sunflowers. The gardeners divide their garden into several plots, each with a different soil type (e.g., sandy, loamy, clayey). They carefully plant sunflower seeds in each plot and label them accordingly. The friends create a watering schedule, where some plots receive regular watering, while others are watered less frequently. Over the growing season, the gardeners meticulously measure and record the height of each sunflower plant at regular intervals. They also note all observations, such as the overall health and vigor of the plants, the number of flowers produced, and any insect or disease issues. By the end of the growing season, the gardeners have accumulated a wealth of data, including soil type for each plot, watering schedule for each plot, Sunflower plant height measurements over time, observations on plant health and flower production. With this comprehensive dataset, the friends can now analyse the information to uncover patterns and insights. They might discover that certain soil types and watering regimes are more conducive to optimal sunflower growth and flowering, which could help them plan their garden layout and maintenance strategies for the upcoming season.

This illustration highlights the systematic nature of data collection in statistics, where researchers or observers carefully control and record variables to understand their relationships and draw meaningful conclusions. This organised and structured approach to data gathering is essential for any statistical analysis, whether it is in the realm of gardening, scientific research, or business analytics.

Since the invention of computers, people have used the term data to refer to computer information, and this information was either transmitted or stored. But that is not the only definition of data; there exist other types of data as well. So, what is the data? Data can be texts or numbers written on papers, or it can be bytes and bits inside the memory of electronic devices, or it could be facts that are stored inside a person's mind. Therefore, data is a collection of facts, figures or observations that are organised in a way that can be used, analysed, and interpreted. It is the foundation upon which we build our understanding of the world around us. In a more concrete sense, data can take many forms, such as Numerical values (e.g., temperature readings, sales figures, population statistics), Textual information (e.g., customer reviews, news articles, survey responses), Categorical or qualitative data (e.g., gender, product categories, opinion

ratings), Multimedia content (e.g., images, videos, audio recordings) and Time-series data (e.g., stock prices, weather patterns, website traffic).

Data can help businesses better understand their customers, improve their advertising campaigns, personalize their content and improve their bottom lines. The advantages of data are many, but you can't access these benefits without the proper data analytics tools and processes. While raw data has a lot of potential, you need data analytics to unlock the power to grow your business. The father of information theory Claude Shannon, an American Mathematician is responsible for the origins of the concept of Data in computing. Introduced this concept through his paper "A Mathematical Theory of Communication" in 1948.

1.3.1.1 Characteristics of Data

1. They are facts obtained by reading, observation, Counting, measuring and weighing etc. which are recordable.
2. Data are derived from external and internal sources of the Organisation
3. Data may be produced as an automatic bye-product of some routine but essential operation such as production of an invoice.
4. The source of data needs to be given considerable attention because if the data is wrong the resulting information will be worthless.

1.3.1.2 Formats of Data

The Data are stored and processed by computers. They are:

1. Text which consists of strings of characters.
2. Numbers.
3. Audio, namely speech, and music.
4. Pictures – monochrome and colour.
5. Video is sequence of pictures such as movies or animation. Usually, video data has an accompanying soundtrack which is synchronized with the pictures.

1.3.1.3 Types of Data

A. Qualitative or Categorical Data

Qualitative or Categorical Data is data that can't be measured or counted in the form of numbers. These types of data are sorted by category, not by number. That's why it is also known as Categorical Data. These data consist of audio, images, symbols, or text. The gender of a person, i.e., male, female, or others, is qualitative data. Qualitative data tells about the perception of people. This data helps market researchers understand the customers' tastes and then design their ideas and strategies accordingly. **Some of the**

examples of qualitative data are language you speak, Favourite holiday destination, Opinion on something (agree, disagree, or neutral) and Colours.

The Qualitative data are further classified into two parts:

a. **Nominal Data**

Nominal Data is used to label variables without any order or quantitative value. The colour of hair can be considered nominal data, as one colour can't be compared with another colour. The name "nominal" comes from the Latin name "nomen," which means "name." With the help of nominal data, we can't do any numerical tasks or can't give any order to sort the data. These data don't have any meaningful order; their values are distributed to distinct categories.

Examples of Nominal Data are Colour of hair (Blonde, Red, Brown, Black, etc.); Marital status (Single, Widowed, Married); Nationality (Indian, German, American); Gender (Male, Female, Others) and Eye Color (Black, Brown, etc.)

b. **Ordinal Data**

Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can't do any arithmetical tasks on them.

The ordinal data is qualitative data for which their values have some kind of relative position. These kinds of data can be considered as "in-between" the qualitative data and quantitative data. The ordinal data only shows the sequences and cannot use for statistical analysis. Compared to the nominal data, ordinal data have some kind of order that is not present in nominal data.

Examples of Ordinal Data are when companies ask for feedback, experience, or satisfaction on a scale of 1 to 10, letter grades in the exam (A, B, C, D, etc.), ranking of peoples in a competition (First, Second, Third, etc.), economic status (High, Medium, and Low), education level (Higher, Secondary, Primary)

Difference between Nominal and Ordinal Data

- ◊ Nominal data can't be quantified, neither they have any intrinsic ordering
- ◊ Ordinal data give some kind of sequential order by their position on the scale
- ◊ Nominal data is qualitative data or categorical data
- ◊ Ordinal data is said to be "in-between" of qualitative data and quantitative data
- ◊ Nominal data don't provide any quantitative value, neither we can perform any arithmetical operation
- ◊ Ordinal data provide sequence and can assign numbers to ordinal data but cannot perform the arithmetical operation
- ◊ Nominal data cannot be used to compare with one another

◊ Ordinal data can help to compare one item with another by ranking or ordering

Examples: Eye colour, housing style, gender, hair colour, religion, marital status, ethnicity, etc

Example: Economic status, customer satisfaction, education level, letter grades, etc

B. Quantitative Data

Quantitative data can be expressed in numerical values, which makes it countable and includes statistical data analysis. These kinds of data are also known as Numerical data. It answers the questions like, "how much," "how many," and "how often." For example, the price of a phone, the computer's ram, the height or weight of a person, etc., falls under the quantitative data.

Quantitative data can be used for statistical manipulation and these data can be represented on a wide variety of graphs and charts such as bar graphs, histograms, scatter plots, boxplot, pie charts, line graphs, etc.

Examples of Quantitative Data are Height or weight of a person or object, Room Temperature, Scores and Marks (Ex: 59, 80, 60, etc.) and Time

The Quantitative data are further classified into two parts:

a. **Discrete Data**

The term discrete means distinct or separate. The discrete data contain the values that fall under integers or whole numbers. The total number of students in a class is an example of discrete data. These data can't be broken into decimal or fraction values.

The discrete data are countable and have finite values; their subdivision is not possible. These data are represented mainly by a bar graph, number line, or frequency table.

Examples of Discrete Data:

◊ Total numbers of students present in a class

◊ Cost of a cell phone

◊ Numbers of employees in a company

◊ The total number of players who participated in a competition

◊ Days in a week

b. **Continuous Data**

Continuous data are in the form of fractional numbers. It can be the version of an android phone, the height of a person, the length of an object, etc. Continuous data represents information that can be divided into smaller levels. The continuous variable can take any value within a range.

The key difference between discrete and continuous data is that discrete data contains the integer or whole number. Still, continuous data stores the fractional numbers to record different data such as temperature, height, width, time, speed, etc.

Examples of Continuous Data : Height of a person, Speed of a vehicle, "Time-taken" to finish the work , Wi-Fi Frequency, Market share price

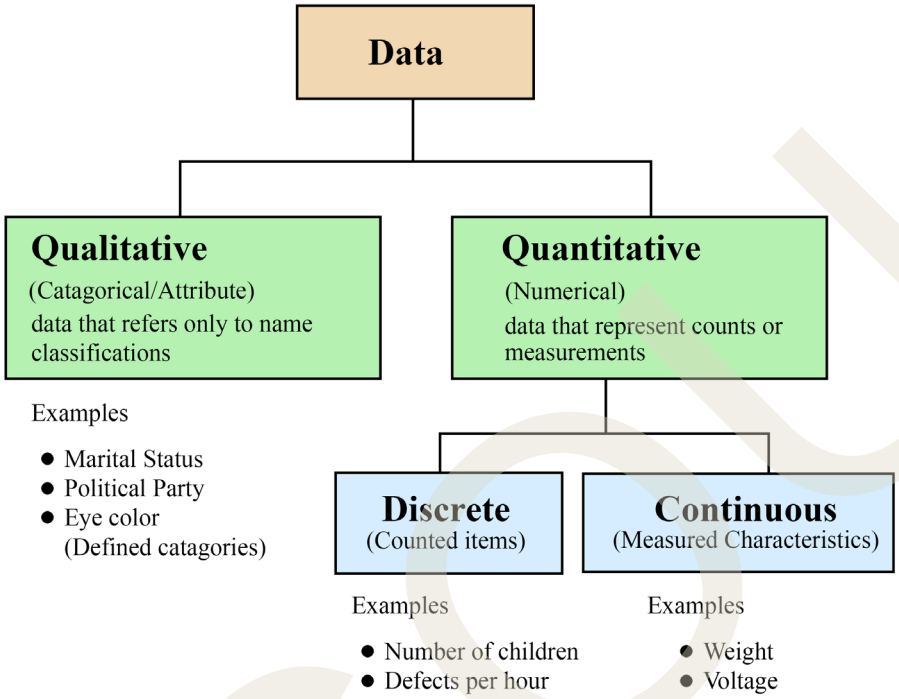


Fig 1.3.1 Types of Data

1.3.2 Data Collection

Data collection is the process of collecting and evaluating information or data from multiple sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities. It is an essential phase in all types of research, analysis, and decision-making, including that done in the social sciences, business, and healthcare. Accurate data collection is necessary to make informed business decisions, ensure quality assurance, and keep research integrity. During data collection, the researchers must identify the data types, the sources of data, and what methods are being used. We will soon see that there are many different data collection methods. There is heavy reliance on data collection in research, commercial, and government fields. Before an analyst begins collecting data, they must answer three questions first. They are;

- ◊ What is the goal or purpose of data collection?
- ◊ What kinds of data are they planning on gathering?
- ◊ What methods and procedures will be used to collect, store, and process the information?

1.3.3 Data Collection Methods

Primary and secondary methods of data collection are two approaches used to gather information for research or analysis purposes. Let's explore each data collection method in detail.

1.3.3.1 Primary Data Collection

Primary data collection involves the collection of original data directly from the source or through direct interaction with the respondents. This method allows researchers to obtain firsthand information specifically tailored to their research objectives.

Techniques of Primary data collection

There are various techniques for primary data collection. They are as follows:

a. Surveys and Questionnaires: Researchers design structured questionnaires or surveys to collect data from individuals or groups. These can be conducted through face-to-face interviews, telephone calls, mail, or online platforms.

b. Interviews: Interviews involve direct interaction between the researcher and the respondent. They can be conducted in person, over the phone, or through video conferencing. Interviews can be structured (with predefined questions), semi-structured (allowing flexibility), or unstructured (more conversational).

c. Observations: Researchers observe and record behaviors, actions, or events in their natural setting. This method is useful for gathering data on human behavior, interactions, or phenomena without direct intervention.

d. Experiments: Experimental studies involve the manipulation of variables to observe their impact on the outcome. Researchers control the conditions and collect data to draw conclusions about cause-and-effect relationships.

e. Focus Groups: Focus groups bring together a small group of individuals who discuss specific topics in a moderated setting. This method helps in understanding opinions, perceptions, and experiences shared by the participants.

1.3.3.2 Secondary Data Collection

Secondary data collection involves using existing data collected by someone else for a purpose different from the original intent. Researchers analyse and interpret this data to extract relevant information.

Sources of Secondary Data Collection

Secondary data can be obtained from various sources, including:

a. Published Sources: Researchers refer to books, academic journals, magazines, newspapers, government reports, and other published materials that contain relevant data.

b. Online Databases: Numerous online databases provide access to a wide range of secondary data, such as research articles, statistical information, economic data, and social surveys.

c. Government and Institutional Records: Government agencies, research institutions, and organisations often maintain databases or records that can be used for research purposes.

d. Publicly Available Data: Data shared by individuals, organisations, or communities on public platforms, websites, or social media can be accessed and utilised for research.

e. Past Research Studies: Previous research studies and their findings can serve as valuable secondary data sources. Researchers can review and analyse the data to gain insights or build upon existing knowledge.

1.3.4 The Importance/Need for Ensuring Accurate and Appropriate Data Collection

Accurate data collecting is crucial to preserving the integrity of research, regardless of the subject of study or preferred method for defining data (quantitative, qualitative). Errors are less likely to occur when the right data gathering tools are used (whether they are brand-new ones, updated versions of them, or already available).

Following are the effects of data collection done incorrectly,

◊ Erroneous conclusions that squander resources

◊ Decisions that compromise public policy

◊ Incapacity to correctly respond to research inquiries

◊ Bringing harm to participants who are humans or animals

◊ Deceiving other researchers into pursuing futile research avenues

◊ The study's inability to be replicated and validated

When these study findings are used to support recommendations for public policy, there is the potential to result in disproportionate harm, even if the degree of influence from flawed data collecting may vary by discipline and the type of investigation.

1.3.5 Challenges in Data Collection

There are some prevalent challenges faced while collecting data, let us explore a few of them to understand them better and avoid.

a. Data Quality Issues: The main threat to the broad and successful application of machine learning is poor data quality. Data quality must be your top priority if you want to make technologies like machine learning work for you. Let's talk about some of the most prevalent data quality problems in this blog article and how to fix them.

b. Inconsistent Data: When working with various data sources, it's conceivable that the same information will have discrepancies between sources. The differences could be in formats, units, or occasionally spellings. The introduction of inconsistent data might also occur during firm mergers or relocations. Inconsistencies in data have a tendency to accumulate and reduce the value of data if they are not continually resolved. Organisations that have heavily focused on data consistency do so because they only want reliable data to support their analytics.

c. Data Downtime: Data is the driving force behind the decisions and operations of data-driven businesses. However, there may be brief periods when their data is unreliable or not prepared. Customer complaints and subpar analytical outcomes are only two ways that this data unavailability can have a significant impact on businesses. A data engineer spends about 80% of their time updating, maintaining, and guaranteeing the integrity of the data pipeline. In order to ask the next business question, there is a high marginal cost due to the lengthy operational lead time from data capture to insight.

   Schema modifications and migration problems are just two examples of the causes of data downtime. Data pipelines can be difficult due to their size and complexity. Data downtime must be continuously monitored, and it must be reduced through automation.

d. Ambiguous Data: Even with thorough oversight, some errors can still occur in massive databases or data lakes. For data streaming at a fast speed, the issue becomes more overwhelming. Spelling mistakes can go unnoticed, formatting difficulties can occur, and column heads might be deceptive. This unclear data might cause a number of problems for reporting and analytics.

e. Duplicate Data: Streaming data, local databases, and cloud data lakes are just a few of the sources of data that modern enterprises must contend with. They might also have application and system silos. These sources are likely to duplicate and overlap each other quite a bit. For instance, duplicate contact information has a substantial impact on customer experience. If certain prospects are ignored while others are engaged repeatedly, marketing campaigns suffer. The likelihood of biased analytical outcomes increases when duplicate data are present. It can also result in ML models with biased training data.

f. Too Much Data: While we emphasize data-driven analytics and its advantages, a data quality problem with excessive data exists. There is a risk of getting lost in an abundance of data when searching for information pertinent to your analytical efforts. Data scientists, data analysts, and business users devote 80% of their work to finding and organising the appropriate data. With an increase in data volume, other problems with data quality become more serious, particularly when dealing with streaming data and big files or databases.

g. Inaccurate Data: For highly regulated businesses like healthcare, data accuracy is crucial. Given the current experience, it is more important than ever to increase the data quality for COVID-19 and later pandemics. Inaccurate information

does not provide you with a true picture of the situation and cannot be used to plan the best course of action. Personalized customer experiences and marketing strategies underperform if your customer data is inaccurate.

Data inaccuracies can be attributed to a number of things, including data degradation, human mistake, and data drift. Worldwide data decay occurs at a rate of about 3% per month, which is quite concerning. Data integrity can be compromised while being transferred between different systems, and data quality might deteriorate with time.

h. Hidden Data: The majority of businesses only utilise a portion of their data, with the remainder sometimes being lost in data silos or discarded in data graveyards. For instance, the customer service team might not receive client data from sales, missing an opportunity to build more precise and comprehensive customer profiles. Missing out on possibilities to develop novel products, enhance services, and streamline procedures is caused by hidden data.

i. Finding Relevant Data: Finding relevant data is not so easy. There are several factors that we need to consider while trying to find relevant data, which include Relevant Domain, Relevant demographics and Relevant Time period and so many more factors that we need to consider while trying to find relevant data. Data that is not relevant to our study in any of the factors render it obsolete and we cannot effectively proceed with its analysis. This could lead to incomplete research or analysis, re-collecting data again and again, or shutting down the study.

j. Deciding the Data to Collect: Determining what data to collect is one of the most important factors while collecting data and should be one of the first factors while collecting data. We must choose the subjects the data will cover, the sources we will be used to gather it, and the quantity of information we will require. Our responses to these queries will depend on our aims, or what we expect to achieve utilising your data. As an illustration, we may choose to gather information on the categories of articles that website visitors between the ages of 20 and 50 most frequently access. We can also decide to compile data on the typical age of all the clients who made a purchase from your business over the previous month. Not addressing this could lead to double work and collection of irrelevant data or ruining your study as a whole.

k. Dealing With Big Data: Big data refers to exceedingly massive data sets with more intricate and diversified structures. These traits typically result in increased challenges while storing, analysing, and using additional methods of extracting results. Big data refers especially to data sets that are quite enormous or intricate that conventional data processing tools are insufficient. The overwhelming amount of data, both unstructured and structured, that a business faces on a daily basis.

The amount of data produced by healthcare applications, the internet, social networking sites social, sensor networks, and many other businesses are rapidly

growing as a result of recent technological advancements. Big data refers to the vast volume of data created from numerous sources in a variety of formats at extremely fast rates. Dealing with this kind of data is one of the many challenges of Data Collection and is a crucial step toward collecting effective data.

l. Low Response and Other Research Issues: Poor design and low response rates were shown to be two issues with data collecting, particularly in health surveys that used questionnaires. This might lead to an insufficient or inadequate supply of data for the study. Creating an incentivized data collection program might be beneficial in this case to get more responses.

1.3.6 Steps in the Data Collection Process

In the Data Collection Process, there are 5 key steps. They are explained briefly below .

1. Decide What Data You Want to Gather

The first thing that we need to do is decide what information we want to gather. We must choose the subjects the data will cover, the sources we will use to gather it, and the quantity of information that we would require. For instance, we may choose to gather information on the categories of products that an average e-commerce website visitor between the ages of 30 and 45 most frequently searches for.

2. Establish a Deadline for Data Collection

The process of creating a strategy for data collection can now begin. We should set a deadline for our data collection at the outset of our planning phase. Some forms of data we might want to continuously collect. We might want to build up a technique for tracking transactional data and website visitor statistics over the long term, for instance. However, we will track the data throughout a certain time frame if we are tracking it for a particular campaign. In these situations, we will have a schedule for when we will begin and finish gathering data.

3. Select a Data Collection Approach

We will select the data collection technique that will serve as the foundation of our data gathering plan at this stage. We must take into account the type of information that we wish to gather, the time period during which we will receive it, and the other factors we decide on to choose the best gathering strategy.

4. Gather Information

Once our plan is complete, we can put our data collection plan into action and begin gathering data. In our DMP, we can store and arrange our data. We need to be careful to follow our plan and keep an eye on how it's doing. Especially if we are collecting data regularly, setting up a timetable for when we will be checking in on how our data gathering is going may be helpful. As circumstances alter and we learn new details, we might need to amend our plan.

5. Examine the Information and Apply Your Findings

It's time to examine our data and arrange our findings after we have gathered all of our information. The analysis stage is essential because it transforms unprocessed data into insightful knowledge that can be applied to better our marketing plans, goods, and business judgments. The analytics tools included in our DMP can be used to assist with this phase. We can put the discoveries to use to enhance our business once we have discovered the patterns and insights in our data.

Recap

1. Data:
    - Gardeners collected data on soil types, watering patterns, and sunflower growth
    - Data included numerical values, text, images, and time-series information
    - Data can help businesses understand customers, improve campaigns, and grow
2. Characteristics of Data:
    - Data are facts obtained through observation, measurement, and counting
    - Data can come from external and internal sources
    - Data can be an automatic byproduct of routine operations
    - Attention must be paid to the data source to ensure quality
3. Formats of Data:
    - Data can be stored and processed as text, numbers, audio, pictures, and video
4. Types of Data:
    - Qualitative/Categorical Data: Nominal and Ordinal
    - Quantitative Data: Discrete and Continuous
5. Data Collection Methods:
    - Primary Data: Surveys, interviews, observations, experiments,

focus groups

- Secondary Data: Published sources, online databases, government/institutional records, past research

6. Importance of Accurate Data Collection:
   - Ensures integrity of research and decision-making
   - Helps avoid erroneous conclusions and harmful outcomes
7. Challenges in Data Collection:
   - Data quality issues, inconsistent data, data downtime, ambiguous data, duplicate data, too much data, inaccurate data, hidden data, finding relevant data, deciding what data to collect, dealing with big data, low response rates
8. Steps in the Data Collection Process:
   - Decide what data to gather
   - Establish a deadline for data collection
   - Select a data collection approach
   - Gather the information
   - Analyse the data and apply the findings

Objective Questions

1. What is the term used to refer to data that can be measured or counted in the form of numbers?
2. Which type of data has a natural ordering, but cannot be used for statistical analysis?
3. What is the name of the American mathematician who introduced the concept of data in computing?
4. What is the term used to refer to data that cannot be measured or counted in the form of numbers?
5. Which type of data consists of distinct or separate values that fall under integers or whole numbers?

6. What is the term used to refer to data that can be divided into smaller levels and represented by fractional numbers?
7. What is the term used to refer to data that can be labeled without any order or quantitative value?
8. Which data collection method involves the manipulation of variables to observe their impact on the outcome?
9. What is the term used to refer to data collected from existing sources for a purpose different from the original intent?
10. What is the term used to refer to data collected directly from the source or through direct interaction with respondents?
11. Which data collection method involves direct interaction between the researcher and the respondent?
12. What is the term used to refer to the process of collecting and evaluating information or data from multiple sources?

Answers

1. Quantitative
2. Ordinal
3. Claude Shannon
4. Qualitative
5. Discrete
6. Continuous
7. Nominal
8. Experiments
9. Secondary data
10. Primary data
11. Interviews
12. Data collection

Assignments

1. What are the two main types of data?
2. What are the characteristics of data?
3. What are the different formats in which data can be stored and processed?
4. What are the differences between nominal and ordinal data?
5. What are the two main methods of data collection?
6. What are the various techniques of primary data collection?
7. What are the different sources of secondary data collection?
8. Why is ensuring accurate and appropriate data collection important?
9. Explain the steps involved in the data collection process.
10. Discuss the common challenges faced during the data collection process and suggest strategies to address them.

Suggested Reading

1. Rea, L. M., & Parker, R. A. (2014). Designing and conducting survey research: A comprehensive guide (4th ed.). Jossey-Bass.
2. Fowler, F. J. (2013). Survey research methods (5th ed.). SAGE Publications.
3. Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE Publications.
4. Levy, P. S., & Lemeshow, S. (2013). Sampling of populations: Methods and applications (4th ed.). Wiley.
5. Babbie, E. (2016). The practice of social research (14th ed.). Cengage Learning.

Reference

1. Miller, D. C., & Salkind, N. J. (2002). Handbook of research design and social measurement (6th ed.). SAGE Publications.
2. Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, phone, mail, and mixed-mode surveys: The tailored design method (4th ed.). Wiley.
3. Bernard, H. R. (2017). Research methods in anthropology: Qualitative and quantitative approaches (6th ed.). Rowman & Littlefield.
4. Salant, P., & Dillman, D. A. (1994). How to conduct your own survey. Wiley.
5. Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). Survey methodology (2nd ed.). Wiley.

BLOCK - 02
Statistical Measures

Unit - 1

Measures of Central Tendency

Learning Outcomes

Upon the completion of this unit, the learner will be able to:

- ◊ familiarise the concept and significance of measures of central tendency (mean, median, mode) in describing and analysing data.
- ◊ get an idea regarding the computational methods and formulas for calculating different measures of central tendency (mean, median, mode) for individual, discrete, and continuous data sets.
- ◊ explore the empirical relationships and patterns among the mean, median, and mode, and their implications for understanding the shape and characteristics of a data distribution.
- ◊ recognise the applications and importance of measures of central tendency in various business decisions.

Prerequisite

Imagine you're a teacher, and you want to analyse the test scores of your students to get a sense of their overall performance. You collect the following test scores from your class:

85, 92, 78, 82, 90, 85, 80, 88

As you look at this data, you realise that you need a way to summarise the typical or central test score in your class. This is where measures of central tendency come into play.

The first measure of central tendency you consider is the mean, or the average. To calculate the mean, you add up all the test scores (85 + 92 + 78 + 82 + 90 + 85 + 80 + 88 = 680) and then divide by the total number of students (8). This gives you 85 as the average test score. The mean provides a good overall sense of your students' performance, but you notice that there are a few scores that are quite a

bit higher or lower than the average. This leads you to consider another measure of central tendency: the median. To find the median, you first need to arrange the test scores in numerical order: 78, 80, 82, 85, 85, 88, 90, 92. Since there are 8 scores, the median is the middle value, which is 85. The median tells you that half the students scored 85 or below, and half scored 85 or above. Finally, you look at the mode, which is the value that appears most frequently in the data. In this case, the mode is 85, as that score appears twice. As you reflect on these measures of central tendency, you realise they each provide a slightly different perspective on your students' performance:

The mean of 85 gives you an overall average test score.

The median of 85 indicates the middle score, with half the students above and half below.

The mode of 85 shows the most common or frequent test score.

These measures help you understand the central or typical performance of your students, which can be valuable information as you make decisions about teaching strategies, grading policies, and student interventions. By considering the different measures of central tendency, you gain a more comprehensive understanding of the patterns and trends in your students' test scores.

This example illustrates how measures of central tendency can be used in an educational context to analyse and summarise student performance data. The same principles can be applied to a wide range of other datasets, such as sales figures, survey responses, or financial data, to gain insights into the typical or central values within the data.

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

Keywords

Mean, Arithmetic Mean, Geometric Mean, Harmonic Mean, Median, Mode

Discussion

The knowledge regarding properties of data is essential for describing data such as measures of central tendency, measures of dispersion and measures of shape. This unit focuses on measures of central tendency. There is a tool which represents the basic features of data is referred to as "average". An average value is a single value that describes an entire group of values. In other words, it is a single value within the range of data that is used to represent all the values in the series. In simple terms, the average of a statistical series is the value of the variable which is representative of the entire series.

2.1.1 Measures of Central Tendencies

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

In statistics there are various type of measures of central tendency, broadly classified as follows:

1. Mathematical Averages
    a. Mean or Arithmetic Mean
        i. Simple
        ii. Weighted
    b. Geometric Mean
    c. Harmonic Mean
2. Positional Averages
    a. Median
    b. Mode

Mathematical Averages

Mean or Arithmetic Mean

Arithmetic mean of a set of observations is their sum divided by the number of observations. It is generally denoted by $\bar{x}$ or AM. Population mean is denoted by μ. Therefore,

$$AM = \frac{\text{Sum of all the Observations}}{\text{Number of Observations}}$$

Arithmetic mean is of two types:

◊ Simple Arithmetic mean

◊ Weighted Arithmetic mean

a. Calculation of Simple Arithmetic Mean

Arithmetic mean can be computed differently for three different types of series. In other words, Arithmetic mean can be computed differently for individual series, discrete frequency distribution and continuous frequency distribution.

i. Computation of Arithmetic mean for Individual Series: The Arithmetic mean of an individual series can be calculated by dividing the sum of the observations by the number of the observations. In a series $x_1 + x_2 + \cdots . + x_n$ the arithmetic mean can be calculated by:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} xi$$

Illustration 2.1.1

A rainwear manufacturing company wants to launch some new products in the State. The rainfall in the State (cm) for the past 10 years is given in the table below. Find the average rainfall in the State for the last 10 years.

Table No:1 Rainfall for the last 10 years

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Rainfall (cm)	30	27	50	76	37	14	34	75	60	121

Solution:

For computing average rainfall, We have to first compute the total rainfall in 10 years and then divide this total by the total number of years, i.e., 10 years.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} xi$$

Year	Rainfall (x)
2005	110
2006	120
2007	130
2008	135
2009	140
2010	150
2011	160
2012	170
2013	180
2014	190
Total	**1485**

$$\therefore \bar{x} = \frac{1485}{10} = 148.5$$

Hence, the average rainfall in 10 years in the State is 148.5 cm.

ii. Computation of Arithmetic mean for discrete frequency distribution: In this type of distribution every term is multiplied by its corresponding frequency and the total sum of these products is divided by the sum of frequencies. Hence for a given series $x_1 + x_2 + \cdots + x_n$ with corresponding frequencies $f_1 + f_2 + \ldots + f_n$, the Arithmetic mean is given by:

$$\text{AM} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i xi}{\sum_{i=1}^{n} f_i}$$

Illustration.2.1.2

The weekly earnings of 187 employees of a company is given in the following table. Find the mean of the weekly earnings.

Weekly earnings of 187 employees

Weekly earnings	100	120	140	160	180	200	210
No. of Employees	5	8	12	16	22	44	80

Solution:

In discrete series, each value must be multiplied by its frequency. The next step is to divide the sum of this product by the total sum of the frequencies to arrive at the arithmetic mean.

Table: Product of the weekly earnings and number of employees

Weekly Earnings (in Rs.) (x)	**Number of Employees (f)**	**f × x**
100	5	500
120	8	960
140	12	1680
160	16	2560
180	22	3960
200	44	8800
210	80	16800
Total	$\sum f = 187$	$\sum f \times x = 35260$

$$AM = \frac{\sum fx}{\sum f} = \frac{35260}{187} = \text{Rs. } 188.55$$

Hence, the average weekly earning is Rs. 188.55

iii. Computation of Arithmetic mean for continuous frequency distribution: The process of computing arithmetic mean for a continuous frequency distribution is the same as the process of computing arithmetic mean for a discrete series. In continuous frequency distribution, the class intervals are given. We take the midpoint of each class interval as the value of *x*. The remaining step are the same as that for computing the arithmetic mean of a discrete series. The formula given below is used for computing the arithmetic mean for a continuous series.

$$AM = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Illustration.2.1.3

From the following table, find the arithmetic mean.

Table: Data related to Class intervals and frequencies

Class Intervals	Frequencies
0-10	5
10-20	7
20-30	19
30-40	12
40-50	5
50-60	2
60-70	7

Solution:

For computing arithmetic mean in a continuous frequency distribution, we need to compute the midpoint of class intervals (x). The midpoints are multiplied by the corresponding frequencies (f_x). The sum of this product is obtained and is divided by the sum of frequencies.

Table: Data given in the question is arranged with class midpoint, frequencies and the product of midpoints and frequencies.

Size	Midpoint (x)	Frequencies (f)	(f × x)
0-10	5	5	25
10-20	15	7	105
20-30	25	19	475
30-40	35	12	420
40-50	45	5	225
50-60	55	2	110
60-70	65	7	455
Total		$\sum f = 57$	$\sum (f \times x) = 1815$

$$AM = \frac{\sum fx}{\sum f} = \frac{1815}{57} = 31.84$$

Merits and Demerits of Arithmetic Mean

It is important to understand merits and demerits to have an idea about its appropriate application. Some of the merits and demerits of arithmetic mean are listed below.

Merits

1. It is rigidly defined.
2. It is easy to calculate and understand.
3. It is based upon all the observations.
4. It is capable of further algebraic treatment.
5. Arrangement of data in ascending or descending order is not necessary.
6. Of all the averages, arithmetic mean is the least affected by fluctuations of sampling

Demerits

1. It is highly affected by extreme values.
2. It cannot be determined by inspection.
3. When dealing with qualitative characteristics, such as intelligence, honesty, beauty, etc., arithmetic mean cannot be used. In such cases, median can be used.
4. In extremely asymmetrical (skewed) distributions, the arithmetic mean is not a suitable measure.

b. Weighted Arithmetic Mean

In the computation of Arithmetic Mean, equal importance is given to all the items of a series. However, there are cases where all the items are not of equal importance and importance itself is relative by nature. In other words, some items of a series are more important as compared to the other items in the same series. In such cases, it becomes important to assign different weight to different items. The weighted mean can be used to calculate an average that takes into account the importance of each value with respect to the overall total. For example, to get an idea of the change in the cost of living of a certain group of people, a simple mean of the prices of the commodities consumed by them will not be an appropriate tool for measuring average price, as all the commodities may not be of equal importance. For instance, wheat, rice, and pulses may be more important when compared with cigarettes and other luxury items.

The formula for calculating weighted mean is;

$$\bar{x}_w = \frac{\sum(w \times x)}{\sum w}$$

Where 'x' is the value of the item and 'w' the weights attached to the corresponding items.

Illustration.2.1.4

A company manufactures three types of products: A, B, and C. The revenue generated by each product and the corresponding weights (based on the company's strategic importance) are given in the following table:

Product	Revenue (in millions $)	Weight
A	20	0.4
B	15	0.3
C	25	0.2

Calculate the weighted mean revenue for the company.

Solution:

To solve this problem, we need to apply the weighted mean formula:

Product	Revenue (in millions $) (x)	Weight (w)	x x w
A	20	0.4	8
B	15	0.3	4.5
C	25	0.2	5
		$\sum w = 0.9$	$\sum w \times x = 17.5$

$$\bar{x}_w = \frac{\sum(w \times x)}{\sum w}$$

$$= \frac{17.5}{0.9} = 19.4$$

Therefore, the weighted mean revenue for the company is $19.44 million.

Geometric Mean

The geometric mean is the n^{th} root of the product of n observations in the data set. The fundamental formula for its computation is:

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

When there are more than two items, the computation is simplified by using a logarithm. The formula above can be expressed as:

$$GM = \text{Antilog of } \frac{1}{N}(\log x_1 + log\ x_2 + \cdots + log\ x_n)$$

When there is a multiplicative relationship between the data or when the data is compounded, the geometric mean performs well. When the data is nonlinear and particularly when a log transformation is used, geometric mean is used.

Advantages of geometric mean

The following are the advantages of geometric mean:

i. The Geometric Mean is significant since it gives less weight to extreme numbers. As a result, the impact of extremely small and extremely high values is minimised.
ii. It can be further algebraically treated.
iii. It can be used to calculate average, percentage changes, ratios, etc.
iv. It is based on all the observation of the series.
v. It can be used to measure relative changes.
vi. The best average in the construction of index numbers is the geometric mean.
vii. It is rigidly defined.

Limitations of geometric mean

The geometric mean has the following limitations:

i. The geometric mean will not be calculated if some of the observations are negative.
ii. It is tough for a layman to comprehend.
iii. If one or more observations are zero, the geometric mean computation is meaningless because the observation's product is always 0, and hence the geometric mean is zero.
iv. It can sometimes give a value that is not in the series.

A. For individual series

$$GM = \text{antilog of } \frac{\Sigma \log x}{N}$$

Illustration.2.1.5

Wages of 10 workers in a factory given below

85, 15, 500, 70, 75, 250, 45, 8, 36, 40

Find geometric mean.

Solution

x	log x
85	1.9294
15	1.1761
500	2.6990
70	1.8451
75	1.8751
250	2.3979
45	1.6532
8	0.9031
36	1.5563
40	1.6021
	$\sum$ log x = 17.6373

The value of log x is determined from logarithm table

GM = antilog of $\frac{\sum \log x}{N}$

= antilog of $\frac{17.6373}{10}$

= antilog of 1.76373

= 58.04

B. For discrete series

GM = antilog of $\frac{\sum (f \times \log x)}{N}$

Where,

x – Value of the variable

N – Number of items

f – Frequency

Illustration.2.1.6

Following are the price and demand for Apple per kilogram. Find geometric mean of the following.

Price	130	350	260	250	175	150
Demand	12	2	4	5	8	10

Solution

x	**f**	**log x**	**f × log x**
130	12	2.1139	25.3668
350	2	2.5441	5.0882
260	4	2.4150	9.66
250	5	2.3979	11.9895
175	8	2.2430	17.944
150	10	2.1761	21.761
	41		**∑f × log x =91.8095**

Geometric Mean = antilog of $\frac{\Sigma(f \times \log x)}{N}$

= antilog of $\frac{91.8095}{41}$

= antilog of 2.2392

=173.5

C. **For continuous series**

Geometric Mean = antilog of $\frac{\Sigma(f \times \log x)}{N}$

Where, x – Mid value

Illustration 2.1.7

From the following data calculate Geometric mean

Income (in' 000)	0-10	10-20	20-30	30-40
No. of families	5	8	3	4

Solution

Income	**Mid value (x)**	**f**	**log x**	**f × log x**
0-10	5	5	0.6990	3.4950
10-20	15	8	1.1761	9.4088
20-30	25	3	1.3979	4.1937
30-40	35	4	1.5441	6.1764
		20		**23.2739**

$$\text{GM} = \text{antilog of } \frac{\Sigma f \times \log x}{N}$$

$$= \text{antilog of } \frac{23.2739}{20}$$

$$= \text{antilog of } 1.1637$$

$$= 14.58$$

Income is shown in thousands. Therefore, the Geometric Mean = 14580

Harmonic Mean

The Harmonic mean of a number of observations is the reciprocal of the arithmetic mean of the reciprocal of the given observations.

Harmonic mean can be defined as "the reciprocal of the arithmetic average of the reciprocal of the value of a variable".

$$\therefore \text{H.M.} = \frac{1}{\frac{1}{N}\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)} = \frac{N}{\Sigma\frac{1}{x}}$$

Where,

H.M. – Harmonic Mean

N – Number of items

x – Value of the variable

When we wish to average units like speed, rates, and ratios, we use the harmonic mean.

Advantages of harmonic mean

The benefits of harmonic mean are as follows:

i. It gives the smallest item the most weight.
ii. It is quite beneficial for averaging certain ratios and rates because it measures relative changes.
iii. It is based on all the observations.
iv. It is rigidly defined.
v. It is possible to calculate it even if a series contains any negative numbers.

Limitations of harmonic mean

The following are the limitations of harmonic mean:

i. It is very difficult to calculate.
ii. It does not accurately reflect the statistical series.
iii. It is tough for a layman to comprehend.
iv. It is impossible to calculate if any of the items are zero.
v. This is merely a summary figure; the actual item in the series may not be shown.
vi. It has a very limited algebraic treatment.

A. For individual observation

$$H.M = \frac{N}{\sum \frac{1}{x}}$$

Illustration.2.1.8

The speeds of five buses in a city are given below.

Speed (km/hr)	15	18	20	22	17

Find the average speed.

Solution

x	$\frac{1}{x}$
15	0.0666
18	0.0555
20	0.05
22	0.04545
17	0.05882
$\sum \frac{1}{x} = 0.27637$	

$$H.M = \frac{N}{\sum \frac{1}{x}}$$

$$= \frac{5}{0.27637}$$

$$= 18.09$$

Average speed = 18.09 km/hr

B. For discrete observation

$$H.M = \frac{N}{\sum f \times \frac{1}{x}}$$

Illustration.2.1.9

Calculate the harmonic mean of the scores on the English class test, as shown below.

Marks	11	12	13	14	15
No. of students	8	7	4	5	2

Solution

X	f	$\frac{1}{x}$	$f \times \frac{1}{x}$
11	8	0.0909	0.7272
12	7	0.0833	0.5831
13	4	0.0769	0.3076
14	5	0.0714	0.3570
15	2	0.0666	0.1332
	26		**2.1081**

$$H.M = \frac{N}{\sum f \times \frac{1}{x}}$$

$$= \frac{26}{2.1081}$$

$$= 12.33$$

C. For continuous observations

$$H.M = \frac{N}{\sum f \times \frac{1}{x}}$$

Where,

x – Mid Value of the classes

Illustration.2.1.10

From the following data, calculate Harmonic mean

Class	0-10	10-20	20-30	30-40
Frequency	2	3	4	2

Solution

Class	Mid Value (x)	F	$\frac{1}{x}$	$f \times \frac{1}{x}$
0-10	5	2	0.2	0.4
10-20	15	3	0.066	0.2
20-30	25	4	0.04	0.16
30-40	35	2	0.0285	0.0571
		11		**0.8171**

$$H.M = \frac{N}{\sum f \times \frac{1}{x}}$$

$$= \frac{11}{0.8171}$$

$$= 13.462$$

2.1.2 Positional Averages

The term "positional average" refers to an average calculated from a set of observations in which one number based on its position is chosen to represent the complete set. The positional averages discussed in this unit are listed below.

a. Median

b. Mode

Median

When data is arranged in ascending or descending order, the median is the value of the data's middle most observations.

Median may be defined as "that value of the variable which divides the group into two equal parts, one part comprising all the values greater and other, all values being less than the median". The number of data points affects how the median is calculated. The middle value is the median for an odd number of data, and the average of the two middle values is the median for an even amount of data.

Advantages of median

i. It is rigidly defined.

ii. It is simple to calculate.

iii. Extreme values or outliers have no effect on it.

iv. It is possible to calculate it for open-ended classes.

v. When dealing with qualitative data where no numerical measurements are provided but it is possible to rank the objects in some order, the median is the only average that can be employed.

vi. The absolute sum of the individual values' deviations from the median is always the minimum.

Disadvantages of median

i. Median is not suitable for further mathematical treatment

ii. In the case of ungrouped data with an even number of observations, the median cannot be estimated precisely. The arithmetical average of the two middle elements is the median in this case.

iii. It sometimes generates a value that is not seen anywhere else in the series.

iv. To calculate the median, the data must be arranged in ascending or descending order.

v. The median is less stable than the mean, especially in small samples.

vi. The median, as a positional average, does not take into account every single item in the distribution.

Computation of median

1. **For individual series**

Steps

i. Sort the data into ascending or descending order.

ii. Use the formula.

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

Illustration.2.1.11

The marks obtained by a student in five examinations are given below.

Marks	35	37	25	28	40

Find the median mark.

Solution

Arrange the data in ascending order

25	28	35	37	40

Apply the formula

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

$$= \left(\frac{5+1}{2}\right)^{th} \text{ item}$$

$$= 3^{rd} \text{ item}$$

The 3rd item in the series is 35.

∴ Median mark is 35

Illustration.2.1.12

The following table shows the income of six families. Find their median income.

Income(Rs)	10000	12000	11000	20000	15000	17000

Solution

Arrange the data in ascending order

10000	11000	12000	15000	17000	20000

Apply the formula

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

$$= \left(\frac{6+1}{2}\right)^{th} \text{ item}$$

$$= 3.5^{th} \text{ item}$$

However, there is not a single item in the series with a position of 3.5. As a result, we use the median as the average of the third and fourth elements in the series.

Median = Mean of 3rd and 4th item

$$= \frac{12000+15000}{2}$$

$$= \frac{27000}{2}$$

$$= 13500$$

Median income of the family is 13500.

2. For discrete series

Steps

i. Arrange the data in ascending or descending order

ii. Calculate cumulative frequency (cf)

iii. Determine $\frac{N+1}{2}$

Where N is the total frequency

iv. Median is the value for the $\left(\frac{N+1}{2}\right)^{th}$ item of the data

Illustration.2.1.13

The daily wage of 115 employees is shown in the table below. Find out what the median wage is.

Wage	500	600	700	800	900	1000	1100	1200
No. of employees	8	14	15	18	20	15	14	11

Solution

Wages	**f**	**cf**
500	8	8
600	14	22
700	15	37
800	18	55
900	20	75
1000	15	90
1100	14	104
1200	11	115
	N=115	

$$\text{Median} = \frac{N+1}{2}$$
$$= \frac{115+1}{2}$$
$$= \frac{116}{2}$$
$$= 58^{th} \text{ item}$$

∴ Median is the value in the data which comes in the 58th position, which is the value of the item having cumulative frequency 58. Since cumulative frequency of 58 comes under the cumulative frequency 75, median is the value in the data that comes in the 75th position,

∴ Median = 900

3. For continuous series

Steps

i. Convert inclusive classes to the exclusive class (if any)

ii. Calculate the cumulative frequencies (cf)

iii. Calculate $\frac{N}{2}$, where N is the total frequency

iv. Identify the class having cumulative frequency $\frac{N}{2}$

v. Find median by using this formula;

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

Where,

l – Lower limit of the median class

m – Cumulative frequency of the class preceding the median class.

f – Frequency of the median class

c – Class interval of the median class

Illustration.2.1.14

The following table shows the household income of 80 families.

Income (in'000)	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of household	12	10	13	19	13	8	5

Find median income

Solution

The cumulative distribution table is

Class	f	Cf
0-10	12	12
10-20	10	22
20-30	13	35
30-40	19	54
40-50	13	67
50-60	8	75
60-70	5	80
	N = 80	

$$\frac{N}{2} = \frac{80}{2} = 40$$

The class having cumulative frequency 40 is 30-40

∴ Median class is 30-40

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

$$= 30 + \frac{40 - 35}{19} \times 10$$

$$= 30 + \frac{50}{19}$$

$$= 30 + 2.63$$

$$= 32.631$$

Illustration.2.1.15

Find the median wage of the following distribution

Wages (₹)	20-30	30-40	40-50	50-60	60-70
No. of labourers	3	5	20	10	5

Solution

Class	f	Cf
20-30	3	3
30-40	5	8
40-50	20	28
50-60	10	38
60-70	5	43
	N = 43	

$$\frac{N}{2} = \frac{43}{2} = 21.5$$

The class having cumulative frequency 21.5 is 40-50

∴ Median class is 40-50

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

$$= 40 + \frac{(21.5 - 8)}{20} \times 10$$

$$= 40 + \frac{135}{20}$$

$$= 40 + 6.75$$

$$= 46.75$$

Illustration.2.1.16

The table below shows the distribution of marks obtained in English by 265 students in the Science and Commerce streams.

Mark:	0-9	10-19	20-29	30-39	40-49	50-59	60-69
No. of students:	15	40	50	60	45	40	15

Find Median

Solution

Here the classes are of the inclusive type. Before computing the median, the inclu-

sive class should be converted into an exclusive class to get the actual class limit.

Marks	Actual class	f	cf
0-9	-0.5 - 9.5	15	15
10-19	9.5 – 19.5	40	55
20-29	19.5 – 29.5	50	105
30-39	29.5 - 39.5	60	165
40-49	39.5 – 49.5	45	210
50-59	49.5 – 59.5	40	250
60-69	59.5 – 69.5	15	265
		N = 265	

$$\frac{N}{2} = \frac{265}{2} = 132.5$$

The class having cumulative frequency 132.5 is 29.5 – 39.5

∴ Median class is 29.5 - 39.5

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

$$= 29.5 + \frac{132.5 - 105}{60} \times 10$$

$$= 29.5 + \frac{275}{60}$$

$$= 29.5 + 4.583$$

$$= 34.083$$

Illustration.2.1.17

Calculate the median mark from the following frequency table giving the distribution of marks of 80 students.

Mark	0-9	10-19	20-29	30-39	40-49
No. of students	3	10	28	36	3

Solution

Here the classes are of the inclusive type. Before computing the median, the inclusive class should be converted into an exclusive class to get the actual class limit.

Marks	Actual class	f	cf
0-9	-0.5 - 9.5	3	3
10-19	9.5 – 19.5	10	13
20-29	19.5 – 29.5	28	41
30-39	29.5 - 39.5	36	77
40-49	39.5 – 49.5	3	80
		N = 80	

$$\frac{N}{2} = \frac{80}{2} = 40$$

The class having cumulative frequency 40 is 19.5 – 29.5

∴ Median class is 19.5 - 29.5

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

$$= 19.5 + \frac{(40 - 13)}{28} \times 10$$

$$= 19.5 + \frac{270}{28}$$

$$= 19.5 + 9.64$$

$$= 29.14$$

Mode

Mode of a data is defined as the value that appears the most frequently in the data. It is the data observation with the highest frequency. Mode means norm or fashion. Mode is also known as the business average or fashionable average.

Advantages of mode

i. It is possible to calculate it graphically.

ii. It determines which value in a series is the most representative.

iii. It is unaffected by the series' extreme values.

iv. It has a lot of applications in the sphere of business and commerce.

v. It is simple and clear to compute and comprehend.

vi. It is not required to know the values of all the items in a series to calculate mode.

vii. The position of mode is likewise not a problem with open ended classes.

viii. The only average that works with categorical data is the mode.

Disadvantages of mode

i. Because it is not rigidly defined, it may have different results in some instances.

ii. Further algebraic treatment is not possible.

iii. It is not based on all evidence.

iv. It is ill-defined, indefinite, and ambiguous.

v. It can only be calculated from series with unequal class intervals if they are equalised.

vi. It is influenced by sampling fluctuations.

Computation of mode

i. **For individual series**

The mode in individual observations is the most recurring value in a series.

Illustration.2.1.18

Eleven people aged 18, 17, 19, 18, 17, 18, 21, 22, 18, 23, and 21 years old took part in a cricket match. Determine the mode of the data.

Solution

Here the observation 18 appears 4 times, 17 and 21 appears 2 times and all others are appeared in a single time. So, the value which appears a maximum number of times is 18.

∴ Mode = 18 years

ii. For Discrete series

Observation with highest frequency is considered as the mode in the discrete series.

Illustration.2.1.19

The age distributions in the following table illustrate the age of employees in various departments. Determine the mode.

Age	21	22	23	24	25	26
No. of employees	5	15	50	30	21	17

Solution

The age 23 has the highest frequency. Therefore, 23 is the mode.

Grouping Table and Analysis Table

The item with the highest frequency is referred to as a mode. However, if the maximum frequency is repeating or if the maximum frequency occurs at the beginning or end of the distribution or if there are irregularity in the distribution it may be impossible to find the mode simply by looking at it. In rare circumstances, the frequency concentration may be more concentrated around a frequency that is lower than the highest frequency. A grouping table and an analysis table should be developed to determine the correct modal value in such circumstances.

Steps for calculation

i. Construct a six-column grouping table.

ii. In column (1), record the frequency in relation to the item.

iii. The frequencies in column (2) are arranged in twos, starting at the top. Their totals are calculated, and the highest total is highlighted.

iv. The frequencies are grouped in twos again in column (3), leaving the first frequencies. The highest total is once again noted.

v. The frequencies in column (4) are arranged in threes, starting at the top. Their totals are calculated, and the highest total is highlighted.

vi. The frequencies are grouped in threes again in column (5), leaving the initial frequency. Their totals are calculated, and the highest total is highlighted.

vii. The frequencies are grouped in threes again in column (6), with the first and second frequencies leaving. After totaling the frequencies, the highest total is identified and highlighted again.

viii. Create an analysis table to find the modal value or modal class that the largest frequencies cluster around for the longest periods of time. Place the column number on the left-hand side of the table and the item sizes on the right-hand side. Mark 'X' in the relevant box corresponding to the values they represent to input the values against which the highest frequencies are found. The mode is the set of values with the most 'X' marks against them.

Illustration.2.1.20

The following table shows the monthly income of 148 families. Calculate the mode value.

Income (in'000):	10	15	20	25	30	35	40
No. of Families:	7	10	35	30	25	33	8

Solution

Since there are irregularity in the distribution, we must construct the grouping table and analysis table because determining the modal value is tough by examination.

(a) **Grouping table**

Income (In'000) x	f (1)	Grouping in twos		Grouping in threes		
		(2)	(3)	(4)	(5)	(6)
10	7	17		52		
15	10		45		75	
20	35	65				90
25	30		55	88		
30	25	58			66	
35	33		41			
40	8					

In column 1, the highest frequency is 35, which corresponds to the 20. So, we put X mark in 20. In column 2 the highest frequency is 65 corresponds to 35 and 30. So we put X mark in 20 and 25. In column 3 the highest frequency is 55 corresponds to 30 and 25. So we put X mark in 25 and 30. In column 4 the highest frequency is 88 corresponds to 30, 25 and 33. So we put X mark in 25, 30 and 35. In column 5 the highest frequency is 75 corresponds to 10, 35 and 30. So we put X mark in 15, 20 and 25. In column 6 the highest frequency is 90 corresponds to 35, 30 and 25. So we put X mark in 20, 25 and 30.

(b) Analysis table

Variable	10	15	20	25	30	35	40
F column							
1			X				
2			X	X			
3				X	X		
4				X	X	X	
5		X	X	X			
6			X	X	X		
Total	-	1	4	5	3	1	-

The greatest total (5) is noted to be against the value of 25. As a result, the modal mark is 25.

Mode = 25

iii. For continuous series

Steps

a. Locate the modal class having highest frequency or by preparing analysis table.

b. Apply this formula

$$\text{Mode} = l + \frac{(f_1 - f_0)}{2f_1 - f_0 - f_2} \times c$$

Where,

l – Lower limit of the modal class.

f_1 – Frequency of the modal class.

f_0 – Frequency of the preceding class to the modal class.

f_2 – Frequency of the succeeding class to the modal class.

c – Class interval of the modal class

Illustration.2.1.21

The following table shows the frequency distribution of the marks of 72 students. Find the mode.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of students	3	5	7	10	12	15	14	4	2

Solution

Here 50-60 is the model class. c = 10, f_1 =15, f_2=14, f_0 = 12

$$\text{Mode} = l + \frac{(f_1 - f_0)}{2f_1 - f_0 - f_2} \times c$$

$$= 50 + \frac{(15-12)}{2\times15-12-14} \times 10$$

$$= 50 + \frac{30}{4}$$

$$= 57.5$$

Illustration.2.1.22

The following table shows the monthly income of 125 families. Calculate the mode value.

Income (in'000)	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Families	4	8	18	30	20	10	30	3	2

Solution

Here the maximum frequency 30 is repeating, we use grouping table and an analysis table

a. Grouping Table

Income (In'000)	f (1)	Grouping in twos		Grouping in threes		
		(2)	(3)	(4)	(5)	(6)

0-10	4	12		30		
10-20	8		26		56	
20-30	18	48				68
30-40	30		50	60		
40-50	20	30			60	
50-60	10		40			43
60-70	30	33		35		
70-80	3		5			
80-90	2					

(b) Analysis table

Variable / F column	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
1				X			X	
2			X	X				
3				X	X			
4				X	X	X		
5					X	X	X	
6			X	X	X			
Total	-	-	2	5	4	2	2	-

The model class is identified as 30-40. The following formula can be used to calculate mode.

$$\text{Mode} = l + \frac{(f_1 - f_0)}{2f_1 - f_0 - f_2} \times c$$

$l = 30$

$f_1 = 30$

$f_0 = 18$

$f_2 = 20$

$c = 10$

$$\text{Mode} = 30 + \frac{(30-18)}{(2 \times 30)-18-20} \times 10$$

$$= 30 + \frac{120}{22}$$

$$= 30 + 5.45$$

$$= 35.45$$

Calculation of mode by Graphical Method

In graphical method, the series is represented by the rectangular diagram. The value of the mode can be determined graphically only in continuous series.

The following are the steps of calculating mode by graphical method.

1. According to the given frequency distribution, draw a histogram.
2. The highest rectangle in this diagram is modal-class.
3. Join the top right and left edges of this rectangle with the top right edge of the rectangle representing the preceding class and the top left edge of the rectangle representing the succeeding class respectively.
4. From the point of intersection of both the lines draw a straight line downwards perpendicular on the X-axis. The point where the straight line meets the X-axis gives us the modal-value.

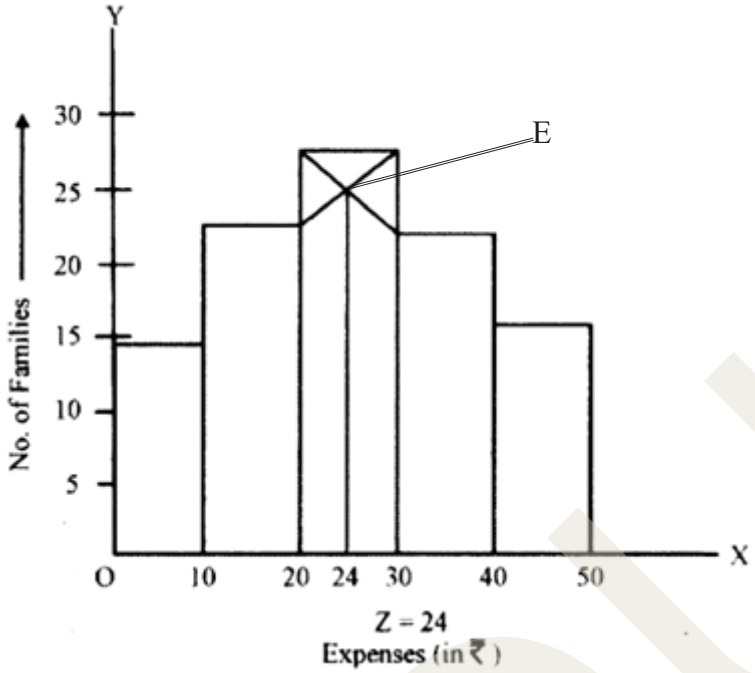
Illustration.2.1.23

Find out the mode of the following frequency distribution by the graphical method

Expense in ₹	0-10	10-20	20-30	30-40	40-50
No. of families	14	23	27	21	15

Solution

The figure shows the expenditure on X axis and the number of families on the Y axis. The rectangle of 20-30 class is highest. The lines joining both top vertices of the rectangle, with the opposite right and left top vertices of the preceding and succeeding rectangles, intersect one another at the point E. On putting the perpendicular from point E to the X axis, the X axis is intersected on the mode value of the distribution. Therefore Mode = 24.



Empirical Relation between median and mode

The relationship between mean, median, and mode depends upon the shape of the frequency curve, in other words, the relationship between mean, median, and mode depends upon the type of frequency distribution. For a symmetrical distribution, mean, median, and mode all coincide. That is, mean = median = mode. In the case of a negatively (left) skewed curve, mean<median<mode, whereas in the case of a positively skewed (right) curve,

mean > medan > mode

For a moderately asymmetrical frequency distribution, the empirical relationship between mean, median, and mode is given by Karl Pearson and is defined as

Mode = 3(Median) -2(Mean)

If any two values are known, the third value can be easily determined.

Applications of Measures of central tendencies in Application in Business decisions

Measures of central tendency, including the mean, median, and mode, play a crucial role in business decision-making processes. They provide valuable insights into the central or typical values of various business data, allowing organisations to make informed decisions. Here are some applications of measures of central tendency in business decisions:

1. Market analysis and consumer behavior:
    - o Mean and median values can be used to analyse consumer spending

patterns, income levels, and purchasing behavior within target markets.

- The mode can help identify the most frequently occurring preferences, choices, or behaviors among consumers.

2. Product pricing and sales strategies:
    - The mean and median prices of similar products in the market can guide pricing decisions for a company's offerings.
    - The mode can reveal the most popular price point or the price at which the highest number of sales occur.
3. Employee compensation and performance evaluation:
    - The mean and median salaries can help determine fair compensation levels for different job roles and experience levels within an organisation.
    - The mode can identify the most common or typical performance rating among employees, which can inform performance management strategies.
4. Inventory management and production planning:
    - The mean and median demand for products can help forecast inventory levels and production quantities.
    - The mode can identify the most frequently ordered products or the most popular product configurations.
5. Financial analysis and budgeting:
    - The mean and median values of financial metrics, such as revenue, expenses, and profits, can provide insights into a company's financial performance and assist in budget planning.
    - The mode can reveal the most frequent financial outcomes or patterns, which can aid in risk assessment and scenario planning.
6. Quality control and process improvement:
    - The mean and median values of product dimensions, defect rates, or process cycle times can help identify deviations from desired targets

and guide quality improvement efforts.

- The mode can highlight the most common types of defects or process issues, enabling targeted corrective actions.

7. Marketing and advertising campaigns:
   - The mean and median values of customer engagement metrics, such as impressions, clicks, or conversions, can evaluate the effectiveness of marketing campaigns.
   - The mode can identify the most successful or popular marketing channels, messages, or tactics for future campaigns.

By incorporating measures of central tendency into their data analysis and decision-making processes, businesses can gain valuable insights into typical or representative values, identify patterns and trends, and make more informed decisions across various functional areas, such as marketing, operations, finance, and human resources.

Recap

- ◊ Mean- is the sum of the observations divided by the number of observations.
- ◊ Median - middle number in a list of numbers sorted in ascending or descending order.
- ◊ Mode- most often occurring value in a series.

Objective Questions

1. Which measure of central tendency is used for calculating the average speed of a vehicle?
2. What is the most widely used measure for central tendency?
3. What is the name of an average calculated from a set of observations in which one number based on its position is chosen to represent the complete set?

4. Name the average that is commonly referred to as the "middle most value of observation."

5. Name the average, often known as the "most recurring value in a series."

Answers

1. Harmonic Mean
2. Arithmetic mean
3. Positional average
4. Median
5. Mode

Assignments

1. What is arithmetic average? State the advantages and limitations.
2. Explain the concept of combined arithmetic mean.
3. Write a short note on the concept of combined arithmetic mean.
4. What is geometric mean?
5. List 3 situations where harmonic mean can be used.
6. What is harmonic mean?
7. Explain the relationship between mean, median and mode.
8. What is a median?
9. State the merits and demerits of using mode.
10. What do you mean by a positional average? Explain any 2 positional averages.
11. The following information shows the number of traffic accidents that occurred in 520 cities. Calculate the mean.

No of accidents	2	4	6	8	10	12	14	16	18	20
No of cities	38	104	140	78	48	42	28	24	16	2

Ans: 7.78

12. The following is a list of 200 people's weekly earnings. Determine the mean.

Weekly wages	1000-2000	1200-1400	1400-1600	1600-1800	1800-2000	2000-2200	2200-2400	2400-2600
No of persons	3	21	35	57	40	24	14	6

Ans. 1768

13. Find Geometric mean

x:	18	16	22	12

Ans.16.61

14. Speed of four cycles are as follows

x:	3.4	5.2	4.8	6.1	4.1

Calculate Harmonic Mean.

Ans. 4.484

15. Calculate Geometric mean

Class	0-10	10-20	20-30	30-40	40-50
Frequency	8	15	25	6	16

Ans. 21.93

16. Find out the median of the following observation

x:	46	64	87	41	58	35	77	55	90	33	92

If 92 is replaced with 99 and 41 with 43 in the above data, find the new median.

Ans.58

17. The following table represents the income of 122 families. Calculate Median income.

Income	1000	1500	3000	2000	2500	1800
No. of family	24	26	16	20	6	30

Ans. 1800

Suggested Reading

1. Gupta, S.C. and Kapoor, V.K. *Fundamentals of Applied Statistics*. Sultan Chand and Sons, Delhi, 4th Edition, January 2007.
2. S.P.Guptha-(2012). *Statistical Methods*. Sultan Chand &Sons,New Delhi
3. Sancheti and V.K. Kapoor. (2014). *Business Mathematics. Sultan Chand & Sons*
4. Wikes, F.M (1998). *Mathematics for Business*, *Finance and Economics*. Thomson Learning

Reference

1. Gupta, S.C. and Kapoor, V.K. *Fundamentals of Mathematical Statistics*, Sultan Chand and Sons, Delhi, XI edition, 2002.
2. Dr. P.R. Vittal (2012). *Business Maths & Statistics*. Margham Publications
3. C.R.Kothari (2013). *Quantitative Techniques*. Vikas Publishing House
4. Dixon, W.S. and F.J. Massey. (1951). *Introduction to Statistical Analysis*, McGraw Hill Book Company, Inc. New York.
5. Fisher, R.A. and F. Yates. (1948). *Statistical Tables for Biological, Agricultural and Medical Research*, 3rd Ed. Haftner Publishing Company, New York.

Unit - 2

Measures of Dispersion

Learning Outcomes

After the completion of this unit, the learner will be able to;

- ◊ familiarise with various types of dispersion.
- ◊ computation of Range, Quartile deviation, Mean Deviation and standard deviation.
- ◊ rectify errors that occur during the computation of standard deviations.

Prerequisite

Dabur India was founded by SK.Burman and started operations initially as a small pharmacy in Kolkata in 1884. Dabur was incorporated as a private limited company in 1936 by the Dabur Group to produce cosmetics and toilet preparations. It became a public limited company in 1986 after a reverse merger with Vidogum limited. The company sells a host of personal care products including hair oil, soap, shampoo toothpowder, toothpaste, health supplements such as Chyawanprash and honey, digestives ayurveda-based over-the-counter medicines, and other consumer care products. These products are marketed under the brand Dabur, Vatika, Hajmola and Anmol and are positioned on the ayurvedic wellness platform.

Dabur responded to the changing dynamics of the business environment in India by appointing its first non-family CEO Ninu Khanna in 1998. The company promotes its brands in the various regional languages of India and is focused on creating special products with a distinct local touch. The company has been able to sustain its growth momentum in key categories like hair care, oral care and health supplements reported a 15% growth during the year 2007- 2008 whereas its foods business grew by 19%. The Consumer Health Division also marked a turnaround, reporting a 12% growth in the second half of the 2007-2008 fiscal and 5.4% growth to the year.

Dabur is also keen to maintain and expand its foothold in the international market. It has shown remarkable presence in the Middle East and North African countries and has doubled its business in Pakistan. Table 2.2.1 lists the net exports of Dabur India Lite for the past 12 years.

Table 2.2.1: Net Export of Dabur India Ltd from 1996 – 2007

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Net Exports (in Million Rupees)	329.4	447.6	164.9	151.5	286.4	245.4	286.4	331.4	203.6	309.5	211.8	583.2

The value of net exports varies greatly over the years with the lowest export at. Rs 151.5 million in 1999 and the highest export at Rs. 583.2 million in 2007. The measures of mathematical averages and positional averages alone will not describe this data adequately. We need to have some tools that will measure the scatteredness of data to analyse the variations. This unit focuses on the various measures of dispersion.

Keywords

Measures of Dispersion, Range, Quartile deviation, Mean deviation, Standard deviation

Discussion

The various measures of central tendency discussed in the above unit provide information about a particular point of the data set. They give us an idea about the central position of data. If we have two distributions with the same mean, it becomes difficult to ascertain whether the two distributions are identical or different, if these two distributions are different, then it is difficult to ascertain which parameter can be used to measure the difference of these two distributions.

2.2.1 Measures of Dispersion

The meaning of dispersion is “scatteredness”. Suppose we have three distributions

with the same mean; Curve A, obtained from the first distribution, is less spread or less scattered than curve B (obtained from the second distribution), and curve A is also less spread or less scattered as compared to curve C (obtained from the third distribution) (Figure 2.2.1). By measures of central tendency, data characteristics cannot be specifically described. If we study mean alone, one important characteristic(scatteredness) will be missed. Additionally, by studying only the mean we will not be able to measure the difference between these three distributions. The same is true with median and mode, which tells us only one aspect of the data in terms of middle position and value with maxim frequency, respectively. So, a tool is required to measure the scatteredness of the data. The extent the degree to which data tends to spread around an average is called dispersion or variation. In other words, the degree to which numerical data tends to spread around an average value is called variation or dispersion of data. Figure 2.2.1 exhibits the difference between three distributions having the same mean but different dispersion.

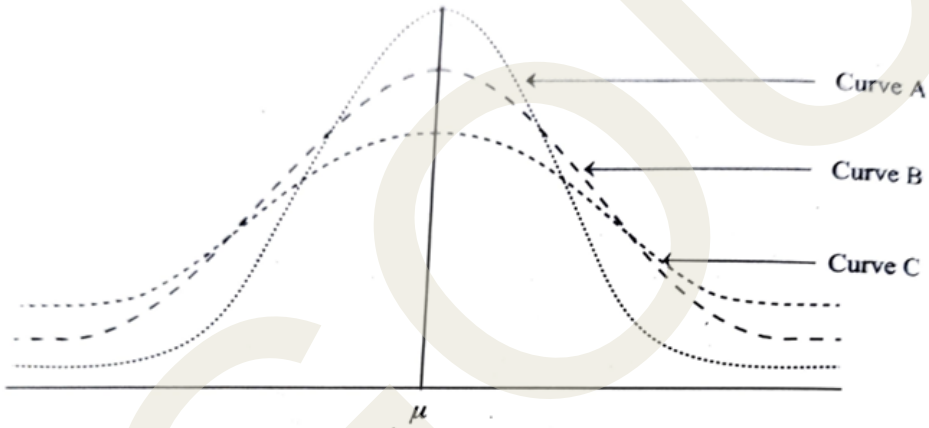


Figure 2.2.1: Three distributions A, B & C with same mean and different dispersion

The averages give an understanding of the distribution's central tendency, but it's also important to understand how the variables are clustered around or spread away from the average. The term "dispersion" refers to the degree to which variables deviate from its average. The degree of scatter or variations of the variable around a central value is referred to as dispersion.

According to A.L Bowley, "Dispersion is the measure of variation of the items."

According to Spiegel, "The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data."

Broadly, statistical techniques that measure dispersion are of two types. In the first category, we study statistical techniques to measure deviation of data value from a measure of central tendency which usually the mean or median. These statistical techniques are referred to as measures of dispersion (or variation or deviation).

Properties of a Good Measure of Dispersion

Like any other measure of central tendency, the measure of dispersion should possess some important Prerequisite. The important characteristics that a measure of dispersion should possess are the following:

1. It should be simple to understand and should be rigidly defined.
2. It should be based on all the observations.
3. Fluctuations of sampling should not affect it.
4. It should be suitable for further mathematical treatment.
5. It should not be affected by extreme values.

2.2.2 Absolute and Relative Measures of Dispersion

Absolute and relative measures of dispersion are two approaches used to quantify and compare the spread or variability of data within a dataset. Let's explore each of these measures in more detail:

a. Absolute Measures of Dispersion: Absolute measures of dispersion provide a direct measurement of the spread or variability in the original unit of measurement. These measures give a numerical value that represents the actual range or extent of variation in the data. Some commonly used absolute measures of dispersion include:

i. The Range
ii. The Quartile Deviation
iii. The Mean Deviation
iv. The Standard Deviation

b. Relative Measures of Dispersion: Relative measures of dispersion express the spread or variability of data relative to a reference point, typically a measure of central tendency. These measures provide a way to compare the variability across different datasets, even if they have different scales or units of measurement. Some commonly used relative measures of dispersion include:

i. Relative Range
ii. Relative Quartile Deviation
iii. Relative Mean Deviation
iv. Coefficient of variation

Relative measures of dispersion are particularly useful when comparing datasets with different scales or units, as it provides a standardised way to evaluate the spread relative to the central tendency.

Both absolute and relative measures of dispersion have their advantages and are chosen based on the specific context and requirements of the analysis. Absolute measures provide direct insights into the spread of the data, while relative measures facilitate comparisons and Standardisation across different datasets.

Difference between Absolute and Relative Measures of Dispersion

The difference between absolute and relative measures of dispersions are as follows:

Aspect	**Absolute Measures of Dispersion**	**Relative Measures of Dispersion**
Definition	Measures that quantify the spread of data in its original units.	Measures that express the spread of data relative to the mean or another reference point.
Calculation	Calculated using the actual values of the data.	Calculated using ratios or percentages.
Examples	Range, Mean Deviation, Variance, Standard Deviation	Coefficient of Variation, Relative Range, Coefficient of Mean Deviation
Interpretation	Provides information on the spread of data in the original units.	Helps to compare the dispersion of data across different datasets or variables.
Unit of Measurement	Retains the same unit as the data.	Unitless or expressed as a percentage.
Comparative Analysis	Suitable for comparing datasets with similar units.	Useful for comparing datasets with different units or scales.
Mathematical Properties	Can sum up the measures of dispersion.	Measures may not be directly additive or have mathematical properties like additivity.
Calculation Complexity	Often simpler to calculate compared to relative measures	May require additional calculations or transformations.

Methods of Measuring Dispersion

In the previous unit we discussed that only one measure of central tendency is not sufficient to study data. There are various tools for measuring central tendency, such as mean, median, mode, etc., each having some specialty. The same is true for the measures of dispersion. Only one measure of dispersion is not sufficient as its use is case specific. The following are some of the important and widely used methods of measuring dispersion:

- ◊ Range
- ◊ Quartile deviation
- ◊ Mean deviation
- ◊ Standard deviation

2.2.3 Range

Range is the simplest measure of dispersion. It is defined as the difference between the smallest and the greatest values in a distribution. In other words, range is the value of the highest observation - the value of the lowest observation. Symbolically,

R = L-S

Where, L is the largest observation, S the smallest observation and R is the range.

Range is an absolute measure of dispersion. The relative measure of dispersion for range is called the ***coefficient of range*** and is calculated by the following formula:

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

$$= \frac{\text{Largest Observation} - \text{Smallest Observation}}{\text{Largest Observation} + \text{Smallest Observation}}$$

Advantages of Range

a. It is the most basic method of determining dispersion.

b. It is simple to comprehend and calculate.

c. It is rigidly defined.

d. Even if some items in the midst of a series are missing, its calculation is unaffected.

Disadvantages of Range

a. It prioritizes only the two extreme values.

b. On many occasions, it is not a reliable measure of dispersion.

c. It is influenced by sampling fluctuations.

d. It cannot be further algebraically treated.

e. It does not take into account the frequencies of the distribution.

f. Range cannot be found for open ended distributions

Usage of Range

a. Range is useful in the following situations,

b. Range can be utilised as an efficient quality control method.

c. Range is used to describe the difference between a commodity's highest and lowest price. It is the most widely used measure of variability in our daily lives.

d. For weather forecasts, the meteorological department uses a range.

e. Range can be applied in areas where the data have small variations

Computation of Range

i. For individual series

Range = L - S

Where,

L = Largest value

S = Smallest value

Illustration.2.2.1

Below are the prices of 1 kg of apples for the first six months. Find Range and Coefficient of Range.

Month	January	February	March	April	May	June
Price/kg	120	115	150	130	175	160

Solution

Range = L-S

L =175

S = 115

Range = 175-115

= 60

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

$$= \frac{175-115}{175+115}$$

$$= \frac{60}{290}$$

$$= 0.206$$

ii. **For discrete series**

Illustration.2.2.2

From the following data relating to the monthly income of 60 people, determine the range and coefficient of range.

Income	210	240	290	360	440	510	500	350	290
No of persons	5	10	15	7	3	10	2	3	5

Solution

Range = L-S

= 510 – 210

= 300

Coefficient of Range = $\frac{510 - 210}{510+210}$

$$= \frac{300}{720}$$

= 0.416

iii. For continuous series

Illustration.2.2.3

From the following data calculate Range and Coefficient of range

Mark	10-20	20-30	30-40	40-50	50-60	60-70
No of Person	14	12	15	8	6	5

Solution

Range = L-S

= 70 – 10

$$= 60$$

$$\text{Coefficient of Range} = \frac{70 - 10}{70 + 10}$$

$$= \frac{60}{80}$$

$$= 0.75$$

2.2.4 Quartile Deviation

The interquartile range is a measure of dispersion based on the upper quartile Q_3 and lower quartile Q1. The quartile deviation is a measure of dispersion based on quartiles. It is the half of the difference between the upper and lower quartile. It is obtained by dividing interquartile range by 2. Therefore, it is also known as semi- inter quartile range.

Inter quartile range = $Q_3 - Q_1$

$$QD = \frac{Q_3 - Q_1}{2}$$

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Where,

QD – Quartile deviation

Q_1 – Lower quartile

Q_3 – Upper quartile

Quartile deviation is a better measure than range since it makes use of 50 per cent of the data compared to range which is based on only the highest and lowest values.

Advantages of Quartile Deviation

◊ It is easy to understand and calculate

◊ It can be calculated for open-ended classes.

◊ It is unaffected by extreme values.

Disadvantages of Quartile Deviation

◊ It is not based on all observation

◊ It is not capable of further algebraic treatment.

Computation of Quartile Deviation

i. For individual series

Illustration.2.2.4

Compute the inter-quartile range, quartile deviation, and coefficient of quartile deviation from the following data:

X	14	13	9	7	12	17	8	10	6	15	18	20	21

Solution

Arrange the data in ascending order

X	6	7	8	9	10	12	13	14	15	17	18	20	21

n =13

Q_1 = value of $\left(\frac{n+1}{4}\right)^{th}$ item

$= \left(\frac{13+1}{4}\right)^{th}$ item

$= 3.5^{th}$ item

$= 3^{rd}$ item + 0.5 (4^{th} item -3^{rd} item)

= 8 + 0.5 (9-8)

= 8.5

Q_3 = value of $\left(\frac{3(n+1)}{4}\right)^{th}$ item

$= 3 \times 3.5^{th}$ item

$= 10.5^{th}$ item

$= 10^{th}$ item + 0.5 (11^{th} item – 10^{th} item)

= 17 + 0.5 (18-17)

= 17.5

Inter-quartile range = $Q_3 - Q_1$

= 17.5 - 8.5

= 9

Quartile Deviation = $\frac{Q_3 - Q_1}{2}$

$= \frac{17.5-8.5}{2}$

$= \frac{9}{2}$

$= 4.5$

Coefficient of Quartile Deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

$= \frac{17.5-8.5}{17.5+8.5}$

$= \frac{9}{26}$

$= 0.346$

Illustration.2.2.5

The following are the prices of a drug in seven different medical stores.

Price:	350	300	425	450	400	600	500

Find first and third quartiles, quartile deviation, and coefficient of quartile deviation

Solution

Arrange the data in ascending order

300	350	400	425	450	500	600

$n = 7$

$Q_1 = \left(\frac{n+1}{4}\right)^{th}$ item in the series

$= \left(\frac{7+1}{4}\right)^{th}$ item in the series

$= 2^{nd}$ item

$= 350$

$Q_3 = \left(\frac{3(n+1)}{4}\right)^{th}$ item in the series

$= \left(\frac{3(7+1)}{4}\right)^{th}$ item in the series

$= 6^{th}$ item

$= 500$

Inter-quartile range $= Q_3 - Q_1$

$= 500 - 350$

$= 150$

Quartile Deviation $= \frac{Q_3 - Q_1}{2}$

$$= \frac{500 - 350}{2}$$

$$= \frac{150}{2}$$

$$= 75$$

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{500 - 350}{500 + 350}$$

$$= \frac{150}{850}$$

$$= 0.176$$

ii. For discrete series

Illustration.2.2.6

Below are the heights (in inches) of 49 people.

Height (in inches)	58	59	60	61	62	63	64	65	66
No of persons	2	3	6	15	10	5	4	3	1

Calculate inter-quartile range, quartile deviation, and coefficient of quartile deviation.

Solution

Height (in inches)	**Frequency**	**Cumulative Frequency**
58	2	2
59	3	5
60	6	11
61	15	26
62	10	36
63	5	41
64	4	45
65	3	48
66	1	49
	49	

N = 49

Q_1 = Series having cf $\left(\frac{49+1}{4}\right)$

= Series having cf 12.5

= 61

$Q_3 = \left(\frac{3(49+1)}{4}\right)$

= Series having cf 37.5

= 63

Inter-quartile range = $Q_3 - Q_1$

= 63 - 61

= 2

Quartile Deviation = $\frac{Q_3 - Q_1}{2}$

$= \frac{63 - 61}{2}$

$= \frac{2}{2}$

= 1

Coefficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

$= \frac{63 - 61}{63 + 61}$

$= \frac{2}{124}$

= 0. 016

i. For continuous series

Illustration.2.2.7

The salaries of 270 employees are given below. Calculate the inter-quartile range, quartile deviation, and coefficient of quartile deviation.

Solution

Salary (in'000)	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No of employees	13	17	50	60	55	45	23	7

Solution

X	f	c.f.
0-10	13	13
10-20	17	30
20-30	50	80
30-40	60	140
40-50	55	195
50-60	45	240
60-70	23	263
70-80	7	270
	N = 270	

Q_1 Class $= \left(\frac{n}{4}\right)^{th}$ Class

$= \left(\frac{270}{4}\right)^{th}$ Class

$= 67.5^{th}$ Class

$= 20 - 30$

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

$$= 20 + \frac{(67.5-30)}{50} \times 10$$

$$= 20 + \frac{375}{50}$$

$$= 20 + 7.5$$

$$= 27.5$$

Q_3Class $= 3\left(\frac{n}{4}\right)^{th}$ Class

$= 3 \times 67.5^{th}$ class

$= 202.5^{th}$ class

$= 50 - 60$

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

$$= 50 + \frac{(202.5 - 195)}{45} \times 10$$

$$= 50 + \frac{75}{45}$$

$$= 50 + 1.67$$

$$= 51.67$$

$Q_1 = 27.5$

$Q_3 = 51.67$

Inter quartile range $= Q_3 - Q_1$

$= 51.67 - 27.5$

$= 24.17$

Quartile Deviation $= \frac{Q_3 - Q_1}{2}$

$$= \frac{51.67 - 27.5}{2}$$

$$= \frac{24.17}{2}$$

$= 12.085$

Coefficient of Quartile Deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

$$= \frac{51.67 - 27.5}{51.67 + 27.5}$$

$$= \frac{24.17}{79.17}$$

$= 0.305$

Illustration 2.2.8

Calculate the quartile deviation for the following data of annual income of 100 families.

Annual Income	Less than 499	500-999	1000-1999	2000-2999	Above 3000
No of families	5	25	40	20	10

Ann. Income	Actual class	f	cf
Less than 499	Less than 499.5	5	5
500-999	499.5-999.5	25	30
1000-1999	999.5-1999.5	40	70
2000-2999	1999.5-2999.5	20	90
Above 3000	Above 2999.5-	10	100
		N = 100	

Q_1 Class = $\left(\frac{n}{4}\right)^{th}$ Class

= $(\frac{100}{4})^{th}$ Class

= 25^{th} Class

= 500-999 class

$$Q_1 = l_1 + \frac{\left(\frac{n}{4} - m_1\right)}{f_1} \times c_1$$

$$= 499.5 + \frac{(25-5)}{25} \times 1000$$

$$= 499.5 + 800$$

$$= 1299.5$$

Q_3 Class = $\left(\frac{3n}{4}\right)^{th}$ Class

= 75^{th} class

= 2000-2999 class

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

$$= 1999.5 + \frac{(75 - 70)}{20} \times 1000$$

$$= 1999.5 + 250$$

$$= 2249.5$$

Inter quartile range = $Q_3 - Q_1$

$= 2249.5 - 1299.5$

$= 950$

$$\text{Quartile Deviation} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{2249.5 - 1299.5}{2}$$

$$= \frac{950}{2}$$

$= 475$

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{2249.5 - 1299.5}{2249.5 + 1299.5}$$

$$= \frac{950}{3549}$$

$= 0.268$

2.2.5 Mean Deviation

The arithmetic mean of the absolute deviation of the observations from an assured average is called mean deviation. In other words, the arithmetic average of the deviations of items in a series taken from its central value, ignoring the plus and minus sign, is known as mean deviation. It is also known as Average deviation.

$$\text{Mean deviation} = \frac{\sum|x-A|}{n}$$

Where,

A = Any average. i.e., Mean or Median or Mode

|x-A| is read as modulus (x-A) is the modulus or absolute value of the deviation obtained after ignoring the negative sign.

Coefficient of Mean Deviation

The mean deviation is divided by the average to get the coefficient of mean deviation. If deviations are taken from mean, we divide it by mean, if the deviations are taken from median, then it is divided by median and if the deviations are taken from mode, then we divide mean deviation by mode. It is calculated to compare the data of two series.

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

Where,

MD - Mean Deviation

A - Any average. i.e., Mean or Median or Mode

Coefficient of Mean variation

$$\text{Coefficient of Mean variation} = \frac{MD}{A} \times 100$$

Advantages of mean deviation

a. It is based on all observation

b. It is rigidly defined

c. Extreme values have little impact on it.

d. It truly represents the average of deviations of the items of a series by removing the irregularities in the distribution.

e. It can be calculated from any value

f. In the realm of business and trade, it is quite useful.

g. It is concise to compute and comprehend

Disadvantages of mean deviation

a. It cannot be treated mathematically any further.

b. Ignoring signs of deviations may create artificiality

c. When the actual value of an average is a fraction, it is difficult to calculate.

d. It cannot be used for open end classes.

e. The volatility of sampling has a significant impact on it.

Computation of mean deviation, coefficient of mean deviation and coefficient of mean variation

i. For individual series

Steps,

For calculating Mean Deviation

a. Calculate the average (mean, median, or mode) required to calculate the mean deviation.

b. Take the deviation of an item from the average (mean, median, or mode). i.e., $|x - A|$

c. Calculate the total value of the deviation. i.e $\sum|x\text{-}A|$

d. Divide the total value of deviation obtained by the number of observations.

$$MD = \frac{\sum|x\text{-}A|}{n}$$

Illustration 2.2.9

From the following data, compute the mean deviation from the mean, the coefficient of mean deviation, and the coefficient of mean variation.

X	8	11	5	4	5	0	2	6	9	3	2

Solution

$$\text{Mean deviation about mean} = \frac{\Sigma|x-\text{Mean}|}{n}$$

$$\text{Mean} = \frac{\Sigma x}{n}$$

$$= \frac{55}{11}$$

$$= 5$$

X	**$\lvert x - \text{Mean}\rvert$**
8	3
11	6
5	0
4	1
5	0
0	5
2	3
6	1
9	4
3	2
2	3
$\sum x = 55$	**$\Sigma\lvert x - A\rvert = 28$**

$$\text{Mean deviation} = \frac{\Sigma|x-\text{Mean}|}{n}$$

$$= \frac{28}{11}$$

$$= 2.545$$

$$\text{Coefficient of Mean Deviation} = \frac{MD}{mean}$$

$$= \frac{2.545}{5}$$

$$= 0.509$$

$$\text{Coefficient of Mean Variation} = \frac{MD}{mean} \times 100$$

$$= \frac{2.545}{5} \times 100$$

$$= 50.90\ \%$$

Illustration.2.2.10

The prices of one kilogram of orange in various markets in rupees are listed below.

Price	120	130	140	110	160	150	190	180	170	200

Find

a. Mean deviation about median

b. Coefficient of Mean deviation

c. Coefficient of Mean variation.

Solution

$$\text{Mean deviation about median} = \frac{\sum|x\text{-Median}|}{n}$$

Arrange the data in an ascending order

X	110	120	130	140	150	160	170	180	190	200

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

$$= \left(\frac{10+1}{2}\right)^{th} \text{ item}$$

$$= 5.5^{th} \text{ item}$$

$$= \frac{(5^{th} \text{ item} + 6^{th} \text{ item})}{2}$$

$$= \frac{(150+160)}{2}$$

$$= 155$$

X	$\|x - \text{Median}\|$
110	45
120	35
130	25
140	15
150	5
160	5
170	15
180	25
190	35
200	45
	$\sum\|x - A\| = 250$

$$\text{Mean deviation} = \frac{\sum|x\text{-Median}|}{n}$$

$$\text{Mean deviation about median} = \frac{250}{10}$$

$$= 25$$

$$\text{Coefficient of Mean Deviation} = \frac{MD}{\text{median}}$$

$$= \frac{25}{155}$$

$$= 0.1613$$

$$\text{Coefficient of Mean variation} = \frac{MD}{\text{median}} \times 100$$

$$= \frac{25}{155} \times 100$$

$$= 16.13\ \%$$

Illustration.2.2.11

Eleven people aged 18, 17, 19, 18, 17, 18, 21, 22, 18, 23, and 21 years took part in a cricket match. Calculate the mean deviation about mode.

Solution

The observation 18 appears four times, the observations 17 and 21 appear twice, and the others appear once. As a result, the value that appears the most is 18.

∴ Mode = 18 years

X	$\|x - Mode\|$
18	0
17	1
19	1
18	0
17	1
18	0
21	3
22	4
18	0
23	5
21	3
	$\sum\|x - A\| = 18$

Mean deviation = $\frac{\sum|x\text{-Mode}|}{n}$

Mean deviation about mode = $\frac{18}{11}$

= 1.636

ii. For discrete series

Steps,

For calculating the Mean Deviation.

Compute the average (mean, median, or mode) required to calculate the mean deviation

a. Take the deviation of an item from the average (mean, median, or mode) i.e.

$|x - A|$

b. Multiply $|x - A|$ by their frequency 'f'. Thus we get $f|x - A|$ to each observation

c. Calculate the total value of the deviation multiplied by its frequency

i.e. $\sum f|x-A|$

d. Divide the total obtained by the number of observations.

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

Illustration.2.2.12

The weights of 20 students in a class are shown below.

Weight (in kg)	49	53	54	55	66	68	70	80
No of student	1	2	4	5	3	2	2	1

Calculate

a. Mean deviation about mean

b. Coefficient of Mean deviation

c. Coefficient of Mean variation.

Solution:

Weight (in kg)	**No of student (f)**	**f *x**	**|x − Mean|**	***f*** **|x − Mean|**
49	1	49	11	11
53	2	106	7	14
54	4	216	6	24
55	5	275	5	25
66	3	198	6	18
68	2	136	8	16
70	2	140	10	20
80	1	80	20	20
Total	**N= 20**	**1200**		**$\sum f\|x-A\| = 148$**

$$\bar{x} = \frac{\Sigma f x}{N}$$

$$= \frac{1200}{20}$$

$$= 60$$

Mean weight = 60 kg

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

$$MD = \frac{148}{20}$$

$$= 7.4$$

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

$$= \frac{7.4}{60}$$

$$= 0.123$$

$$\text{Coefficient of Mean variation} = \frac{MD}{A} \times 100$$

$$= \frac{7.4}{60} \times 100$$

$$= 12.33\ \%$$

Illustration.2.2.13

Find the Mean deviation from median, Coefficient of Mean deviation, Coefficient of Mean variation for the following data

x	10	11	12	13	14
f	3	12	18	12	3

Solution

X	f	Cum f	|x – Median|	f |x – Median|
10	3	3	2	6
11	12	15	1	12
12	18	33	0	0
13	12	45	1	12

14	3	48	2	6
Total	**n= 48**			$\sum f \lvert x - A \rvert = 36$

Median $= \left(\frac{n+1}{2}\right)^{th}$ term

$= \frac{49^{th}}{2}$ term

$= 24.5^{th}$ term

$= 12$

$MD = \frac{\sum f \lvert x - A \rvert}{N}$

$= \frac{36}{48}$

$= 0.75$

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

$= \frac{0.75}{12}$

$= 0.0625$

Coefficient of Mean variation $= \frac{MD}{A} \times 100$

$= \frac{0.75}{12} \times 100$

$= 6.25\ \%$

iii. For continuous series

Steps,

For calculating the Mean Deviation.

a. Calculate the required average (mean, median, or mode) to calculate mean deviation.

b. Calculate the item's deviation from the average (mean, median, or mode), i.e., |x-A|, where x is the midpoint.

c. Multiply |x-A| by the frequency 'f' of each observation to get $f|x - A|$.

d. Calculate the deviation's total value multiplied by its frequency

i.e $\sum f \lvert x - A \rvert$

Divide the total obtained from the number of observation

$$MD = \frac{\sum f\,|x-A|}{n}$$

Illustration.2.2.14

Calculate the mean deviation of the following data using the median. Find the coefficient of mean variation and the coefficient of mean deviation as well.

Price of mango per kg	100-120	120-140	140-160	160-180	180-200
Demand	4	6	10	8	5

Solution

$$\text{Mean deviation about median} = \frac{\sum f|x-\text{Median}|}{n}$$

Class	Mid Value (x)	f	cf	|x − median|	f|x − median|
100-120	110	4	4	43	172
120-140	130	6	10	23	138
140-160	150	10	20	3	30
160-180	170	8	28	17	136
180-200	190	5	33	37	185
Total		**33**			**661**

n = 33

Median= $\left(\frac{n}{2}\right)^{th}$ class

= $(\frac{33}{2})^{th}$ class

= 16.5

The class having cumulative frequency 16.5 is 140-160

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

$$= 140 + \frac{16.5 - 10}{10} \times 20$$

$$= 140 + \frac{130}{10}$$

$$=140 +13$$

$$=153$$

$$MD = \frac{661}{33}$$

$$= 20.03$$

$$\text{Coefficient of Mean Deviation} = \frac{20.03}{153}$$

$$= 0.130$$

$$\text{Coefficient of Mean variation} = \frac{MD}{A} \times 100$$

$$= \frac{20.03}{153} \times 100$$

$$= 13.09\ \%$$

Illustration.2.2.15

Compute the mean deviation from mode of the following data. Also find the coefficient of mean variation and the coefficient of mean deviation as well.

Profit per shop (Rs)	0-10	10-20	20-30	30-40	40-50	50-60
No. of shop	12	18	27	20	17	6

Solution:

Class	**F**	**Mid Value (x)**	**|x – mode|**	**f|x – mode|**
0-10	12	5	20.63	247.56
10-20	18	15	10.63	191.34
20-30	27	25	0.63	17.01
30-40	20	35	9.37	187.4
40-50	17	45	19.37	329.29
50-60	6	55	29.37	176.22
Total		**100**		**1148.82**

$$\text{Mode} = l + \frac{(f_1 - f_0)}{2f_1 - f_0 - f_2} \times c$$

$$\text{Mode} = 20 + \frac{(27 - 18)}{2 \times 27 - 18 - 20} \times 10$$

$$= 20 + \frac{9}{16} \times 10$$

$$= 20 + 5.63$$

$$= 25.63$$

$$\text{MD} = \frac{1148.22}{100}$$

$$= 11.48$$

$$\text{Coefficient of Mean Deviation} = \frac{11.48}{25.63} = 0.45$$

Coefficient of Mean variation = $0.45 \times 100 = 45\ \%$

2.2.6 Standard deviation

Standard deviation is the positive square root of the mean of the squares of deviation from the arithmetic mean. It is denoted by the Greek letter σ (sigma). It cannot be negative. Karl Pearson was the first to introduce the concept of standard deviation. It is the most used methods of dispersion since it is free from some defects of other measures of dispersion.

Advantages of standard deviation

a. It is rigidly defined.

b. It is based on all observation.

c. Never disregards the plus or minus sign

d. It can be subjected to more mathematical analysis.

e. The changes in sampling have little effect on it.

f. It allows us to compare and contrast two or more series and determine their consistency or stability.

g. It is used in testing of hypothesis

Disadvantages of standard deviation

a. A layman would find it difficult to comprehend.

b. It is complex to calculate since it incorporates several mathematical models.

c. It cannot be used to compare the dispersion of two or more series of observations with different units of measurement.

Coefficient of variation

The coefficient of variation is calculated by dividing the standard deviation by the arithmetic mean, which is given as a percentage. It is the most popular way of comparing the consistency or stability of two or more sets of data.

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

Computation of Standard deviation and Coefficient of variation

i. For individual series

Standard deviation = $\sqrt{\frac{\sum d^2}{N}}$

Where,

d - Deviation of the item from their actual mean $(x - \bar{x})$

N - Total number of items

d^2 – Squares of deviation taken from actual mean

Illustration.2.2.16

A study looked into how many hours students spent studying before an exam. A total of eleven students were chosen. The students' responses are 8, 6, 3, 0, 5, 9, 2, 1, 3, 5, 2. Calculate the standard deviation and coefficient of variation for the number of hours that students have spent studying.

Solution

$$\bar{x} = \frac{\sum x}{n} = \frac{44}{11} = 4$$

X	d (x -4)	d^2
8	4	16
6	2	4
3	-1	1
0	-4	16
5	1	1
9	5	25

2	-2	4
1	-3	9
3	-1	1
5	1	1
2	-2	4
$\sum x = 44$		**$\sum d^2 = 82$**

$$\text{Standard deviation} = \sqrt{\frac{\sum d^2}{N}}$$

$$= \sqrt{\frac{82}{11}}$$

$$= \sqrt{7.4545}$$

$$= 2.73$$

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$= \frac{2.73}{4} \times 100$$

$$= 68.26\ \%$$

ii. For discrete series

$$\textbf{Standard deviation} = \sqrt{\frac{\sum f \times x^2}{N} - \bar{x}^2}$$

Where,

N - Total frequency

f – Frequency

Illustration.2.2.17

An arithmetic test was given to 100 kids. The following is the time in minutes required to finish the test:

Time (in minute)	18	19	20	21	22	23	24	25	26	27
No of students	3	7	11	14	18	17	13	8	5	4

Calculate the standard deviation of their test completion time as well as the coefficient of variation.

x	f	f x	x^2	$f x^2$
18	3	54	324	972
19	7	133	361	2527
20	11	220	400	4400
21	14	294	441	6174
22	18	396	484	8712
23	17	391	529	8993
24	13	312	576	7488
25	8	200	625	5000
26	5	130	676	3380
27	4	108	729	2916
	N = 100	**2238**		**50562**

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

$$= \frac{2238}{100}$$

$$= 22.38$$

$$\text{Standard deviation} = \sqrt{\frac{\sum fx^2}{N} - \bar{x}^2}$$

$$= \sqrt{\frac{50562}{100} - 22.38^2}$$

$$= \sqrt{505.62 - 500.8644}$$

$$= \sqrt{4.7556}$$

$$= 2.181$$

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$= \frac{2.181}{22.38} \times 100$$

= 9.75%

iii. For continuous series

Standard deviation = $\sqrt{\frac{\sum fx^2}{N} - \bar{x}^2}$

Where,

N - Total frequency

f – Frequency

x – Mid value

Illustration.2.2.18

Below are the profits earned by 100 sole proprietorship businesses.

Profit in '000	0-10	10-20	20-30	30-40	40-50	50-60
No of companies	8	12	20	30	20	10

Calculate the standard deviation and the coefficient of variation of the data.

Solution

Profit	**Mid value (x)**	**f**	**x^2**	**fx**	**fx^2**
0-10	5	8	25	40	200
10-20	15	12	225	180	2700
20-30	25	20	625	500	12500
30-40	35	30	1225	1050	36750
40-50	45	20	2025	900	40500
50-60	55	10	3025	550	30250
				$\sum fx = 3220$	**$\sum fx^2 = 122900$**

N = 100

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

$$= \frac{3220}{100}$$

$$= 32.2$$

$$\text{Standard deviation} = \sqrt{\frac{\sum fx^2}{N} - \bar{x}^2}$$

$$= \sqrt{\frac{122900}{100} - 32.2^2}$$

$$= \sqrt{192.16}$$

$$= 13.86$$

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$= \frac{13.86}{32.2} \times 100$$

$$= 43.04\ \%$$

Illustration.2.2.19

The results of five distinct class tests for two students, Rahul and Manu, are shown here.

Rahul	20	22	17	23	28
Manu	10	20	18	12	15

Determine which of the two students, Rahul or Manu, is the most consistent in terms of scoring.

Solution

$$\text{Rahul's } \bar{x} = \frac{\sum fx}{N}$$

$$= \frac{110}{5}$$

$$= 22$$

$$\text{Manu's } \bar{x} = \frac{\sum fx}{N}$$

$$= \frac{75}{5}$$

$$= 15$$

Rahul			Manu		
X	**d (x -22)**	**d^2**	**X**	**d (x -15)**	**d^2**
20	-2	4	10	-5	25
22	0	0	20	5	25
17	-5	25	18	3	9
23	1	1	12	-3	9
28	6	36	15	0	0
$\sum x = 110$		**$\sum d^2 = 66$**	**$\sum x = 75$**		**$\sum d^2 = 68$**

Rahul

$$\text{Standard deviation} = \sqrt{\frac{\sum d^2}{N}}$$

$$= \sqrt{\frac{66}{5}}$$

$$= \sqrt{13.2}$$

$$= 3.63$$

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$= \frac{3.63}{22} \times 100$$

$$= 16.5\ \%$$

Manu

$$\text{Standard deviation} = \sqrt{\frac{\sum d^2}{N}}$$

$$= \sqrt{\frac{68}{5}}$$

$$= \sqrt{13.6}$$

$$= 3.69$$

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$= \frac{3.69}{15} \times 100$$

$$= 24.6\ \%$$

In comparison to Manu, Rahul is more consistent in his scoring because his coefficient of variation is lower.

Combined standard deviation

The following formula can be used to calculate the combined standard deviation of two or more groups:

$$\sigma_{1.2} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

Where,

$\sigma_{1.2}$ – Combined standard deviation

σ_1 – Standard deviation of the first series

σ_2 – Standard deviation of the second series

$D_1 - (\bar{x}_1 - \bar{x}_{1.2})$

$D_2 - (\bar{x}_2 - \bar{x}_{1.2})$

$\bar{x}_{1.2}$ – Combined mean

$\bar{x}_{1.2} - \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1+N_2}$

N_1 – Number of items of the first series

N_2 – Number of items of the second series

Illustration.2.2.20

Calculate the combined standard deviation of the two Factories using the given information.

	Factory A	**Factory B**
Mean	63	54
SD	8	7
Number of items	50	40

Solution

$$\sigma_{1.2} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

$$\bar{x}_{1.2} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1+N_2}$$

$$= \frac{(50 \times 63)+(40 \times 54)}{50+40}$$

$$= \frac{3150+2160}{90}$$

$$= \frac{5310}{90}$$

$$= 59$$

$$d_1 = (\bar{x}_1 - \bar{x}_{1.2})$$

$$= (63 - 59)$$

$$= 4$$

$$d_2 = (\bar{x}_2 - \bar{x}_{1.2})$$

$$= (54 - 59)$$

$$= -5$$

$$\sigma_{1.2} = \sqrt{\frac{(50 \times 8^2) + (40 \times 7^2) + (50 \times 4^2) + (40 \times -5^2)}{50 + 40}}$$

$$= \sqrt{\frac{(50 \times 64) + (40 \times 49) + (50 \times 16) + (40 \times 25)}{90}}$$

$$= \sqrt{\frac{3200 + 1960 + 800 + 1000}{90}}$$

$$= \sqrt{\frac{6960}{90}}$$

$$= \sqrt{77.33}$$

$$= 8.79$$

Illustration.2.2.21

Analysis of the monthly wages of two hospitals gave the following information.

	Hospital I	**Hospital II**
No. of staff	550	600
Average wages	60	48.5
Variance	100	144

Obtain the average wage and combined standard deviation of the two hospitals together.

Solution

$$\sigma_{1.2} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

$$\bar{x}_{1.2} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2}$$

$$= \frac{(550 \times 60) + (600 \times 48.5)}{550 + 600}$$

$$= \frac{33000 + 29100}{1150}$$

$$= \frac{62100}{1150}$$

$$= 54$$

$$d_1 = (\bar{x}_1 - \bar{x}_{1.2})$$

$$= (60 - 54)$$

$$= 6$$

$$d_2 = (\bar{x}_2 - \bar{x}_{1.2})$$

$$= (48.5 - 54)$$

$$= -5.5$$

$$\sigma_{1.2} = \sqrt{\frac{(550 \times 100) + (600 \times 144) + (550 \times 6^2) + (600 \times (-5.5)^2)}{550 + 600}}$$

$$= \sqrt{\frac{55000 + 86400 + 19800 + 18150}{1150}}$$

$$= \sqrt{\frac{179350}{1150}}$$

$$= \sqrt{155.96}$$

$$= 12.49$$

Correction in mean and standard deviation

Sometimes, there can be errors in the data we use to calculate the average (mean) and how spread out the data is (standard deviation). These errors might only become clear after we've done the calculations. So, to get accurate results, we need to adjust the mean and standard deviation by considering the correct values for those observations.

Illustration.2.2.22

The mean and standard deviation of 11 observations were calculated as 5 and 3.67, respectively. But later, it was identified that one item having a value of 2 was misread as 13. Calculate the correct mean and standard deviation.

Solution

Incorrect $\sum x = \bar{x} \times n$

$= 5 \times 11 = 55$

Correct $\sum x$ = Incorrect $\sum x$ – wrong item + correct item

$= 55 - 13 + 2$

$= 44$

Correct $\bar{x} = \frac{44}{11}$

$= 4$

Calculation of the correct Standard Deviation

$$= \sqrt{\frac{\sum x^2}{N} - (\bar{x})^2}$$

$$3.67 = \sqrt{\frac{\sum x^2}{11} - (5)^2}$$

Squaring both sides

$$3.67^2 = \frac{\sum x^2}{11} - 25$$

$$13.4689 + 25 = \frac{\sum x^2}{11}$$

$$38.4689 = \frac{\sum x^2}{11}$$

$$\sum x^2 = 38.4689 \times 11$$

Incorrect $\sum x^2 = 423.1579$

Correct $\sum x^2$ = Incorrect $\sum x^2$ - square of wrong item + square of correct item.

$= 423.1579 - 13^2 + 2^2$

$= 423.1579 - 169 + 4$

$= 258.1579$

$$\text{Correct SD} = \sqrt{\frac{258.1579}{11} - 4^2}$$

$$= \sqrt{\frac{258.1579}{11} - 16}$$

$$= \sqrt{23.4689 - 16}$$

$$= \sqrt{7.4686}$$

$$= 2.73$$

Illustration.2.2.23

For a group of 200 candidates the mean and standard deviation of scores were found to be 40 and 15 respectively. Later on it was discovered that the score 43 and 35 were wrongly written as 34 and 53 respectively. Find the corrected mean and standard deviation corresponding to the corrected figure.

Solution

Incorrect $\sum x = \bar{x} \times n$

$= 40 \times 200$

$= 8000$

Correct $\sum x$ = Incorrect $\sum x$ – wrong item + correct item

$= 8000 – (34 + 53) + (43 + 35)$

$= 8000 – 87 + 78$

$= 7991$

Correct $\bar{x} = \frac{7991}{200}$

$= 39.955$

Calculation of the correct Standard Deviation

$$= \sqrt{\frac{\sum x^2}{N} - (\bar{x})^2}$$

$$15 = \sqrt{\frac{\sum x^2}{200} - (40)^2}$$

Squaring both sides

$$15^2 = \frac{\sum x^2}{200} - 1600$$

$$1600 + 225 = \frac{\sum x^2}{200}$$

$$1825 = \frac{\sum x^2}{200}$$

$\sum x^2 = 1825 \times 200$

Incorrect $\sum x^2 = 365000$

Correct $\sum x^2$ = Incorrect $\sum x^2$ - square of wrong item + square of correct item.

$= 365000 - (34^2 + 53^2) + (43^2 + 35^2)$

$= 365000 - 3965 + 3074$

$= 364109$

Correct SD $= \sqrt{\frac{364109}{200} - 39.955^2}$

$= \sqrt{224.143}$

$= 14.971$

Recap

- ◊ Dispersion – measures to quantify data deviation from averages.
- ◊ Common measures of dispersion include range, mean deviation, quartile deviation, and standard deviation.
- ◊ Absolute measures of dispersions -quantify original unit spread.
- ◊ Relative measures of dispersion - allow comparisons regardless of units.
- ◊ The range - difference between the largest and the least numbers in the set.
- ◊ Quartile Deviation - half of the difference between the upper and lower quartile.
- ◊ Inter-quartile range - The difference between upper quartile and lower
- ◊ The mean deviation - statistical measure used to calculate the average deviation from the average value of the series.
- ◊ The standard deviation - statistic that calculates the square root of the variance and measures the dispersion of a dataset relative to its mean.
- ◊ The standard deviation - the value of standard deviation cannot be negative.
- ◊ The coefficient of variation (CV) - ratio of the standard deviation to the mean.

Objective Questions

1. Which measure of dispersion provides information about the spread of data in its original units?
2. Which measure of dispersion is less affected by outliers?
3. What does dispersion measure in a dataset?
4. Which measure of dispersion provides a direct understanding of variability?
5. Which measure of dispersion allows comparisons regardless of units or scales?

Answers

1. Absolute measures of dispersion
2. Relative measures of dispersion
3. Spread or variability
4. Absolute measures of dispersion.
5. Relative measures of dispersion

Assignments

1. Define measures of dispersion and explain their importance in statistical analysis.
2. Discuss the concept of range as a measure of dispersion. Highlight its strengths and limitations.
3. Marks obtained by ten students in a class test is given below.

Marks	20	25	40	30	35	45	70	65	60	80

Find Mean Deviation from mean and coefficient of Mean Deviation.

Ans. M.D = 17.4, Coefficient of M.D = 0.37

4. The following table shows the number of books read by students in a B.com class consisting of 28 students, in a month.

No of Books	0	1	2	3	4
No of students	2	6	12	5	3

Calculate mean deviation about mode of number of books read.

Ans. 0.75

5. Calculate mean deviation about mean of the number of telephone calls received at an exchange:

No of calls	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	4	6	10	20	10	6	4

Ans. 11.33

6. Find standard deviation and coefficient of variation from the following data.

X	10	20	30	40	50	60	70	80	90	100

Ans. S.D= 28.72, CV = 52.21%

7. The score of two batters Saju and Raju in 10 innings during a certain season are as follows

Saju	25	50	45	30	70	42	36	48	34	60
Raju	10	70	50	20	95	55	42	60	48	80

Find which of the two batters, Saju or Raju is more consistent in scoring? Also state who is better run getter?

Ans. Saju - $\bar{x}$ = 44, SD = 13.08, CV = 29.37% Raju - $\bar{x}$ = 53, SD =24.35, CV = 45.94. Saju is more consistent in Scoring as his CV is low. Raju is a better run getter as his arithmetic mean is higher than that of Saju.

8. A factory produces two types of batteries X and Y. in an experiment relating to their life the following result were obtained.

Length of life (Hrs)	Number of battery X	Number of Battery Y
500-700	5	4
700-900	11	30
900-1100	26	12
1100-1300	10	8
1300-1500	8	6

a) Which battery X or Y has more average life?

b) Which is more consistent?

Ans. (a) Battery X has more average life.

(b) Battery X is more consistent.

9. A student is obtained the mean and Standard Deviation of 100 observation as 40 and 5.1 respectively. It was later discovered that the value of item 40 was misread as 50. Calculate the correct mean and standard deviation.

Ans. Mean = 39.9, S.D = 4.3

Suggested Reading

1. C. B Gupta & Vijay Gupta (2004). *An Introduction to Statistical Methods*. Vikas Publishing
2. Neiswanger, William A. *Elementary Statistical Methods*, The Macmillian Co., New York.
3. Richard I. Levin and David S. Rubin. *Statistics for Management*. Seventh Edition. Pearson - Prentice Hall of India, New Delhi.
4. Riggleman and Frisbee. *Business Statistics*, McGraw Hill Book Co., New York.

Reference

1. Dr. S M Shukla and Dr. Sahai (2020) - *Principles of statistics*, Sahitya Bhavan Publication, Delhi
2. C.R.Kothari (2013). *Quantitative Techniques*. Vikas Publishing House.
3. S.P.Guptha-(2012). *Statistical Methods*. Sultan Chand &Sons,New Delhi

BLOCK - 03
Correlation and Regression Analysis

Unit - 1
Correlation Analysis

Learning Outcomes

At the conclusion of this unit, the learner will be able;

- ◊ to understand the meaning of correlation,
- ◊ to be aware of the significance of correlation,
- ◊ to be familiar with different types of correlation.

Prerequisite

Most of you are familiar with the term variables. A 'variable' means values which vary. In previous units "measures of central tendency" and "measures of dispersion" the study is related to the characteristics of only one variable such as height, weight, ages, marks, wages etc. In real life, there may be situations where you need to understand and analyse relationship between more than one variable. For instance, you may need to check whether your body weight and the blood sugar level have any relationship. In such cases we may have to resort to analysing two variables weight and sugar level together to arrive at an interpretation about their relationship. In this unit, we are discussing the concept of correlation and the possibility of using correlation analysis.

Keywords

Positive Correlation, Negative Correlation, Linear and Non-linear Correlation, Univariate analysis, Bi-variate analysis

Discussion

As mentioned earlier, we often analyse variables individually in many situations. For example, analysis of marks obtained by a group of students in an examination, wages of employees in a factory. These type of analysis of data or statistical measures are known as univariate analysis.

Univariate analysis: The study related to the characteristics of only one variable is known as Univariate analysis. Measures of Central Tendency and Measures of Dispersion are the statistical tools comes under this category.

There are many situations where we might need to evaluate a variable in relation to another variable. For instance, parents advise their children to work hard so that they may get good marks. They are relating good marks with hard work. i.e., they know very well that good marks depend on hard work. Consider another example of a child's age and height. Even though knowledge of a child's age does not make it possible to estimate his height but it does help in estimating or forecasting the height with less error. This is done through the analysis of the relationship between the variables age and height. Such analysis is referred to bi-variate analysis.

Bi-variate analysis: The statistical analysis related to the study of the relationship between two variables is known as bi-variate analysis. For example, study of relationship between income and expenditure of a group of families include bi-variate analysis.

3.1.1 Meaning of Correlation

Let us examine a situation where a company invests money in advertising the products. In order to understand whether the advertisement is effective or not, the company needs to find out if the advertisement expense and the sales levels are related or not. If the company's sales increases when the investment in advertisement is increased, then we can assume that advertisement is effective. When you plot the figures corresponding to the two variables, it may be little clearer (refer figure 3.1.1). The figure indicates that there is some connection between the variables advertising and sales. This can be established statistically using correlation analysis.

Correlation is a measure of association between two variables. It can be understood as the movement in one variable in relation with the movement in another variable. The movement can be direct or inverse or none. You can find positive, negative and no correlation in the universe.

Correlation Definition:

When two variables are correlated that a change in one is accompanied by a change in the other in such a way that an increase in one variable follows increase/decrease in the other variable or decrease in one variable follows decrease/increase in the other variable. Then the variables are said to be correlated.

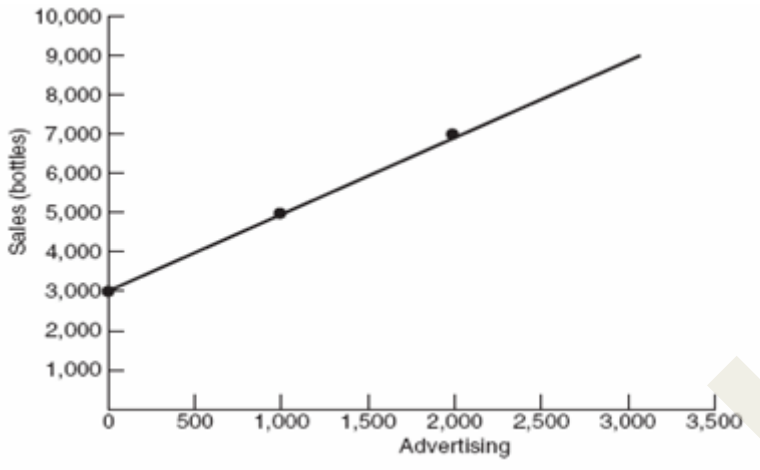


Figure 3.1.1 Relationship between sales and advertisement

3.1.2 Types of Correlation

In the earlier example, as we discussed about the relationship between advertisement and sales both variables are moving together in one direction. When advertisement increases, sales also increase. Will that be the case in every relationship? The answer is 'No.' Correlation between different variables can be of different nature. Hence, correlation can be of many types. Generally, correlation is classified into following types.

I. Positive and Negative

II. Linear and Non-linear

III. Partial and Total

IV. Simple and Multiple

3.1.2.1 Positive and Negative Correlation

Consider the two variables income and expenditure. We know increasing in the value of one variable is accompanied by an increase in the value of the other or decrease in the value of one variable is accompanied by the decrease in the other variable. i.e., if the two variables tend to move together in the same direction, then the correlation is called positive or direct correlation.

Example 1. The more time you spend running on a tread mill, the more calories you will burn.

Example 2. Quantity of money in circulation and price of commodity

Example 3: Taller people have larger shoe sizes and shorter people have smaller shoe sizes

If increasing or decreasing in the value of one variable is accompanied by a decreasing or increasing in the value of another variable i.e., if the two variables tend to move together in opposite direction, then the correlation is called negative or inverse correlation.

Negative correlation means that there is an inverse relationship between two variables-when one variable decreases, the other variable increases or one variable increase and the other decreases.

Example 1: If a train increases speed, the length of time to get to the final point decreases.

Example 2: If a car decreases speed, travel time to a destination increases.

Example 3: when experience increases typing error decreases.

3.1.2.2 Linear and Non-linear correlation

In order to understand the concept of linear and non-linear correlation, consider the values of two variables X and Y as below;

X	2	4	6	8	10	12	14
Y	4	8	12	16	20	24	28

If you observe the two arrays of data above, when variable X increased from 2 to 4, variable Y increased from 4 to 8. Again, when X increased from 4 to 6, Y increased from 8 to 12. In both cases the rate of increase is constant. Hence, we can deduce that the ratio of change between two variables is the same. If we plot these points on a graph, we will get a straight line. This type of relationship is called Linear correlation.

If the ratio of change between the two variables is a constant, then the correlation between the variables is called Linear correlation.

X	1	3	6	10	11	12	14
Y	3	6	10	15	20	26	29

Now think about another situation as depicted above where along with increase in one variable the other variable also increases but not in a constant rate. Such relationship is opposite to the Linear correlation. If the amount of change in one variable does not bear a constant ratio of the amount of change in the other, then the relationship is called Curvilinear correlation (Non-linear correlation) and if such variables are plotted, the resultant graph will show a curve nature.

3.1.2.3 Simple and Multiple Correlation

Imagine you are a student and you have noticed that the more time you spend studying for a test, the higher your test scores tend to be. This relationship between your study time and your test scores is a simple correlation. In other words, when one thing (study time) goes up, the other thing (test scores) also tends to go up. This is a basic way to measure how two variables are related to each other. Thus, simple correlation means relationship between two variables.

Eg: Price and demand

Now, let us say you realise that your test scores are not just affected by your study time alone. You also realise that when you get enough sleep the night before the test, your scores tend to be even better, regardless of how much time you spent studying. So now there are two factors at play: study time and sleep.

Multiple correlation takes into account multiple factors that might influence a result. In this case, it is not just about how much you study; it is also about how much you sleep. It is like understanding that your test scores are influenced by more than one thing, and you want to figure out how each of those factors contributes to the final result.

To put it simply, simple correlation looks at how one thing changes when another thing changes, while multiple correlation looks at how multiple things together influence something. The relationship of more than two variables is called multiple correlation. Eg: Yield of crops, rainfall, fertilizer used etc.

3.1.2.4 Partial and Total correlation

Imagine you are tracking the relationship between the amount of exercise you do and your weight. You notice that as you exercise more, your weight tends to go down. This is a total correlation because you are looking at the direct relationship between two variables; exercise and weight without considering any other factors.

Now let us make it a bit more interesting. You start noticing that some of your friends who exercise a lot have higher weights than you would expect based on the total correlation you observed earlier. You realise that diet might also be playing a role here. Some of your friends exercise a lot but also eat a lot, which could explain why their weight is not dropping as much as you would expect just based on exercise.

Partial correlation considers a third variable that might be influencing the relationship between the two main variables. In this case, the third variable is diet. By considering the partial correlation between exercise and weight while holding diet constant, you are trying to see the 'pure' relationship between exercise and weight without the confounding effect of diet.

In simple words, if we study only two variables and eliminates some other factors or variable is called partial correlation. Eg: Study price and demand and eliminating supply side. If we consider all facts related to the study, then correlation is called total correlation.

Recap

- ◊ Correlation - measure of association between two variables.
- ◊ Positive correlation- two variables move together in the same direction.
- ◊ Negative correlation - inverse relationship between two variables.
- ◊ Linear Correlation - ratio of change between the two variables is a constant.
- ◊ Non-Linear correlation - relation between two variables doesn't follow a straight line.
- ◊ Simple correlation - how one variable changes with another variable.
- ◊ Multiple correlation - relationship of more than two variables.
- ◊ Partial Correlation - study only two variables and eliminates another variable which may affect the relationship.
- ◊ Total correlation - study all variables related to another without omitting any.

Objective Questions

1. What term describes a relationship between two variables that does not follow a straight line?
2. What type of correlation shows a consistent increase in both variables simultaneously?
3. Which type of correlation demonstrates a consistent decrease in both variables together?
4. Which type of correlation indicates that a change in one variable corresponds to a proportional change in the other variable?
5. What term is used to describe a relationship between two variables that does not follow any specific pattern, appearing random or independent of each other?
6. Which correlation measures the strength and direction of the relationship between two variables while controlling for the influence of a third variable?

7. Which correlation represents the complete association or relationship between two variables without considering the impact of any other related variables?

Answers

1. Non-linear correlation.
2. Positive correlation.
3. Negative correlation.
4. Linear correlation.
5. No correlation.
6. Partial Correlation
7. Total Correlation

Self- Assessment Questions

1. What is correlation analysis?
2. Discuss the concept of a perfect positive correlation.
3. Explain the difference between linear and non-linear correlations.
4. How does one interpret a negative correlation in a practical scenario?
5. Elaborate on the concept of no correlation (zero correlation) between two variables. Provide an example from everyday life where zero correlation might be observed.
6. Discuss how understanding positive correlation can be beneficial in making informed decisions.
7. Define the concept of nonlinear correlation and provide a practical example where such a correlation would be observed.

Assignments

1. In a study analysing the relationship between exercise and heart health, how would you interpret a strong positive correlation between the frequency of exercise and cardiovascular fitness?
2. How would you interpret a moderate negative correlation between temperature and sales of ice cream in a particular region?
3. In a study of the relationship between student attendance and academic performance, how would you explain a weak positive correlation between the two variables?
4. If there is a strong positive correlation between hours spent gaming and reported feelings of aggression in children, what implications might this have for parents and educators?

Suggested Reading

1. Sancheti, & Kapoor, V. K. (2014). *Business mathematics*. Sultan Chand & Sons.
2. Wikes, F. M. (1998). *Mathematics for business, finance and economics*. Thomson Learning.
3. Vittal, P. R. (2012). *Business maths & statistics*. Margham Publications.
4. Dixon, W. S., & Massey, F. J. (1951). *Introduction to statistical analysis*. McGraw Hill Book Company, Inc.
5. Fisher, R. A., & Yates, F. (1948). *Statistical tables for biological, agricultural and medical research* (3rd ed.). New York: Hafner Publishing Company.

Reference

1. Gupta, S. P. (1994). *Statistical Methods*, Sultan Chand & Sons, New Delhi, pp. *E10*, 1-61.
2. Gupta, C. B., & Gupta, V. (2009). *Introduction to statistical methods*. Vikas publishing House Pvt. LTD.
3. Goel, A., & Goel, A. *Mathematics and Statistics*, Taxmann Allied Services Pvt.
4. Kothari, C. R. (2013). *Quantitative Techniques*. Vikas Publishing House.

Unit - 2
Measurement of Correlation

Learning Outcomes

At the conclusion of this unit, the learner will be able;

- ◊ to familiarise with Karl Pearson's coefficient of correlation,
- ◊ to learn the rank correlation method of finding correlation,
- ◊ to acquaint with concurrent deviation method of finding correlation.

Prerequisite

Imagine you are a farmer and you are trying to understand the relationship between rainfall and crop yield. By measuring the correlation between the amount of rainfall in a season and the volume of crops harvested, you can determine if there is a connection between the two. If a strong positive correlation is found, it means that when there is more rainfall, the crop yield tends to increase, and when there is less rainfall, the crop yield decreases. This information can help you make informed decisions about when to plant, what crops to choose, and how to manage water resources efficiently, ensuring a more productive and sustainable farming process. On the other hand, if the correlation is weak or non-existent, it suggests that rainfall might not be the only factor affecting crop yield, and you need to consider other variables like soil quality, temperature, or pest control methods. Understanding this correlation can significantly impact the farming practices and overall agricultural productivity. This unit will throw light into various methods one can use to measure the correlation between variables.

Keywords

Scatter diagram, Correlation Coefficient, Rank Correlation, Concurrent Deviation, Probable error

3.2.1 Methods of finding Correlation

There are several methods to find out the relationship between two variables. Depending on the nature of data, objective behind finding the relation, research question, etc., people can choose which method to adopt for finding the correlation. Following are the most used methods to find out correlation.

3.2.1.1 Scatter Diagram

A scatter plot is one of the simplest and most effective ways to visualise the relationship between two continuous variables. It is the diagrammatic representation of the relationship between two variables. One variable is represented along the X axis and the second variable along the Y axis. Each data point is plotted on the graph with one variable on the X axis and the other on the Y axis. For each pair of observations of two variables, we put a dot in the plane. The pattern of points can provide a visual indication of the strength and direction of the correlation. If the points tend to form a line sloping upward from left to right, it suggests a positive correlation. If the points form a line sloping downward, it suggests a negative correlation. If the points are scattered randomly, there may be no significant correlation. (See the figures below)

Figure 3.1.1 Perfect positive correlation

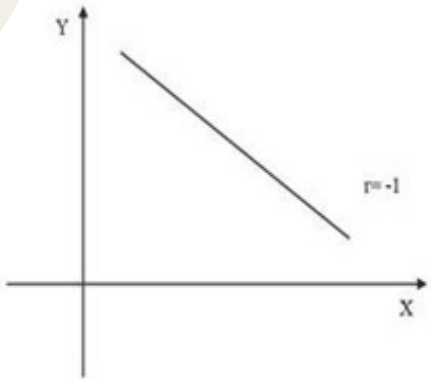


Figure 3.1.2 Perfect negative correlation

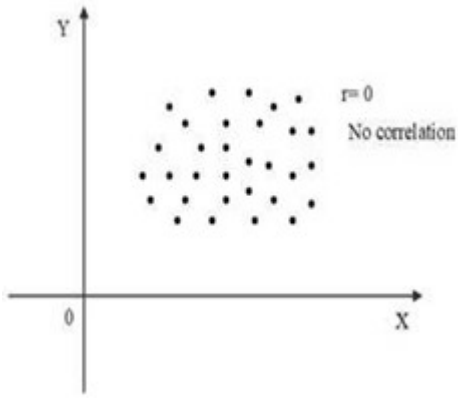


Figure 3.1.3 No correlation

3.2.1.2 Karl Pearson's Coefficient of Correlation

Degree of relationship between two variables is called coefficient of correlation. It is an algebraic method of measuring correlation. Coefficient of correlation is denoted by the symbol r and the value of r lies between −1 and +1.

The correlation coefficient is a value that indicates the strength of the relationship between variables. In case of 'Perfect negative correlation' (value -1), the variables tend to move in opposite directions (i.e., when one variable increases, the other variable strictly decreases).

In the case of 'No correlation' (value 0), the variables do not have a relationship with each other.

In case of 'Perfect positive correlation' (value +1), the variables tend to move in the same direction (i.e., when one variable increases, the other variable also increases).

There are several different kinds of correlation coefficients, but Pearson's is one of the most renowned. Karl Pearson, a biologist, and statistician has given a formula for calculation of coefficient of correlation. It is also known as Product Moment Method. Following are the formulas for finding the correlation coefficient.

1. $r = \dfrac{Cov\,(x,y)}{\sigma(x) \times \sigma(y)}$ where $Cov\,(x,y)$ = covariance of (x,y). Covariance is a statical measure that quantifies the degree to which two variables change together. It is the average of the product of the deviations of the observations from arithmetic mean.

$$i.e, Cov\,(x,y) = \frac{\Sigma((x-\bar{x})\,(y-\bar{y}))}{n}$$

$\sigma(x) = \sqrt{\dfrac{\Sigma(x-\overline{x})^2}{n}}$ is the standard deviation of x

$\sigma(y) = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$ is the standard deviation of y

$\bar{x} = \frac{\sum x}{n}$, $\bar{y} = \frac{\sum y}{n}$ are the arithmetic means

$$i.e, r = \frac{\sum((x-\bar{x})\,(y-\bar{y}))}{n\,\sigma x \times \sigma y}$$

$$= \frac{\frac{\sum((x-\bar{x})\,(y-\bar{y}))}{n}}{\sqrt{\frac{\sum(x-\bar{x})^2}{n}} \times \sqrt{\frac{\sum(y-\bar{y})^2}{n}}}$$

2. $$r = \frac{\sum((x-\bar{x})\,(y-\bar{y}))}{\sqrt{\sum(x-\bar{x})^2} \times \sqrt{\sum(y-\bar{y})^2}}$$

The above formula can be expressed in the following form also

3. $$r = \frac{n\sum xy - \sum x \times \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}}$$

Illustration 3.1.1

Find Karl Pearson's correlation coefficient between x and y for the following data,

$n = 10, \Sigma x = 35, \Sigma x^2 = 203, \Sigma y = 28, \Sigma y^2 = 140, \Sigma xy = 168$

Solution

Karl Pearson's correlation coefficient.

$$r = \frac{n\sum xy - \sum x \times \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}}$$

$$= \frac{10\times168 - 35\times28}{\sqrt{10\times203 - 35^2} \times \sqrt{10\times140 - 28^2}}$$

$$= \frac{10\times168 - 35\times28}{\sqrt{10\times203 - 35^2} \times \sqrt{10\times140 - 28^2}}$$

$$= \frac{1680 - 980}{\sqrt{2030\text{-}1225} \times \sqrt{1400\text{-}784}}$$

$$= \frac{700}{(\sqrt{805}) \times (\sqrt{616})}$$

$$= \frac{700}{28.37 \times 24.82}$$

$$= \frac{700}{704.14}$$

$$= 0.99$$

A correlation coefficient of 0.99 indicates a very strong positive linear relationship between the two variables. This implies that when one variable increases, the other variable tends to increase as well in a consistent and nearly perfect manner. It signifies a highly robust and direct association between the variables.

Illustration 3.1.2

Find Karl Pearson's correlation coefficient between x and y for the following data,

$$n = 15, \quad Cov\,(x,y) = 8.13\,, \quad \sigma x = 3.01, \quad \sigma y = 3.03$$

Solution

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

$$= \frac{8.13}{3.01 \times 3.03}$$

$$= \frac{8.13}{9.12}$$

$$= 0.89$$

Illustration 3.1.3

Find Karl Pearson's correlation coefficient between x and y for the following data,

$$n = 1000, \quad \sigma x = 4.5, \quad \sigma y = 3.6, \quad \Sigma\,(x - \bar{x}) \times (y - \bar{y}) = 4800$$

Solution

$$r = \frac{\Sigma(x - \bar{x})\,(y - \bar{y})}{n\,\sigma(x) \times \sigma(y)}$$

$$= \frac{4800}{1000 \times 4.5 \times 3.6}$$

$$= \frac{4800}{16200}$$

$$= 0.296$$

A Pearson correlation coefficient of 0.296 indicates a weak positive linear relationship between the two variables. This suggests that as one variable increases, the other variable tends to increase as well, but not very strongly. The value signifies a mild and less direct association between the variables.

Illustration 3.1.4

Find Product Moment method of Correlation Coefficient between x and y for the following data,

$$n = 20, \Sigma\,(x - \bar{x})^2 = 136, \Sigma\,(y - \bar{y})^2 = 138, \Sigma\,(x - \bar{x}) \times (y - \bar{y}) = 122$$

Solution

$$r = \frac{\Sigma\,((x\text{-}\bar{x})\,(y\text{-}\bar{y}))}{\sqrt{\Sigma(x\text{-}\bar{x})^2} \times \sqrt{\Sigma(y\text{-}\bar{y})^2}}$$

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

$$= \frac{122}{11.66 \times 11.75}$$

$$= \frac{122}{137.01}$$

$$= 0.89$$

A correlation of 0.89 means the two things are strongly connected: when one goes up, the other usually goes up too, following a straight line. It shows a strong relationship between them.

Illustration 3.1.5

Given : $\sum x = 125$, $\sum y = 100$, $\sum x^2 = 650$, $\sum y^2 = 436$, $\sum xy = 520$ and $n = 25$, obtain the value of Karl Pearson's correlation coefficient $r(X,Y)$.

Solution

$$r(x,y) = \frac{n\sum xy - \sum x \sum y}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$= \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - 125^2}\sqrt{25 \times 436 - 100^2}}$$

$$= \frac{13000 - 12500}{\sqrt{16250 - 15625}\sqrt{10900 - 10000}}$$

$$= \frac{500}{\sqrt{625}\sqrt{900}}$$

$$= \frac{500}{25 \times 30}$$

$$= \frac{500}{750}$$

$$= 0.67$$

Illustration 3.1.6

Calculate the coefficient of correlation for the following table by Karl Pearson's coefficient of correlation.

x	1	2	3	4	5
y	3	1	2	5	4

Solution

$$\bar{x} = \frac{\Sigma x}{n} = \frac{15}{5} = 3$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{15}{5} = 3$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	3	-2	0	0	4	0
2	1	-1	-2	2	1	4
3	2	0	-1	0	0	1
4	5	1	2	2	1	4
5	4	2	1	2	4	1
Total 15	15			6	10	10

i.e., Karl Pearson's coefficient of correlation formula is

$$r = \frac{\Sigma(x-\bar{x})\,(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2} \times \sqrt{\Sigma(y-\bar{y})^2}}$$

$$= \frac{6}{\sqrt{10} \times \sqrt{10}}$$

$$= \frac{6}{10}$$

$$= 0.6$$

Illustration 3.1.7

Calculate the coefficient of correlation for the following table by Karl Pearson's coefficient of correlation.

x	70	69	68	67	66	65	64
y	72	68	70	68	65	67	66

Solution

$$\bar{x} = \frac{\Sigma x}{n} = \frac{469}{7} = 67$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{476}{7} = 68$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
70	72	3	4	9	16	12
69	68	2	0	4	0	0
68	70	1	2	1	4	2
67	68	0	0	0	0	0
66	65	-1	-3	1	9	3
65	67	-2	-1	4	1	2
64	66	-3	-2	9	4	6
469	476			$\Sigma(x-\bar{x})^2=28$	$\Sigma(y-\bar{y})^2=34$	$\Sigma(x-\bar{x})(y-\bar{y})=25$

Karl Pearson's coefficient of correlation formula is

$$r = \frac{\Sigma(x-\bar{x})\,(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2} \times \sqrt{\Sigma(y-\bar{y})^2}}$$

$$= \frac{25}{\sqrt{28} \times \sqrt{34}}$$

$$= \frac{25}{5.2915 \times 5.831}$$

$$= \frac{25}{30.8547}$$

$$= 0.81$$

Illustration 3.1.8

Calculate the coefficient of correlation for the following table by Karl Pearson's coefficient of correlation.

x	9	8	10	12	11	13	14	16	15
y	1	2	3	4	5	6	7	8	9

Solution:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{108}{9} = 12$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{45}{9} = 5$$

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
9	1	-3	-4	9	16	12
8	2	-4	-3	16	9	12
10	3	-2	-2	4	4	4
12	4	0	-1	0	1	0
11	5	-1	0	1	0	0
13	6	1	1	1	1	1
14	7	2	2	4	4	4
16	8	4	3	16	9	12
15	9	3	4	9	16	12
108	45			$\Sigma(x-\bar{x})^2=60$	$\Sigma(y-\bar{y})^2=60$	$\Sigma(x-\bar{x})(y-\bar{y})=57$

Karl Pearson's coefficient of correlation formula is;

$$r = \frac{\sum(x - \bar{x})\,(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2} \times \sqrt{\sum(y - \bar{y})^2}}$$

$$= \frac{57}{\sqrt{60} \times \sqrt{60}}$$

$$= \frac{57}{60}$$

$$= 0.95$$

Illustration 3.1.9

The following are the percentage of marks in Mathematics and Statistics:

Marks in Mathematics:	78	36	98	25	75	82	90	62	65	39
Marks in Statistics:	84	51	91	60	68	62	86	58	53	47

Find Karl Pearson's Coefficient of Correlation.

Solution

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
78	84	13	18	169	324	234
36	51	-29	-15	841	225	435
98	91	33	25	1089	625	825
25	60	-40	-6	1600	36	240
75	68	10	2	100	4	20
82	62	17	-4	289	16	-68
90	86	25	20	625	400	500
62	58	-3	-8	9	64	24
65	53	0	-13	0	169	0
39	47	-26	-19	676	361	494
Total 650	660			5398	2224	2704

$$\bar{x} = \frac{\Sigma x}{n} = \frac{650}{10} = 65$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{660}{10} = 66$$

Karl Pearson's coefficient of correlation formula is

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2} \times \sqrt{\Sigma(y - \bar{y})^2}}$$

$$= \frac{2704}{\sqrt{5398} \times \sqrt{2224}}$$

$$= \frac{2704}{73.471 \times 47.1593}$$

$$= \frac{2704}{3464.84}$$

$$= 0.78$$

3.2.1.3 Spearman's Rank Correlation coefficient

The Spearman Rank Correlation coefficient, often denoted as ρ (rho), is a statistical measure used to assess the strength and direction of the monotonic relationship between two variables. Unlike Pearson's correlation coefficient, which measures linear relationships, Spearman's rank correlation works with ranked or ordinal data and is based on the ranks of the observations rather than their actual values.

Spearman's Rank Correlation coefficient is valuable for non-normally distributed and ordinal data. It detects non-linear relationships, making it versatile for various datasets, especially with outliers. It is crucial in fields like psychology and social sciences where variables can be ranked but not precisely measured.

Calculating Spearman's Rank Correlation Coefficient

Let us use a simple array of data to illustrate the use of Spearman's rank correlation:

Suppose you have data representing the amount of time (in hours) spent commuting to work (X) and the corresponding levels of job satisfaction (Y) for a group of employees:

Commute Time (X)	Job Satisfaction (Y)
1	4
2	3
3	1
4	2
5	5

Let us check the steps for calculating Spearman's Rank Correlation Coefficient:

Step 1: Rank the data for both X and Y, separately, from lowest to highest values. Assign the same rank for tied values.

Commute Time (X)	Rank of X	Job Satisfaction (Y)	Rank of Y
1	1	4	2
2	2	3	3
3	3	1	5
4	4	2	4
5	5	5	1

Step 2: Calculate the differences between the ranks of X and the ranks of Y for each data point.

Commute Time (X)	Rank of X	Job Satisfaction(Y)	Rank of Y	Rank Difference (d)
1	1	4	2	1
2	2	3	3	1
3	3	1	5	2
4	4	2	4	0
5	5	5	1	4

Step 3: Square the rank differences (d) and calculate the sum of squared differences (Σd^2).

$$\Sigma d^2 = 1^2 + 1^2 + 2^2 + 0^2 + 4^2 = 22$$

Step 4: Calculate the Spearman Rank Correlation Coefficient (ρ) using the formula:

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

Where ‘n’ is the number of data points, which in this case is 5.

$$\rho = 1 - \frac{6 \times 22}{5\,(5^2-1)}$$

$$= 1 - \frac{132}{5 \times 24}$$

$$= 1 - \frac{132}{120}$$

$$= 1 - 1.1$$

$$= -0.1$$

The resulting ρ value is approximately -0.1. Since Spearman’s rank correlation ranges from -1 to 1, this suggests a very weak negative correlation between commute time and job satisfaction in this dataset. In other words, as commute time increases, job satisfaction tends to slightly decrease, but the relationship is not very strong.

Illustration 3.1.10

Following are the marks obtained by 10 students of two subjects Mathematics and Physics in a class test. Estimate Spearman’s Rank Correlation.

Name of students	Mathematics	Physics
A	7	6
B	3	2
C	1	1
D	4	5
E	6	8
F	8	7
G	2	3
H	5	4

Solution

Name of students	Mathematics	Rank in Mathematics	Physics	Rank in Physics	d	d^2
A	7	2	6	3	1	1
B	3	6	2	7	1	1
C	1	8	1	8	0	0
D	4	5	5	4	1	1
E	6	3	8	1	2	4
F	8	1	7	2	1	1
G	2	7	3	6	1	1
H	5	4	4	5	1	1
					Σd^2	10

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

$$= 1 - \frac{6 \times 10}{8\,(8^2 - 1)}$$

$$= 1 - \frac{60}{8 \times 63}$$

$$= 1 - \frac{60}{504}$$

$$= 1 - 0.12$$

$$= 0.88$$

A rank correlation coefficient of 0.88 suggests a strong positive relationship between the ranked variables, indicating that they generally change together in the same direction.

Illustration 3.1.11

The data given below relates to the price and demand of a commodity over a period. Compute the correlation coefficient between the Price and Demand.

Price	50	75	60	70	95	90	88
Demand	100	140	110	115	150	134	120

Solution:

Price	Rank of Price	Demand	Rank of Demand	Rank Difference(d)	d^2
50	7	100	7	0	0
75	4	140	2	2	4
60	6	110	6	0	0
70	5	115	5	0	0
95	1	150	1	0	0
90	2	134	3	1	1
88	3	120	4	1	1
					$\Sigma d^2=6$

Spearman's Rank correlation Coefficient $= 1 - \frac{6 \times \Sigma d^2}{n(n^2 - 1)}$

$$= 1 - \frac{6 \times 6}{7(7^2 - 1)}$$

$$= 1 - \frac{36}{336}$$

$$= 1 - 0.1071$$

$$= 0.8929$$

A rank correlation of 0.8929 means a strong tendency for the ranked variables to change together in a similar direction, but not at an even pace. This shows a solid positive relationship between them. The closer the value is to 1, the stronger the connection.

Computation for tied observations (Repeated Rank)

If there exist two or more same or equal values the ranking is said to be tied. In such case an average rank is to be given to each individual item. For example, if the value is repeated twice at the second rank, the common rank is to be assigned to each item is $\frac{2+3}{2} = 2.5$ which is the average of 2 and 3.

When equivalent ranks are assigned to multiple entries, certain adjustments in the formula become necessary for calculating the Rank Correlation coefficient. This adjustment involves adding $\frac{m^3-m}{12}$ to the sum of squared differences $\sum D^2$, where m stands for the number of items which have the common rank. In case, there are more than one such group of items with same rank, the value is added as many times as the number of such groups. The formula in that case is written as

$$\rho = 1 - \frac{6[(\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \dots\dots]}{n(n^2 - 1)}$$

Illustration 3.1.12

Compute the correlation coefficient between the two variables x and y

W	2.7	2.5	2.7	2.5	2.5	2.7	3.5	4.9	5.4	5.4	3.8
y	30	27	50	76	37	14	34	75	60	121	74

Solution

x	y	Rank of x	Rank of y	d	d^2
2.7	30	7	9	2	4
2.5	27	10	10	0	0
2.7	50	7	6	1	1
2.5	76	10	2	8	64
2.5	37	10	7	3	9
2.7	14	7	11	4	16
3.5	34	5	8	3	9
4.9	75	3	3	0	0
5.4	60	1.5	5	3.5	12.25
5.4	121	1.5	1	0.5	0.25
3.8	74	4	4	0	0
					$\Sigma d^2 = 115.50$

In X series 5.4 occurs twice. The rank of 5.4 is $\frac{1+2}{2}$=1.5. The next lower value 4.9 is assigned its rank as 3. Similarly, 2.7 occurs thrice. The rank of 2.7 is calculated as $\frac{6+7+8}{3}$=7 and next 2.5 occurs thrice. The rank of 2.5 is $\frac{9+10+11}{3}$=10. Hence, on these common ranks the coefficient of correlation will have to be corrected by adding $\frac{\sum(m^3-m)}{12}$ to the value of Σd^2.

In respect of the X series this addition will be $\frac{1}{12}(2^3\text{-}2)+\frac{1}{12}(3^3\text{-}3)+\frac{1}{12}(3^3\text{-}3)$ as 5.4 occurs twice, 2.7 occurs thrice and 2.5 occurs thrice. This correction value calculated as below is to be added to rank correlation equation.

$$\rho = 1 - \frac{6[(\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \dots\dots]}{n(n^2 - 1)}$$

$$= 1 - \frac{6[115.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(3^3 - 3)]}{11(11^2 - 1)}$$

$$= 1 - \frac{6[115.5 + \frac{1}{12} \times 6 + \frac{1}{12} \times 24 + \frac{1}{12} \times 24]}{11 \times 120}$$

$$= 1 - \frac{6[115.5 + \frac{1}{2} + 2 + 2]}{1320}$$

$$= 1 - \frac{6[115.5 + 4.5]}{1320}$$

$$= 1 - \frac{6[115.5 + 4.5]}{1320}$$

$$= 1 - \frac{6 \times 120}{1320}$$

$$= 1 - \frac{720}{1320}$$

$$= 1\text{- } 0.545$$

$$= 0.454$$

Illustration 3.1.13

Calculate the coefficient of rank correlation from the following data,

W	48	33	40	9	16	16	65	24	16	57
y	13	31	31	6	15	4	20	9	6	19

Solution

Ranks are assigned as follows for x series;

We see that 16 is repeating three times. i.e, 7th , 8th and 9th rank is repeating. So the average $\frac{7+8+9}{3}=\frac{24}{3}=8$ is assigned to 16. So $m = 3$.

Ranks are assigned for y series;

31 is repeating two times. i.e, 1st and 2nd rank is repeating. So the average $\frac{1+2}{2}=\frac{3}{2}=1.5$ is assigned to 31. So m = 2, 6 is repeating two times. i.e, 8th and 9th rank is repeating. So the average $\frac{8+9}{2}=\frac{17}{2}=8.5$ is assigned to 6 . So m = 2

X	Rank in X	Y	Rank in Y	d	d^2
48	3	13	6	3	9
33	5	31	1.5	3.5	12.25
40	4	31	1.5	2.5	6.25
9	10	6	8.5	1.5	2.25
16	8	15	5	3	9
16	8	4	10	2	4
65	1	20	3	2	4
24	6	9	7	1	1
16	8	6	8.5	0.5	0.25
57	2	19	4	2	4
Total					52

$$\rho = 1 - \frac{6[(\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\ldots]}{n(n^2 - 1)} \ldots$$

$$= 1 - \frac{6[52 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)]}{n(n^2 - 1)}$$

$$= 1 - \frac{6[52 + \frac{1}{12}(24) + \frac{1}{12}(6) + \frac{1}{12}(6)]}{10(10^2 - 1)}$$

$$= 1 - \frac{6[52 + 2 + \frac{1}{2} + \frac{1}{2}]}{10(100 - 1)}$$

$$= 1 - \frac{6[52 + 3]}{10 \times 99}$$

$$= 1 - \frac{6 \times 55}{10 \times 99}$$

$$= 1 - \frac{330}{990}$$

$$= 1 - 0.33$$

$$= 0.67$$

Illustration 3.1.14

Find the rank correlation coefficient for the following data

x:	72	60	63	75	60	60	55	67	68	70
y:	70	60	63	75	58	68	60	80	65	65

Solution

x	y	Rank of x	Rank of y	d	d^2
72	70	2	3	1	1
60	60	8	8.5	0.5	0.25
63	63	6	7	1	1
75	75	1	2	1	1
60	58	8	10	2	4
60	68	8	4	4	16
55	60	10	8.5	1.5	2.25
67	80	5	1	4	16
68	65	4	5.5	1.5	2.25
70	65	3	5.5	2.5	6.25
					$\Sigma d^2 = 50$

$$\rho = 1 - \frac{6[(\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\ldots]}{n(n^2 - 1)} \ldots$$

$$= 1 - \frac{6[50 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)]}{n(n^2 - 1)}$$

$$= 1 - \frac{6[50 + \frac{1}{12}(24) + \frac{1}{12}(6) + \frac{1}{12}(6)]}{10(10^2 - 1)}$$

$$= 1 - \frac{6[50 + 2 + \frac{1}{2} + \frac{1}{2}]}{10(100 - 1)}$$

$$= 1 - \frac{6[50 + 3]}{10 \times 99}$$

$$= 1 - \frac{6 \times 53}{10 \times 99}$$

$$= 1 - \frac{318}{990}$$

$$= 1 - 0.321$$

$$= 0.679$$

Illustration.3.1.15

Given the following aptitude and I.Q. scores for a group of students. Find the coefficient of rank correlation

Aptitude Score	57	58	59	59	60	61	60	64
I.Q. Score	97	108	95	106	120	126	113	110

Solution

X	Y	Rank in X	Rank in Y	d	d^2
57	97	8	7	1	1
58	108	7	5	2	4
59	95	5.5	8	2.5	6.25
59	106	5.5	6	0.5	0.25
60	120	3.5	2	1.5	2.25
61	126	2	1	1	1
60	113	3.5	3	0.5	0.25
64	110	1	4	3	9
					24

Here two correction factors are to be added to the equation, for X series, 59 is repeated twice, so the correction factor is $\frac{2^3 - 2}{12}$ is added and 60 is repeated twice, so the correction factor is $\frac{2^3 - 2}{12}$ is added.

The rank correlation coefficient is

$$r = 1 - \frac{6[(\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\ldots]}{n(n^2 - 1)} \ldots$$

$$= 1 - \frac{6[24 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)]}{8(8^2 - 1)}$$

$$= 1 - \frac{6[24 + \frac{1}{12}(6) + \frac{1}{12}(6)]}{8(8^2 - 1)}$$

$$= 1 - \frac{6[24 + \frac{1}{2} + \frac{1}{2}]}{8(64 - 1)}$$

$$= 1 - \frac{6[25]}{8 \times 63}$$

$$= 1 - \frac{150}{504}$$

$$= 1 - 0.2976$$

$$= 0.7024$$

3.2.1.4 Concurrent Deviation Method

The concurrent deviation method is a technique used for finding the correlation coefficient between two variables. This method involves finding the deviations of the of the values of the variable from its preceding value and does not take into account the exact magnitude of the values of the variables. Thus, we put a plus (+) sign, minus (–) sign or equality (=) sign for the deviation if the value of the variable is greater than, less than or equal to the preceding value respectively. It is a straightforward way to calculate the correlation between two sets of data points. In this method only the directions of deviations are taken and ignore magnitudes. If the deviations of the two variables are concurrent then they move in the same direction, otherwise in the opposite direction.

The formula for finding coefficient of concurrent deviation is:

$$r = \pm\sqrt{\pm\frac{2C - n}{n}}$$

Where; C = The number of pairs of concurrent deviation and

n = The number of pairs of deviations.

It should be clearly noted that here n is not the number of pairs of observations but

it is the number of pairs of deviations and as such it is one less than the number of pairs of observations.

Since $-1 \leq r \leq 1$, the quantity inside the square root, $\pm \frac{2c-n}{n}$ must be positive, otherwise r will be imaginary which is not possible. Thus if $(2c-n)$ is positive, we take positive sign in and outside the square root in $\pm\sqrt{\pm \frac{2c-n}{n}}$ and if $(2c-n)$ is negative, we take negative sign in and outside the square root in $\pm\sqrt{\pm \frac{2c-n}{n}}$.

Illustration 3.1.16

Find out the correlation between two arrays of variables X and Y using the concurrent deviation method.

X	13	18	23	8	21	25	28	10	22
Y	23	11	17	3	23	18	8	23	20

Solution

X	Y	Sign of deviations from the proceeding value in X *dx*	Sign of deviations from the proceeding value in Y *dy*	*dx* × *dy*
13	23	-------	-----	------
18	11	+	-	-
23	17	+	+	+
8	3	-	-	+
21	23	+	+	+
25	18	+	-	-
28	8	+	-	-
10	23	-	+	-
22	20	+	-	-

C = Number of + signs in $dx \times dy$ column = 3

$n = 9 - 1 = 8$ (first pair of observation is not compared)

$$r = \pm\sqrt{\pm\frac{2c-n}{n}}$$

$$= \pm\sqrt{\pm\frac{2\times3-8}{8}}$$

$$= \pm\sqrt{\pm\frac{-2}{8}}$$

$$= \pm\sqrt{\pm(-0.25)}$$

= -0.5

Since $2c - n = -0.25$, i.e., (negative), we take negative sign inside and outside the square root to get, $r = -\sqrt{-(-0.25)} = -0.5$.

Illustration 3.1.17

Calculate the coefficient of concurrent deviations for the following data

Supply	65	40	35	75	63	80	35	20	80	60	50
Demand	60	55	50	56	30	70	40	35	80	75	80

Solution

Supply	Demand	Direction of change in x (dx)	Direction of change in y (dy)	$dx \times dy$
65	60	-------	-----	------
40	55	-	-	+
35	50	-	-	+
75	56	+	+	+
63	30	-	-	+
80	70	+	+	+
35	40	-	-	+
20	35	-	-	+
80	80	+	+	+
60	75	-	-	+
50	80	-	+	-

C = Number of + signs in $dx \times dy$ column = 9

$n = 11 - 1 = 10$ (first pair of observation is not compared)

$$r = \pm\sqrt{\pm\frac{2C-n}{n}}$$

$$= \pm\sqrt{\pm\frac{2\times 9 - 10}{10}}$$

$$= \pm\sqrt{\pm\frac{8}{10}}$$

$$= \pm\sqrt{\pm(0.8)}$$

$$= 0.89$$

Illustration 3.1.18

Calculate the coefficient of concurrent deviations from the data given below :

Year	1993	1994	1995	1996	1997	1998	1999	2000	2001
Supply	160	164	172	182	166	170	178	192	186
Demand	292	280	260	234	266	254	230	190	200

Solution

Year	Supply	Demand	Direction of change in x (dx)	Direction of change in y (dy)	$dx \times dy$
1993	160	292	-------	-----	------
1994	164	280	+	-	-
1995	172	260	+	-	-
1996	182	234	+	-	-
1997	166	266	-	+	-
1998	170	254	+	-	-
1999	178	230	+	-	-
2000	192	190	+	-	-
2001	186	200	-	+	-

C = Number of + signs in $dx \times dy$ column = 0

$n = 9 - 1 = 8$ (first pair of observation is not compared)

$$r = \pm\sqrt{\pm\frac{2C - n}{n}}$$

$$= \pm\sqrt{\pm\frac{2 \times 0 - 8}{8}}$$

$$= \pm\sqrt{\pm\frac{-8}{8}}$$

$$= \pm\sqrt{\pm(-1)}$$

Since $2c - n = -8$, i.e., (negative), we take negative sign inside and outside the square root to get, $r = -\sqrt{-(-1)} = -1$. Hence, there is perfect negative correlation between the supply and the price.

Illustration 3.1.19

Calculate the coefficient of concurrent deviations from the following data .

No. of pairs of observations = 96

No. of pairs of concurrent deviations = 36

Solution

C = Number of pairs of concurrent deviations = 36

$n = 96 - 1 = 95$ (first pair of observation is not compared)

$$r = \pm\sqrt{\pm\frac{2C - n}{n}}$$

$$= \pm\sqrt{\pm\frac{2 \times 36 - 95}{95}}$$

$$= \pm\sqrt{\pm\frac{72 - 95}{95}}$$

$$= \pm\sqrt{\pm\frac{-23}{95}}$$

$$= \pm \sqrt{\pm(-0.242)}$$

$$= - \sqrt{-(-0.242)}$$

$$= - 0.492$$

Note:

Increase in the value is denoted by + sign and decrease by – sign

r is positive when $2C > N$ and r is negative when $2C < N$

3.2.2 Probable Error

In correlation analysis, the probable error refers to an estimate of the possible deviation between the sample correlation and the true correlation in the population. It helps to determine the accuracy and reliability of the correlation coefficient calculated from a sample, indicating the likelihood of the sample correlation differing from the actual correlation within the entire population.

Generally, Probable error is used to measure validity of the value of correlation coefficient. Correlation coefficient may vary for different samples drawn from the same population. But, the numerical value of such variations is expected to be less than the probable error.

Actual correlation coefficient lies between the limits (Coefficient of correlation ± Probable error)

$$\text{Probable Error} = 0.6745 \times \frac{(1 - r^2)}{\sqrt{n}}$$

where r is the correlation coefficient and 'n' number of pairs of observation.

In correlation analysis, the standard error measures the accuracy of the sample correlation coefficient by estimating the variability or dispersion of the calculated correlation from the true correlation in the population. It provides a measure of how the sample correlation may vary from the actual correlation if the study were to be repeated multiple times. The standard error (SE) indicates the precision of the sample correlation coefficient and is calculated using the formula: SE = $\frac{(1-r^2)}{\sqrt{n}}$

Illustration 3.1.20

Find Probable error and Standard error if r = 0.8 and n= 100

$$\text{Probable Error} = \frac{0.6745\,(1 - r^2)}{\sqrt{n}}$$

$$= \frac{0.6745(1 - 0.64)}{\sqrt{100}}$$

$$= 0.6745 \times \frac{0.36}{10} = 0.024$$

Standard Error $= \frac{(1 - r^2)}{\sqrt{n}}$

$$= \frac{1 - 0.64}{\sqrt{100}}$$

$$= \frac{0.36}{10}$$

$$= 0.036$$

Illustration.3.1.21

The correlation coefficient between two variables for 20 items is 0.96. Find the probable error Also determine the limits for population r.

Solution

Given $r = 0.96, \quad n = 20$

$$P.E = 0.6745 \times \frac{1 - (0.96)^2}{\sqrt{20}}$$

$$= 0.6745 \times \frac{1 - 0.9216}{\sqrt{20}}$$

$$= 0.6745 \times \frac{0.0784}{\sqrt{20}}$$

$$= 0.6745 \times 0.0175$$

$$= 0.0118$$

limits for population $r = r \pm P.E.(r)$

$$0.96 \pm 0.0118$$

$$(0.96 - 0.0118, \quad 0.96 + 0.0118)$$

$$(0.9482, \quad 0.9718)$$

Recap

- ◊ Scatter Diagram - diagrammatic representation of the relationship between two variables
- ◊ Correlation coefficient - Degree of relationship between two variables
- ◊ Karl Pearson's Correlation Coefficient -

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} * \sqrt{n\sum y^2 - (\sum y)^2}}$$

- ◊ Spearman's Rank Correlation - $\rho = 1 - \frac{6\Sigma d^2}{n\,(n^2-1)}$
- ◊ Coefficient of Concurrent Deviation - $\pm\sqrt{\pm\frac{2C-n}{n}}$
- ◊ Probable error - estimate of the possible deviation between the sample correlation and the true correlation in the population
- ◊ Standard error - measure of how the sample correlation vary from the actual correlation if the study were to be repeated multiple times

Objective Questions

1. What method involves finding the average of the products of the deviations from the means of two variables?
2. What statistical measure helps evaluate the strength and direction of the linear relationship between two continuous variables?
3. Which correlation method involves ranking the data and finding the differences between the ranks to quantify the relationship between variables?
4. What is a scatter diagram in the context of correlation analysis?
5. What does the probable error indicate in correlation analysis?
6. What does rank correlation evaluate?
7. What does the coefficient of correlation measure?

Answers

1. The method of concurrent deviation.
2. Pearson correlation coefficient
3. Spearman rank correlation coefficient.
4. The visual representation of the relationship between two variables.
5. An estimate of the possible deviation between the sample correlation and the true correlation in the population.
6. Association between variables without assuming linearity.
7. Strength of linear relationship.

Self-Assessment Questions

1. Define the term 'correlation coefficient.'
2. Explain the significance of a correlation coefficient of 0.
3. Describe the steps involved in calculating the Pearson correlation coefficient.
4. Find the correlation coefficient between x and y series for the following data,

$$n = 15, \sigma x = 3.01, \sigma y = 3.03, \sum(x - \bar{x})(y - \bar{y}) = 122$$

5. The correlation coefficient between two variables x and y is 0.48. The covariance is 36 and the variance of x is 16. Find the standard deviation of y.
6. Find Karl Pearson's correlation coefficient between x and y for the following data,

$$n = 10, \Sigma x = 12., \Sigma x2 = 450, \Sigma y = 15, \Sigma y2 = 320, \Sigma xy = 285$$

7. What is a Scatter Diagram? How is it useful in the study of Correlation?

Assignments

1. Find Rank Correlation Coefficient

X	50	95	92	78	72	68	56	52
Y	44	22	29	34	27	37	36	42

2. Apply Spearman Rank Correlation Coefficient for following data

X	12	18	21	13	19	21	17	12	21	8
Y	8	15	15	17	21	25	21	19	8	10

3. Compute coefficient of correlation for the following table by Karl Pearson's coefficient of correlation.

X	12	13	14	15	16	17	18	19	20	21
Y	31	30	30	26	25	25	25	22	19	16

4. Calculate the coefficient of correlation between the Price and Supply

Price	24	29	24	18	30
Supply	17	13	15	19	11

5. Find the coefficient of rank correlation for the following data

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

6. Ten students obtained the following marks in Statistics and Accountancy. Compute Spearman's Rank Correlation Coefficient?

Statistics	115	109	112	87	98	120	98	100	98	118
Accountancy	75	73	85	70	76	82	65	73	68	80

7. Calculate probable error and limits of population from the following data, $n = 7$ *and* $r = 0.96$.

Suggested Reading

1. Sancheti, & Kapoor, V. K. (2014). *Business mathematics*. Sultan Chand & Sons.
2. Wikes, F. M. (1998). *Mathematics for business, finance and economics*. Thomson Learning.
3. Vittal, P. R. (2012). *Business maths & statistics*. Margham Publications.
4. Dixon, W. S., & Massey, F. J. (1951). *Introduction to statistical analysis*. McGraw Hill Book Company, Inc.
5. Fisher, R. A., & Yates, F. (1948). *Statistical tables for biological, agricultural and medical research* (3rd ed.). New York: Hafner Publishing Company.

Reference

1. Gupta, S. P. (1994). *Statistical Methods*, Sultan Chand & Sons, New Delhi, pp. *E10*, 1-61.
2. Gupta, C. B., & Gupta, V. (2009). *Introduction to statistical methods*. Vikas publishing House Pvt. LTD.
3. Goel, A., & Goel, A. *Mathematics and Statistics*, Taxmann Allied Services Pvt.
4. Kothari, C. R. (2013). *Quantitative Techniques*. Vikas Publishing House.

Unit - 3
Regression Analysis

Learning Outcomes

At the conclusion of this unit, the learner will be able to;

- ◊ Familiarise the concept of linear regression and its application,
- ◊ Learn the different types of regression and its significance,
- ◊ Develop the knowledge on methods of finding regression and interpreting the results.

Prerequisite

Imagine you are curious about how the price of a house shifts as the house gets bigger or smaller. When you are planning to buy a house, this analysis can provide you with valuable insights, helping you anticipate how much the price might change based on the size of the house. For sellers, it can guide you in setting a competitive price. For such cases, by examining data on different houses and their prices, how much the price generally changes for each additional square foot of space can be understood. Here, what we are actually exploring is the causal relationship between two variables, price of the house and its size. The variable price of the house is depended on its relative size. For discovering such relationships between variables, a statistical analysis called regression will come to your rescue. This unit will throw light into the meaning and application of regression analysis.

Keywords

Dependent Variable, Independent Variable, Linear regression, Non-linear regression

Discussion

Most of you are familiar with the term 'variable' by now. As we mentioned earlier, variables can be of two types. One is independent variable and the other one is dependent variable.

In any study, the *dependent variable* is the one that researchers are trying to understand or predict. It is the outcome that they are interested in. On the other hand, the *independent variable* is the one that they think might be influencing the dependent variable. It is the factor that researchers manipulate to see how it affects the dependent variable.

For example, let us say you are investigating how the amount of water given to plants affects their growth. In this case, the growth of the plants is the dependent variable because it is the outcome you are interested in. The amount of water you give to the plants is the independent variable because it is what you are changing to see how it impacts the plant's growth.

Understanding the relationship between dependent and independent variables helps us make informed conclusions about how different factors influence specific outcomes. This is where we use regression analysis.

3.3.1 Introduction to Regression Analysis

Regression analysis is like a helpful tool that helps us understand how things are connected. It is like when we want to figure out how changes in one thing might cause changes in another. For example, it can help us understand how exercise affects our weight or how studying impacts our test scores. By using regression analysis, we can make better predictions and understand how different things in our lives might be related.

Generally, Regression analysis means the estimation or prediction of the unknown value of one variable from the known value of the other variable.

According to M. M. Blair "Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data"

Mathematically, we can express regression simply as $y = a + bx$ where a, b is taken from real numbers and x is a variable. The variable x can assume any value from real numbers and thus called independent variable and value of y is depend upon the value of x or its value depends on changes in the independent variable and hence called dependent variables. To understand this better we can take the example of Advertisement expense and Sales of a company (refer figure 2.3.1).

As you can observe from the figure, there is a linear connection between the variables. Here, Advertisement is an independent variable as it can accept any value but

Sales depend upon Advertising and hence Sales is a dependent variable.

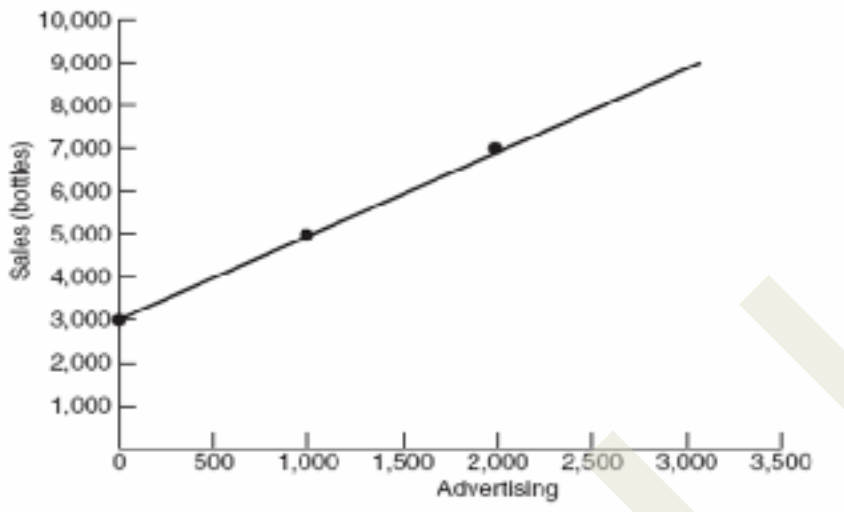


Figure 2.3.1 Relationship between Advertisement and Sales

In this graph, the horizontal axis shows the amount of money spent on advertising, and the vertical axis represents the number of products sold. The points on the graph represent your actual data. The line is one that best fits all the actual data points, showing you the overall trend. If the line slopes upwards, it means that as you spend more on advertising, your sales tend to increase. If the line slopes downwards, it means that increasing advertising spending might not be as effective in boosting sales.

If you check the data points, at '0' Advertisement level, Sales is 3000 which means that the level of Sales will be a minimum 3000 always. When Advertisement increased to 1000, the Sales has become 5000, which you can express as the minimum Sales (3000) + twice the Advertisement expense $(2 \times 1000 = 2000)$. If you can verify this with the third data point, again Sales 7000 = 3000 (minimum sales) + 2 × 2000 (twice the advertisement expense). Now using this degree of relationship between Sales and Advertisement you can estimate or predict the value of Sales at different levels of advertisement and vice versa. The regression equation to indicate this relationship can be written as:

Sales = 3000 + 2 × Advertisement

Simply, regression is the measure of the average relationship between two or more variables. The forecasting is based on average relationship derived statistically by regression analysis. The equation, linear or otherwise is called the regression equation.

3.3.2 Types of Regression

The regression analysis can be classified into

a. Simple and multiple Regression.

b. Linear and nonlinear Regression.

c. Total and partial Regression.

3.3.2.1 Simple and Multiple Regression

Simple regression helps us understand the relationship between two variables, one independent and one dependent. For example, let us consider how the price of a car (dependent variable) might be determined by its age (independent variable). Using simple regression, we can predict how the price changes as the car gets older. Recall the previous example of influence of advertising expenditure on sales turnover. Here two variables are considered for relationship. If y is sales turnover and x is the advertising expenditure then the turnover relationship is expressed as $y = f(x)$

When there are only two variables the regression equation obtained is called simple regression equation.

Multiple regression, on the other hand, helps us understand the relationship between a dependent variable and two or more independent variables. For instance, if you want to predict the income of individuals based on their education level and years of experience, multiple regression can help analyse how both education and experience impact their income. It allows us to see how each of these factors contributes to the overall variation in income.

From the earlier example, turnover z is depended on advertising expenditure (x) and income of people (y) and z is depended variable and others are independent variables. Then functional relationally is expressed as $z = f(x,y)$ and is called multiple regression.

A Multiple regression equation is an equation for estimating the value of a depending variable 'z'.

From the values of the independent variables x any y and is called a regression equation of z on x and y and is denoted by $z = f(x,y)$.

3.3.2.2 Linear and Non-linear Regression

Linear regression is used when the relationship between two variables seems to follow a straight line. For example, if we want to understand how the amount of sugar, we use affects the sweetness of a recipe, and we notice that as we add more sugar, the sweetness increases at a constant rate, then it is a case for linear regression.

Non-linear regression, on the other hand, is like drawing a curved line through a scatter plot. It is used when the relationship between variables does not seem to be a straight line but follows a curve. For example, if we are looking at how a person's height affects their weight and we notice that initially, weight increases rapidly with height, but then levels off, it is a situation for non-linear regression.

In simple terms, linear regression deals with relationships that can be represented by

a straight line, while nonlinear regression deals with relationships that require a curve to accurately represent the data. Linear and non-linear regression can be both simple and multiple depending on the situation.

3.3.2.3 Total and Partial Regression

Total and partial regression are concepts that arise in the context of multiple regression analysis.

Total regression refers to the relationship between the dependent variable and all the independent variables in the model. It examines how the combination of all independent variables collectively influences the dependent variable. For example, in a model that predicts the price of a house based on its size, number of bedrooms, and location, the total regression considers how all these factors together impact the house price.

Partial regression, on the other hand, focuses on the relationship between a specific independent variable and the dependent variable while controlling for the effects of other variables. It helps to understand the unique contribution of each independent variable to the dependent variable. For instance, in the same house price prediction model, partial regression could isolate the influence of the size of the house on the price, considering the effects of the number of bedrooms and location.

3.3.3 Difference between Correlation and Regression

Correlation and Regression are different statistical concepts, although they are related. Here is a concise explanation of their differences:

Point of Difference	***Correlation***	***Regression***
Purpose	Measures the strength and direction of a relationship between two variables	Predicts the value of the dependent variable based on the independent variable(s)
Focus	Focuses on the association between variables	Focuses on the degree of causal relationship between variable(s)
Analysis	Analyses the degree of relationship between variables without establishing cause and effect	Analyses the nature of the relationship and estimates the effect of the independent variable(s) on the dependent variable

Result Range	Ranges from -1 to 1, indicating the strength and direction of the relationship	No specific range for the results; the emphasis is on the estimation and prediction of the dependent variable
Output	Provides a correlation coefficient (r) that indicates the strength and direction of the relationship	Provides an equation that represents the relationship between the variables
Application	Commonly used to examine the relationship between two continuous variables	Commonly used to make predictions, understand cause and effect, and control one variable to understand its effect on another

Recap

- ◊ Dependent variable - the outcome variable, value of which depends on changes in the independent variable.
- ◊ Independent variable - a variable which can accept any existing value under consideration.
- ◊ Regression analysis - the estimation of unknown value of one variable from the known value of the other variable.
- ◊ Simple regression - the relationship between two variables
- ◊ Multiple regression - the relationship between a dependent variable and two or more independent variables
- ◊ Linear regression - relationship follows straight line
- ◊ Non-linear regression - relationship does not follow straight line
- ◊ Total regression - the relationship between the dependent variable and all the independent variables in the model.
- ◊ Partial regression - the relationship between an independent variable and the dependent variable while controlling for other variables.

Objective Questions

1. What does the independent variable represent in a study?
2. What is the main purpose of studying dependent variables in research?
3. Which analysis is used to predict one variable using another variable?
4. What type of regression is used when the relationship between variables can be represented by a straight line?
5. What type of regression is appropriate when the relationship between variables does not follow a straight line?
6. Which regression type is helpful when there are multiple independent variables in the analysis?
7. What kind of regression examines the relationship between the dependent variable and all the independent variables in the model?
8. In which type of regression is the focus on the relationship between a specific independent variable and the dependent variable while controlling for the effects of other variables?
9. What kind of relationship between variables is measured using regression?

Answers

1. The variable that is being manipulated or changed.
2. To observe how they respond to changes in the independent variable.
3. Regression analysis
4. Linear regression.
5. Nonlinear regression.
6. Multiple regression
7. Total regression
8. Partial regression
9. Cause and effect relationship

Self-Assessment Questions

1. Explain independent variable with an example?
2. Explain dependent variable with an example?
3. Explain the concept of linear regression and provide a real-world example where it can be applied.
4. Elaborate on the importance of nonlinear regression in modeling complex relationships between variables.
5. Compare and contrast multiple regression and simple regression, highlighting the advantages and applications of each in real-world data analysis
6. Discuss the significance of using multiple regression in fields like economics, where several independent variables can influence a single dependent variable. Provide an example to illustrate your point.
7. Differentiate between total and partial regression

Assignments

1. A biologist is studying the growth of a population of organisms over time. Initially, the population grows rapidly, but as resources become limited, the growth rate levels off. The biologist wants to understand and predict the growth pattern accurately. Why is it necessary to use nonlinear regression in this population growth study?
2. Suppose you are an agricultural researcher studying the relationship between fertilizer usage and crop yields in a specific region. How might linear regression be useful in this case? How could the findings from the linear regression analysis be practically applied to improve agricultural practices in the region?
3. Suppose you are a real estate analyst aiming to predict housing prices in a particular city. How might multiple regression be useful in this case? How could the results of the multiple regression analysis benefit various stakeholders such as homebuyers, sellers, and real estate developers?
4. How would you predict students' exam scores based on the number of hours they study, considering other potential influencing factors such as attendance and prior academic performance?

Suggested Reading

1. Sancheti, & Kapoor, V. K. (2014). *Business mathematics*. Sultan Chand & Sons.
2. Wikes, F. M. (1998). *Mathematics for business, finance and economics*. Thomson Learning.
3. Vittal, P. R. (2012). *Business maths & statistics*. Margham Publications.
4. Dixon, W. S., & Massey, F. J. (1951). *Introduction to statistical analysis*. McGraw Hill Book Company, Inc.
5. Fisher, R. A., & Yates, F. (1948). *Statistical tables for biological, agricultural and medical research* (3rd ed.). New York: Hafner Publishing Company.

Reference

1. Gupta, S. P. (1994). *Statistical Methods*, Sultan Chand & Sons, New Delhi, pp. *E10*, 1-61.
2. Gupta, C. B., & Gupta, V. (2009). *Introduction to statistical methods*. Vikas publishing House Pvt. LTD.
3. Goel, A., & Goel, A. *Mathematics and Statistics*, Taxmann Allied Services Pvt.
4. Kothari, C. R. (2013). *Quantitative Techniques*. Vikas Publishing House.

Unit - 4

Methods of Regression

Learning Outcomes

At the conclusion of this unit, the learner will be able to;

- ◊ familiarise the concept of regression line,
- ◊ learn different methods of regression analysis,
- ◊ apply the knowledge of regression techniques to practical scenarios,
- ◊ develop the ability to evaluate and interpret the results of regression analysis.

Prerequisite

As you know, is a powerful tool for understanding and modeling relationships between variables. It allows us to make sense of complex relationships in data, predict outcomes, and draw meaningful insights. Are you wondering how to do this with original set of data?

In this unit, we probe into the powerful realm of statistical modeling, where we seek to understand and quantify the relationships between variables. Regression analysis serves as a fundamental tool, allowing us to explore and uncover patterns, make predictions, and gain valuable insights from data. Whether you are interested in predicting sales, understanding the impact of marketing efforts, or exploring the intricate dynamics between variables, regression analysis provides a robust framework for investigation. Our exploration will cover both the theoretical foundations and practical applications of regression from the classic algebraic approach to the graphical representation of data. This will equip you with a diverse set of tools to analyse, interpret, and draw meaningful conclusions from real-world data. So, let us embark on this intellectual adventure and unravel the mysteries hidden within the data through the lens of regression analysis.

Keywords

Regression Line, Regression Equation, Normal Equations, Coefficient of Regression

Discussion

3.4.1 Regression Line/ Line of Best Fit

A regression line is a straight line that best represents the relationship between a dependent variable and one or more independent variables in a dataset. It is the central element of a linear regression model, which aims to minimize the differences between the observed values and the values predicted by the line. The regression line is determined by the regression equation, which mathematically defines the relationship between the variables.

In simple terms, the regression line is a visual representation of how a change in one variable is associated with a change in another. It helps us understand the general trend or pattern in the data and enables us to make predictions about the dependent variable based on the values of the independent variable(s). By fitting the regression line to the data points, we can better grasp the overall relationship between the variables and use this knowledge to make informed decisions or predictions. As mentioned earlier fitting a regression line can be done in many ways.

Properties of Regression line

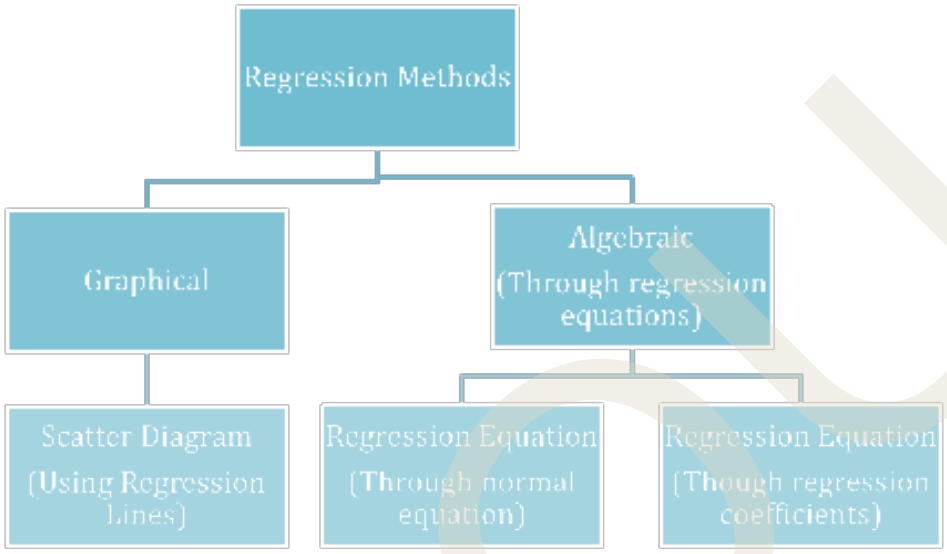
- The regression line is constructed to minimise the sum of squared residuals (the differences between observed and predicted values), ensuring the line provides the best fit to the data.
- The two regression lines cut each other at the point of average of x and average of y (i.e.,x and y)
- When r = 1, the two regression lines coincide each other and give one line.
- When r = 0, the two regression lines are mutually perpendicular.
- The independent variable is not influenced by the errors in the dependent variable, ensuring that causality is properly addressed.
- The regression line can be used to predict values of the dependent variable based on the known values of the independent variable(s).

The regression line assumes a linear relationship between the variables. If the true relationship is nonlinear, the regression line might not accurately represent the

relationship.

3.4.2 Methods of Regression Analysis

Applying actual data points, the relationship between variables can be estimated in different ways using regression analysis. There are mainly two types of regression analysis methods used commonly. These are explained in detail here.



3.4.2.1 Graphical/ Free hand Curve Method

It is the easiest method for obtaining regression line. As per this method independent variable is taken along the horizontal axis and dependent variable along the vertical axis. Original data are plotted in the graph paper. We can draw a smooth free hand line such that the area below and above the line are approximately equal. This is the regression line.

Once the regression line is drawn, it can be used to make predictions about the dependent variable based on the values of the independent variable(s) within the range of the data.

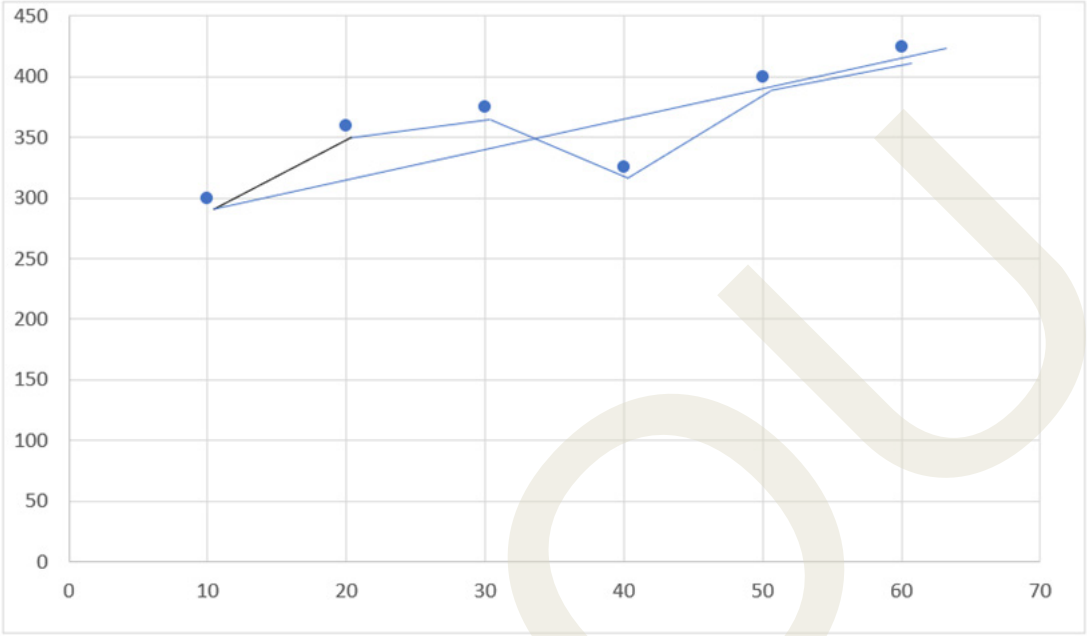
The graphical method provides a visual representation of the relationship between variables and offers an intuitive way to understand how the regression line captures the overall trend in the data. However, it is important to note that while this method is useful for simple visualisation, more complex data may require the use of statistical software to calculate the regression line accurately.

Illustration 3.4.1

The following are the advertisement cost (in lakhs) and the sales (in lakhs). Sketch the regression line of sales on advertisements by free hand curve method.

Advertisement:	10	20	30	40	50	60
Sales	300	360	375	325	400	425

Solution



3.4.2.2 Algebraic Method

Different approaches for solving a set of linear equations are referred to as the algebraic method. The regression line is represented algebraically by the regression equation. Since there are two regression lines, there will be two regression equations. The linear model's regression equation has the following structure.

Regression Equation of y on x $\quad y = a + bx$ (1)

Regression Equation of x on y $\quad x = a + by$ (2)

Where x, y are variables, a, b are constants whose values are to be determined.

Here, when independent variable changes, dependent variable also changes in response. From equation (1) we can estimate y for the known values of x and from equation (2) we can estimate x for the known values of y.

3.4.3 Estimating Regression Equation

As mentioned earlier (figure. 2.4.1), under algebraic method, there are mainly two ways to estimate regression equation. They are;

3.4.3.1 Regression Equation Through Normal Equations

Determining the values of the constants included in the causal connection between the two variables is the first step towards identifying the optimum relationship. The least squares principle can be used to accomplish this.

The total of the squares representing the discrepancies between the estimated and observed values should be as small as possible, according to the least squares principle.

In the case of a linear relationship, we can create two equations known as normal equations with a little algebra and differential calculators. These normal equations can be solved to determine the ideal values for the constants 'a' and 'b.'

Regression Equation of y on x

Normal equations of straight line $x = a + by$ are;

$$\sum x = na + b\sum y$$

$$\sum xy = a\sum y + b\sum y^2$$

Regression Equation of x on y

Normal equations of straight line $y = a + bx$ are;

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

After solving the two equations for finding the constants a and b, we can substitute the values in the regression equations.

Illustration 3.4.2

Find the regression equation that fits the following data.

x	2	3	4
y	-1	-2	-3

Solution

Number of observations n = 3

x	y	x^2	xy
2	-1	4	-2
3	-2	9	-6
4	-3	16	-12
$\sum x = 9$	$\sum y = -6$	$\sum x^2 = 29$	$\sum xy = -20$

Normal equations of y on x, $y = a + bx$ are;

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

$$-6 = 3a + 9b \text{ ---------- (1)}$$

$$-20 = 9a + 29b \text{ ----------(2)}$$

To solve these equations, first multiply equation 1 with 3, the result will be;

$$-18 = 9a + 27b \text{ ---------- (3)}$$

Now, subtract equation 3 from equation 2

$$-20 = 9a + 29b$$

$$-18 = 9a + 27b$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$-2 = 2b$$

Therefore, $b = -1$

Substitute $b = -1$ to equation 1, we get

$$-6 = 3a + 9 \times -1$$

$$3a = -6 + 9 = 3$$

$$a = 1$$

Now, after substituting both these values, the regression equation will be;

$$y = 1 + -1x$$

$$y = 1 - x$$

Illustration 3.4.3

Find two regression equations

X	8	4	10	2	6
Y	7	8	5	11	9

Solution

x	y	x^2	y^2	xy
8	7	64	49	56
4	8	16	64	32
10	5	100	25	50
2	11	4	121	22
6	9	36	81	54
$\sum x=30$	$\sum y=40$	$\sum x^2=220$	$\sum y^2=340$	$\sum xy=214$

Regression equation of x on y, $x = a + by$

The normal equations are;

$$\sum x = na + b\sum y$$

$$\sum xy = a\sum y + b\sum y^2$$

Substituting,

$$30 = 5a + 40b \text{----------(1)}$$

$$214 = 40a + 340b \text{----------(2)}$$

Now multiply equation 1 by 8

$$240 = 40a + 320b \text{ ----------(3)}$$

Subtract equation 3 from 2

$$214 = 40a + 340b$$

$$240 = 40a + 320b$$

$$-26 = 20b$$

$$b = \frac{-26}{20}$$

$$b = -1.3$$

Substitute $b = -1.3$ to equation 1, we get

$$30 = 5a + 40 \times -1.3$$

$$30 = 5a - 52$$

$$5a = 52 + 30$$

$$a = \frac{82}{5}$$

$$a = 16.4$$

Regression equation of x on y is

$$x = a + by$$

$$\boldsymbol{x = 16.4 - 1.3y}$$

Regression equation of y on x is $y = a + bx$

The normal equations are;

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

$$40 = 5a + 30b \text{----------(1)}$$

$$214 = 30a + 220b \text{----------(2)}$$

Now multiply equation 1 by 6

$$240 = 30a + 180b \text{----------(3)}$$

Subtract equation 3 from 2

$$214 = 30a + 220b$$

$$240 = 30a + 180b$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$-26 = 40b$$

$$b = \frac{-26}{40}$$

$$b = -0.65$$

Substituting this value, equation 1 will become;

$$40 = 5a + 30 \times -0.65$$

$$40 = 5a - 19.5$$

$$5a = 19.5 + 40$$

$$a = \frac{59.5}{5}$$

$$a = 11.9$$

Regression equation of y on x is

$$y = a + bx$$

$$\boldsymbol{y = 11.9 - 0.65x}$$

Illustration 3.4.4

From the following data, obtain the two regression equations by the method of least square,

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

Solution

x	y	x^2	y^2	$x \times y$
1	2	1	4	2
2	5	4	25	10
3	3	9	9	9
4	8	16	64	32
5	7	25	49	35
Total=15	25	55	151	88

Regression equation y on x is given by $y = a + b\,x$

two normal equations are

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

Substituting the values in the equation, we get;

$$25 = 5a + 15b \quad \ldots..(1)$$

$$88 = 15a + 55b \ldots..(2)$$

Solving the equations,

Eqn (1) × 3 $\qquad 75 = 15a + 45b \ldots$

Eqn (2) $\qquad 88 = 15a + 55b$

$\qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

Subtracting $\qquad -13 = -10b$

$$b = \frac{13}{10} = 1.3$$

Substituting the value of *b* in eqn. (1) we get

$$25 = 5a + 15 \times 1.3$$

$$25 = 5a + 19.5$$

$$5a = 25 - 19.5 = 5.5$$

$$a = \frac{5.5}{5} = 1.1$$

$$\therefore \quad a = 1.1, \quad b = 1.3$$

Regression equation *y* on *x* is given by $y = 1.1 + 1.3\,x$

Regression equation *x* on *y* is given by $x = a + b\,y$

two normal equations are

$$\sum x = na + b\sum y$$

$$\sum xy = a\sum y + b\sum y^2$$

Substituting the values in the equation, we get;

$$15 = 5a + 25b \ldots\ldots\ldots (1)$$

$$88 = 25a + 151b \ldots\ldots\ldots (2)$$

Solving the equations,

Eqn (1) × 5 $\quad 75 = 25a + 125b$

$\quad 88 = 25a + 151b$

……

Subtracting $\quad -13 = -26\,b$

$$b = \frac{13}{26} = 0.5$$

Substituting the value of *b* in eqn. (1) we get

$$15 = 5a + 25 \times 0.5$$

$$15 = 5a + 12.5$$

$$5a = 15 - 12.5 = 2.5$$

$$a = \frac{2.5}{5} = 0.5$$

$$\therefore \quad a = 0.5, \quad b = 0.5$$

Regression equation x on y is given by $x = 0.5 + 0.5\,y$

Illustration 3.4.5

From the following data, obtain the regression equation of y on x.

x	10	6	10	6	8
y	6	2	10	4	8

Solution

x	y	x^2	y^2	$x \times y$
10	6	100	36	60
6	2	36	4	12
10	10	100	100	100
6	4	36	16	24
8	8	64	64	64
Total = 40	30	336	220	260

Regression equation y on x is given by $y = a + bx$

To determine the value of constants "a" and "b", the following two normal equations are to be solved;

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

Substituting the values in the equation,

we get;

$$30 = 5a + 40b \ldots\ldots\ldots (1)$$

$$260 = 40a + 336b \ldots\ldots\ldots (2)$$

Multiplying equation 1 by 8 we get;

$$240 = 40a + 320b$$

$$260 = 40a + 336b$$

……..

Subtracting

$$-20 = -16\,b$$

$$b = \frac{20}{16}$$

$$= 1.25$$

This value of b can be substituted in equation (1), we get the value of a. That is;

$$30 = 5a + 40 \times 1.25$$

$$30 = 5a + 50$$

$$5a = -20$$

$$a = {}^{-}4$$

Substituting the values of "a" and "b" in the regression equation, we get the regression line of y on x, $y = -4 + 1.25\,x$

3.4.3.2 Regression Equation Through Coefficients

Regression equation can also be written as in terms of regression coefficient. Regression coefficients are estimations of some unknown parameters used to characterise the link between a predictor variable and the related response.

Regression coefficient x on y is denoted as b_{xy} and that of y on x is denoted as b_{yx}.

Regression coefficient x on y measures the change in variable x, corresponding to a unit change in variable y. The regression coefficient of x on y can be calculated as;

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

Regression coefficient *y* on *x* measures the change in variable y, corresponding to a unit change in variable *x*. The regression coefficient of y on *x* can be calculated as;

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

Where, *r* = Karl Pearson's Correlation Coefficient

σ_x = Standard deviation of x series

σ_y = Standard deviation of y series

Using this coefficient, we can solve the regression equation in the following manner.

Regression equation of *x* on y is;

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Regression equation of y on *x* is;

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Using deviations from the actual mean of *x* and y series of data, the regression coefficients b_{xy} and b_{yx} can be easily obtained by using the following formula.

$$b_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(y - \bar{y})^2}$$

$$b_{yx} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

Relation between Correlation coefficient and Regression coefficients

We know that $b_{xy} = r\frac{\sigma_x}{\sigma_y}$ and $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

Therefore,

$$b_{xy} \times b_{yx} = r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x}$$

Cancelling the common items, we get $b_{xy} \times b_{yx} = r^2$

Thus, when you know the value of regression coefficients, correlation coefficient can

be calculated using the following equation;

$$r = \sqrt{(b_{xy} \times b_{yx})}$$

Since the value of the correlation coefficient cannot exceed one, one of the regression coefficients must be less than one. In other words, both the regression coefficients cannot be greater than one. Similarly, both the regression coefficients will have the same sign, that is they will be either positive or negative.

Illustration 3.4.6

Find the correlation coefficient when regression coefficients are, $b_{xy}=\frac{9}{20}$ and $b_{yx}=\frac{4}{5}$

Solution

Correlation coefficient $r = \sqrt{b_{xy} \times b_{yx}}$

$$= \sqrt{\frac{9}{20} \times \frac{4}{5}}$$

$$= \sqrt{\frac{36}{100}}$$

$$= 0.6$$

Illustration 3.4.7

Find the correlation coefficient when regression coefficients are, $b_{xy} = \frac{15}{8}$ and $b_{yx} = \frac{3}{5}$

Solution

Correlation coefficient $r = \sqrt{b_{xy} \times b_{yx}}$

$$r = \sqrt{\frac{15}{8} \times \frac{3}{5}}$$

$$= \sqrt{\frac{45}{40}}$$

$$= \sqrt{1.125}$$

$$= 1.06$$

Properties of Regression Coefficients

i. The sign of both regression coefficients will be the same, meaning they can be either positive or negative. Regression coefficients can never be positive and negative at the same time.

ii. Regression coefficients cannot both be more than one, or, to put it another way, one of the regression coefficients must be smaller than one because the correlation coefficient value cannot be greater than one.

iii. The regression coefficient and the correlation coefficient will have the same sign; that is, if the regression coefficient has a positive sign, r will likewise be positive, and if the regression coefficient has a negative sign, r will also be negative.

iv. Correlation coefficient is the geometric mean between regression coefficients

v. Regression coefficients are scale-dependent. This means that if the units of measurement for the independent or dependent variables are changed, the values of the regression coefficients will also change.

vi. Regression coefficients are independent of the origin. This means that if a constant is added to or subtracted from all values of the independent or dependent variables, the values of the regression coefficients will not change.

vii. The arithmetic mean of b_{xy} and b_{yx} is greater than or equal to coefficient of correlation

viii. If $\sigma_y = \sigma_x$, then coefficient correlation equal to regression coefficient, $r = b_{yx} = b_{xy}$. If $r = 0$, then both b_{yx} and b_{xy} will be zero. If $b_{yx} = b_{xy}$ then it is equal to coefficient of correlation i.e., $r = b_{yx} = b_{xy}$.

Calculation of Regression Coefficient

Let us walk through the stepwise procedure for calculating the regression coefficient using the actual mean method with a simple array of data.

Let us say we have the following data pairs:

x	1	2	3	4	5
y	2	3	1	4	5

Step 1: Calculate the means

First, we find the means of both the X values and Y values.

Mean of x (denoted as $\bar{x}$) $= \frac{(1+2+3+4+5)}{5} = 3$

Mean of y (denoted as $\bar{y}$) $= \frac{(2+3+1+4+5)}{5} = 3$

Step 2: Calculate the deviations from mean for x and y

Next, we find the deviations of each data point from their respective means.

For each data point, we subtract the mean from the corresponding value:

$x - \bar{x}$	$y - \bar{y}$
1 – 3 = -2	2 – 3 = -1
2 – 3 = -1	3 – 3 = 0
3 – 3 = 0	1 – 3 = -2
4 – 3 = 1	4 – 3 = 1
5 – 3 = 2	5 – 3 = 2

Step 3: Find the product of deviations

Next, we multiply the deviations of x and y for each data point and sum them up:

$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) \times (y - \bar{y})$
-2	-1	2
-1	0	0
0	-2	0
1	1	1
2	2	4
		$\sum(x - \bar{x})(y - \bar{y}) = 7$

Step 4: Find the sum of squared deviations of x and y

Then, we find the sum of squares of the deviations of x and y

$x - \bar{x}$	$(x - \bar{x})^2$	$y - \bar{y}$	$(y - \bar{y})^2$
1 – 3 = -2	4	2 – 3 = -1	1
2 – 3 = -1	1	3 – 3 = 0	0
3 – 3 = 0	0	1 – 3 = -2	4
4 – 3 = 1	1	4 – 3 = 1	1
5 – 3 = 2	4	5 – 3 = 2	4
	$\sum (x - \bar{x})^2 = 10$		$\sum (y - \bar{y})^2 = 10$

Step 5: Calculate the regression coefficient

Finally, we can calculate the regression coefficient (bxy and byx) using the formula:

$$b_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2}$$

$$b_{yx} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

Since, the values of numerator and denominator for both equations are same, we need to calculate only once.

$$b_{xy} = b_{yx} = \frac{7}{10}$$

$= 0.7$

Illustration 3.4.8

Compute the two regression equations from the following data. If value of $x = 2.5$, what will be the value of y?

x	1	2	3	4	5
y	2	3	5	4	6

Solution

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})\times(y-\bar{y})$
1	2	-2	-2	4	4	4
2	3	-1	-1	1	1	1
3	5	0	1	0	1	0
4	4	1	0	1	0	0
5	6	2	2	4	4	4
Σx =15	Σy =20	0	0	$\Sigma (x-\bar{x})^2$ = 10	$\Sigma (y-\bar{y})^2$= 10	$\Sigma (x-\bar{x}) \times (y-\bar{y})$ = 9

Mean of x $(\bar{x}) = \frac{\Sigma x}{n} = \frac{15}{5} = 3$

Mean of y $(\bar{y}) = \frac{\Sigma y}{n} = \frac{20}{5} = 4$

Regression Coefficient of x on y is;

$$b_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2}$$

Regression Coefficient of y on x is;

$$b_{yx} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

Since, the values of numerator and denominator for both equations are same, we need to calculate only once.

$$b_{xy} = b_{yx} = \frac{9}{10}$$

$$= 0.9$$

Now, using this coefficient let us try to fit the regression equation.

Regression equation of x on y

$$x - \bar{x} = b_{xy}(y-\bar{y})$$

$$x - 3 = 0.9(y - 4)$$

$$x - 3 = 0.9y - 3.6$$

$$x = 0.9y - 3.6 + 3$$

$$x = 0.9y - 0.6$$

Regression equation y on x

$$y - \bar{y} = b_{xy}(x - \bar{x})$$

$$y - 4 = 0.9(x - 3)$$

$$y - 4 = 0.9x - 2.7$$

$$y = 0.9x - 2.7 + 4$$

$$y = 0.9x + 1.3$$

To find out value of y when $x = 2.5$, we can substitute this value to the above regression equation (equation y on x).

$$y = 0.9 \times 2.5 + 1.3$$

$$y = 2.25 + 1.3$$

$$y = 3.55$$

Illustration 3.4.9

Find two regression equation from the following data

	x	y
Mean	65	67
SD	2.5	3.5

Coefficient of correlation = 0.8

Solution

Regression equation of x on y

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$x - 65 = 0.8 \times \frac{2.5}{3.5}(y - 67)$$

$$x - 65 = 0.57(y - 67)$$

$$x - 65 = 0.57y - 38.19$$

$$x = 0.57y - 38.19 + 65$$

$$x = 0.57y + 26.81$$

Regression equation of y on x

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$y - 67 = 0.8 \times \frac{3.5}{2.5}(x - 65)$$

$$y - 67 = 1.12(x - 65)$$

$$y - 67 = 1.12x - 72.8$$

$$y = 1.12x - 72.8 + 67$$

$$y = 1.12x - 5.8$$

Illustration 3.4.10

If the regression equation between the variables x and y are $3x + 2y - 26 = 0$ and $6x + y - 31 = 0$. Find $(\bar{x}, \bar{y})$.

Solution

The given lines are $3x + 2y - 26 = 0$ .....(1)

$6x + y - 31 = 0$.....(2)

.....

Solving the equations,

Eqn (1) $\times$ 2 $6x + 4y = 52$

Eqn (2) $6x + y = 31$

.....

Subtracting $3y = 21$

$$y = 7$$

Substituting the value of y in eqn. (1) we get

$$3x + 2 \times 7 - 26 = 0$$

$$3x = 12$$

$$x = 4$$

Thus $x = 4, y = 7$

Since the regression lines passes through $(\bar{x}, \bar{y})$, $\bar{x}, = 4, \bar{y} = 7$

Illustration 3.4.11

From the following data, obtain the regression equation of x and y and y on x.

x	10	6	10	6	8
y	6	2	10	4	8

Solution

x	y	$x-8$	$(x-8)^2$	$y-6$	$(y-6)^2$	$(x-8)(y-6)$
10	6	2	4	0	0	0
6	2	-2	4	-4	16	8
10	10	2	4	4	16	8
6	4	-2	4	-2	4	4
8	8	0	0	2	4	0
Total 40	30	0	16	0	40	20

$$n = 5, \ \bar{x} = \frac{\Sigma x}{n} = \frac{40}{5} = 8 \ , \quad \bar{y} = \frac{\Sigma y}{n} = \frac{30}{5} = 6$$

Regression equation y on x is given by the equation

$$(y - \bar{y}\ ) = b_{yx}\ (x - \bar{x})$$

$$b_{yx} = \frac{\Sigma(x - \bar{x})\ (y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

$$= \frac{20}{16} = 1.25$$

$$(y - 6\ ) = 1.25\ (x - 8)$$

$$y = 1.25x - 4$$

Similarly, regression equation x on y is given by the formula

$$(x - \bar{x}\ ) = b_{xy}\ (y - \bar{y})$$

$$b_{xy} = \frac{\Sigma(x - \bar{x})\ (y - \bar{y})}{\Sigma(y - \bar{y})^2}$$

$$= \frac{20}{40} = 0.5$$

$$(x - 8\ ) = 0.5\ (y - 6)$$

$$x = 0.5\ y + 5$$

Recap

- ◊ Regression line - visual representation of how a change in one variable is associated with a change in another
- ◊ Graphical method of regression- through scatter diagram
- ◊ Algebraic method of regression- though normal equations and regression coefficient
- ◊ Regression coefficients - estimations of some unknown parameters used to characterise the link between a predictor variable and the related response.
- ◊ Regression coefficient x on y is denoted as b_{xy} and that of y on x is denoted as b_{yx}
- ◊ Relation between regression and correlation is - $r = \sqrt{(bxy \times byx)}$

Objective Questions

1. What does the slope of a regression line represent?
2. How can you interpret the constant value 'a' in regression?
3. What does a scatter plot visually represent in the graphical method of regression?
4. How is the regression line depicted on a scatter plot?
5. What is the primary objective of the algebraic method of regression?
6. State the simple linear regression equation
7. How is the coefficient interpreted in the context of a regression equation?
8. What does a regression coefficient of zero signify in simple linear regression?
9. State the relationship between regression coefficient and correlation mathematically.

Answers

1. Rate of change or the strength and direction of the relationship.
2. The predicted value of the dependent variable when the independent variable is zero.
3. The relationship between the independent and dependent variables.
4. Straight line that minimises the sum of squared differences between observed and predicted values.
5. To determine the equation of the regression line that best fits the data.
6. $y = a + bx$
7. It represents the proportion of variability in the dependent variable explained by the independent variable.
8. No linear relationship between the independent and dependent variables.
9. $r = \sqrt{byx \times bxy}$

Self-Assessment Questions

1. What is meant by regression?
2. Mention the properties of regression coefficient
3. List the properties of regression lines
4. If the regression coefficients are -4/7 and -7/9, then find the value of correlation coefficient.
5. If the correlation coefficient between two variables X and Y is 0.4 and the regression coefficient of X on Y is 0.2. Find the regression coefficient of Y on X.
6. Prove that $r = \sqrt{byx \times bxy}$
7. Write your inference when the regression lines are perpendicular?
8. Mention the differences between correlation and regression analysis

Assignments

1. You have data on advertising expense (in dollars) and corresponding sales (in units) for a product over several months. Use this data to find the regression coefficients and create a predictive model for sales based on advertising expense.

Advertising	200	400	600	800	1000
Sales	1200	1600	2100	2500	2800

2. You want to understand the relationship between the number of hours students spend studying and their exam scores. Collect data on study hours and exam scores of your classmates and determine the regression coefficients to model this relationship.
3. The lines of regression of y on x and x on y are respectively y = x + 5 and 16x = 9y - 94. Find the variance of x if the variance of y is 16. Also find the covariance of x and y.

4. For a given set of bivariate data, the following results were obtained:

$\bar{x} = 53.2, \bar{y} = 27.9$, $b_{yx} = -1.5$ and $b_{xy} = -0.2$

Find the most probable value of y when $x = 60$. Also find 'r'.

Suggested Reading

1. Sancheti, & Kapoor, V. K. (2014). *Business mathematics*. Sultan Chand & Sons.
2. Wikes, F. M. (1998). *Mathematics for business, finance and economics*. Thomson Learning.
3. Vittal, P. R. (2012). *Business maths & statistics*. Margham Publications.
4. Dixon, W. S., & Massey, F. J. (1951). *Introduction to statistical analysis*. McGraw Hill Book Company, Inc.
5. Fisher, R. A., & Yates, F. (1948). *Statistical tables for biological, agricultural and medical research* (3rd ed.). New York: Hafner Publishing Company.

Reference

1. Gupta, S. P. (1994). *Statistical Methods*, Sultan Chand & Sons, New Delhi, pp. *E10*, 1-61.
2. Gupta, C. B., & Gupta, V. (2009). *Introduction to statistical methods*. Vikas publishing House Pvt. LTD.
3. Goel, A., & Goel, A. *Mathematics and Statistics*, Taxmann Allied Services Pvt.
4. Kothari, C. R. (2013). *Quantitative Techniques*. Vikas Publishing House.

BLOCK - 04
Time Series Analysis

Unit - 1
Introduction to Time Series Analysis

Learning Outcomes

At the conclusion of this unit, the learner will be able to:

- ◊ comprehend the meaning and essence of time series
- ◊ appreciate the significance of time series in data analysis
- ◊ acquire the ability to identify and analyse components within a time series, such as trends, seasonality, and cycles.

Prerequisite

Imagine you are a retail manager closely monitoring monthly sales data for your store. Initially, everything seems steady, but suddenly there is an unexpected surge in sales during the holiday season. As the numbers continue to fluctuate, you find it challenging to pinpoint whether these changes are part of a temporary trend or indicative of a more significant pattern. For finding solution for this issue, you might need to make an effort to understand and interpret your sales data over time. By applying time series techniques, you can identify seasonal spikes, understand customer behaviour over time, and predict future sales trends. This unit discusses the meaning and significance of Time Series as well as several components of Time Series Analysis in detail.

Keywords

Trend, Seasonal trend, Cyclical component, Secular Trend, Irregular Component, Time series

Discussion

There is a claim that our stock market, the Sensex, will reach 280,000 in the year 2050 from its current 52,000 level in 2022.That is, over the long run, the return from all investments is outperformed by the equity market. This claim relies on the analysis of historic data on stock price movements. This means looking at how stock market values change over time. By studying past patterns, analysts predict future trends. In simple terms, it is like analysing the stock market's historical movement to make an educated guess about its future. So, the idea is that, over the long run, investments in the stock market tend to perform well, and this prediction is based on understanding how the market has behaved in the past through the lens of time series analysis. Now let us familiarise the meaning of Time Series.

4.1.1 Time Series - Meaning

Time series is an arrangement of statistical data in chronological order. It involves a set of values recorded at equal time intervals. It is a forecasting method based on changes of data that have taken place in the past. Forecasting means predicting the future values of a variable based on historical values of the same or other variables. If the forecast is based simply on past values of the variable itself, it is called time series forecasting.

Imagine you are tracking the monthly sales of a popular ice cream shop over the past year. Your time series data would look something like this:

Month	Sales
Jan	5000
Feb	5200
Mar	6000
Apr	8500
May	10000
Jun	12500
Jul	15000
Aug	14000
Sep	11200
Oct	9000
Nov	7500
Dec	6200

In this example, each row represents a specific month, and the corresponding sales figures that form a time series. Time series analysis of this data may help forecast future trends and identify seasonal patterns, such as higher sales during the summer.

Definition: According to Mooris Hamburg, "A time series is a set of statistical observations arranged in chronological order."

4.1.2 Utility of Time Series Analysis

The utility of time series analysis lies in its ability to reveal patterns, trends, and insights within chronological data. By examining how values change over time, it helps in making predictions, understanding fluctuations, and identifying potential future scenarios. In fields like finance, economics, and business, time series analysis guides decision-making, enabling businesses to forecast market trends, optimise resource allocation, and respond effectively to changes. It serves as a valuable tool for strategic planning, risk management, and extracting meaningful information from temporal data, providing a clearer understanding of past, present, and future trends. The following points highlights the utilities of time series analysis in general.

i. **Forecasting Future Trends:** Time series analysis allows us to predict future values based on past patterns, aiding in proactive decision-making and planning.

ii. **Pattern Identification:** It helps recognise recurring trends and irregularities, providing valuable insights into the behaviour of data over time.

iii. **Strategic Decision Support:** Businesses use time series analysis to strategically allocate resources, optimise operations, and make informed decisions for improved performance.

iv. **Risk Management:** By understanding trends and fluctuations, organisations can effectively manage and mitigate risks associated with market dynamics or economic changes.

v. **Resource Optimisation:** Time series analysis guides resource planning by identifying peak periods, enhancing efficiency, and aligning operations with demand, contributing to overall operational optimisation.

4.1.3 Components of Time Series

In time series, think of the data like a story. The components are the key characters in this story. First, there is the main character, the trend, showing the overall direction: up, down, or steady. Then, the supportive friend, seasonality, bringing regular and predictable patterns like holidays or weekends. There is also a wildcard, the cyclical component, showing longer term ups and downs that are not as predictable. Lastly, the random part, or noise, adds unpredictability. Understanding these components is

like knowing the characters in our story, helping us make sense of the data's journey through time and make smarter predictions for the future.

Let us familiarise each of these components of time series thoroughly before moving forward to the actual analysis of data using time series.

There are mainly four components to time series;

- ◊ Secular Trend
- ◊ Cyclic Component
- ◊ Seasonal Trend
- ◊ Irregular Component (Erratic)

This may be classified based on time i.e., long term and short term as follows:

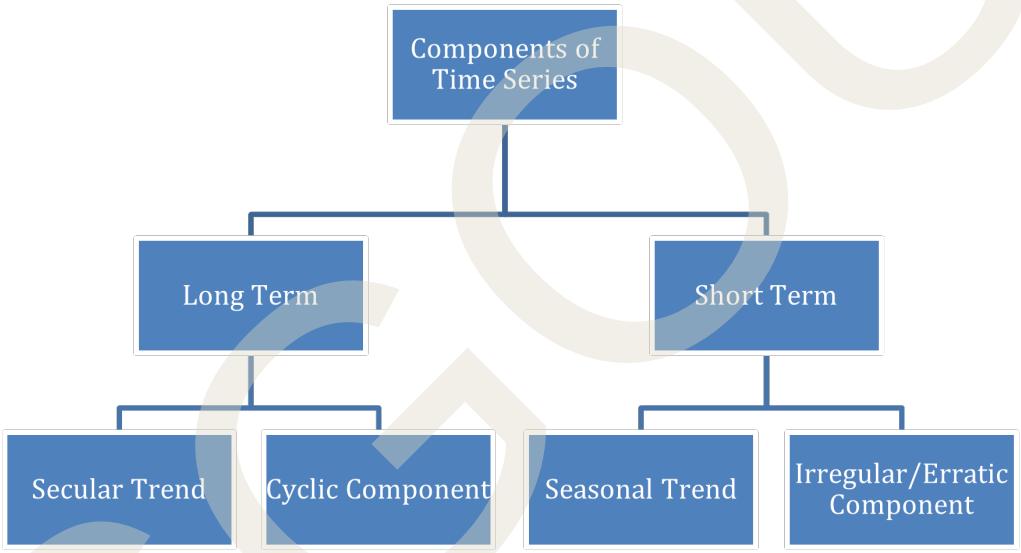


Figure 4.1.1 Components of Time Series

4.1.3.1 Secular Trend

It is a long term movement in Time series. The tendency of time series to increase or decrease in a long period of time is called the secular trend or simply Trend. A secular trend is like the main storyline in our data. It shows the big, long term direction; whether things are consistently going up, staying the same, or decreasing. It is the key theme, helping us understand the fundamental, lasting changes happening over a long period. So, when we talk about a secular trend, we are focusing on the major, enduring pattern behind the ups and downs in our data. It is like the main plot in a movie, guiding our understanding of the overall journey.

Imagine tracking the population of a city over several decades. If the population consistently increases year after year due to factors like urbanization and job opportunities, that is a secular trend. The long term upward movement in population signifies a fundamental and lasting change. Even though there might be short term fluctuations due to events like economic downturns or development projects, the overarching growth represents the city's secular trend. It is like the main plot of the city's demographic story, helping policymakers and planners make decisions for the future based on the enduring pattern of population increase. The following figure will give you the idea of secular trend in population data of India.

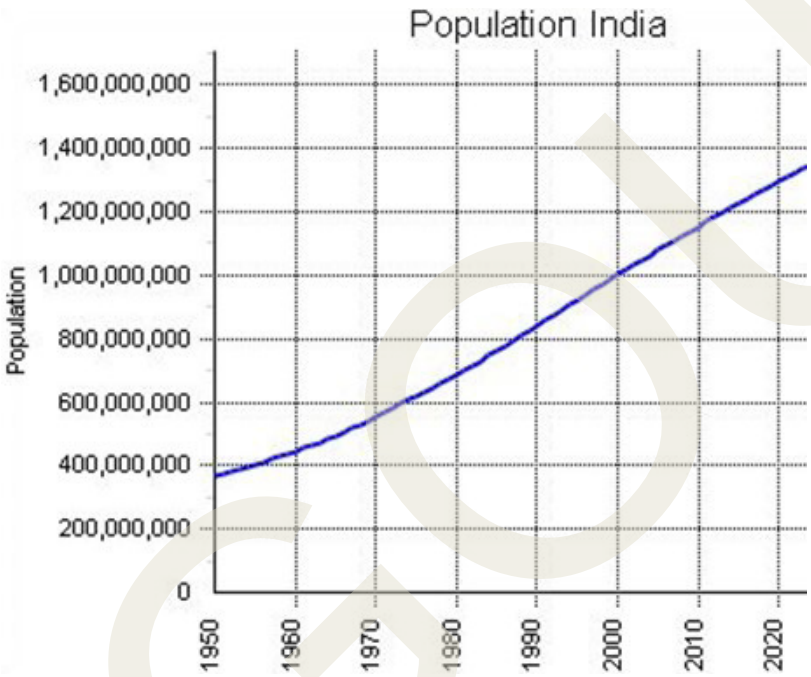


Figure 4.1.2 Secular trend in population of India

4.1.3.2 Seasonal Trend

It is an observable pattern in the data in less than a year. That is identical patterns or regularity over equal interval of time. A seasonal trend is the expected pattern in our data that happens regularly, like a favorite ice cream shop selling more in the summertime. It is like noticing the repeating behaviour like certain things go up or down predictably at specific times, making it easy to anticipate when they will happen again.

Think of coffee sales spiking every winter as people crave warmth. That is a seasonal trend. Just like clockwork, it happens predictably during colder months, showing a pattern in the data (refer the figure below). Businesses use this insight to stock up on coffee beans and meet the expected surge in demand.

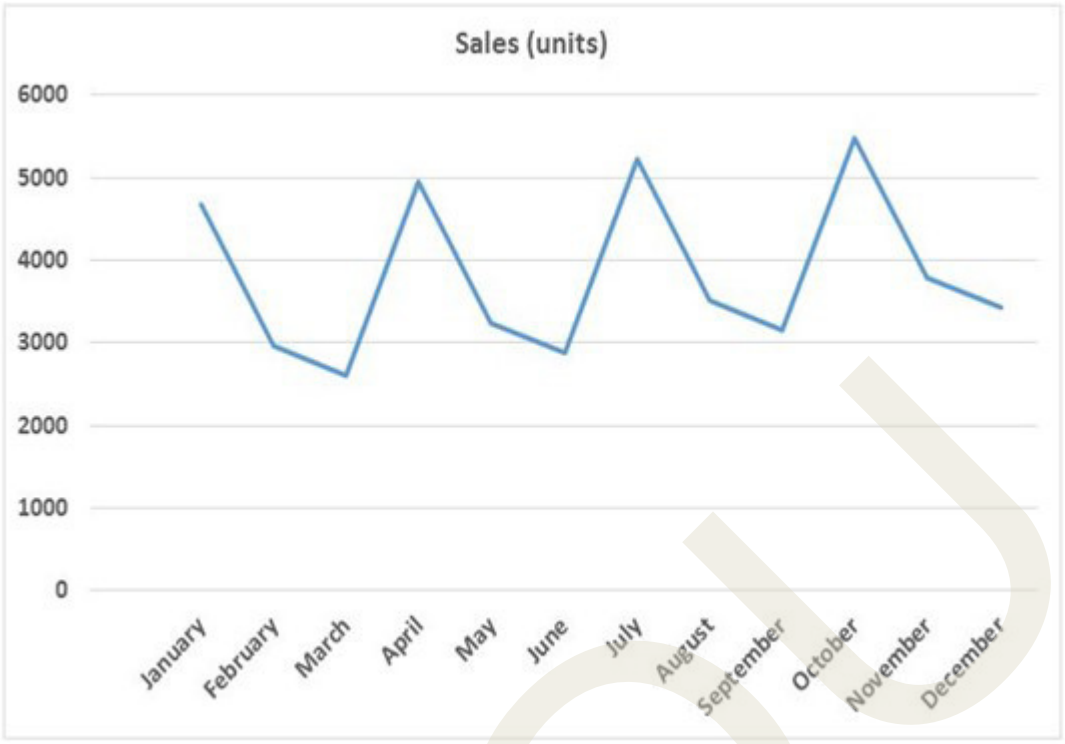


Figure 4.1.3 Seasonal Trend in Time Series

4.1.3.3 Cyclic Component

The cyclic component in a time series is like a wave in our data story. It shows longer-term ups and downs, not as predictable as seasons. Think of it as a rollercoaster ride, with periods of growth followed by decline, but the overall pattern is not fixed. Unlike seasons, cycles might take years and are not linked to specific times. Understanding this component helps us recognise broader economic or business cycles, providing insights into the more extended rolling movements in our data.

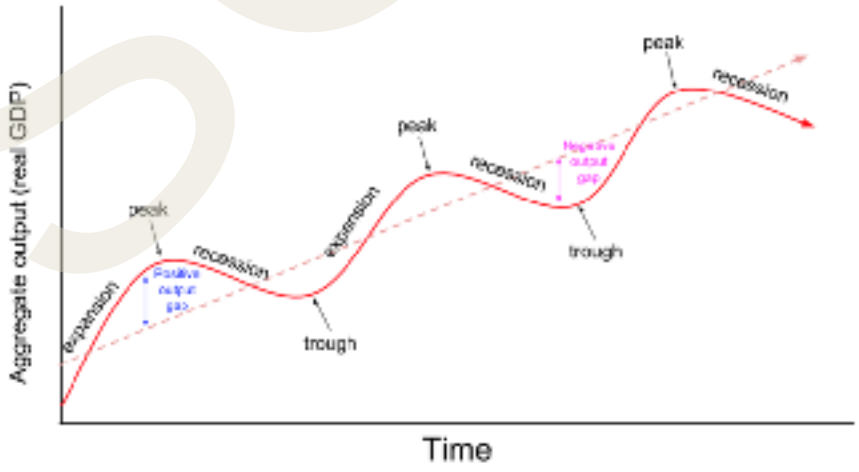


Figure 4.1.4 Cyclic Component in Time Series

Consider the housing market. During economic booms, property values often rise, reflecting a positive cycle. Conversely, during economic downturns, property values might decrease, marking a downturn in the cycle. This cyclical pattern is not as predictable as the change in seasons but can be observed over a more extended period.

4.1.3.4 Irregular Component

The irregular component in a time series is like the unpredictable twists in our data tale. It represents random fluctuations or unexpected events that do not follow a specific pattern. Imagine a sudden spike in ice cream sales due to an unexpected heatwave. This component is essential for acknowledging the unanticipated factors influencing data, making our understanding more realistic and allowing businesses to adapt to sudden, irregular changes in their operations or markets.

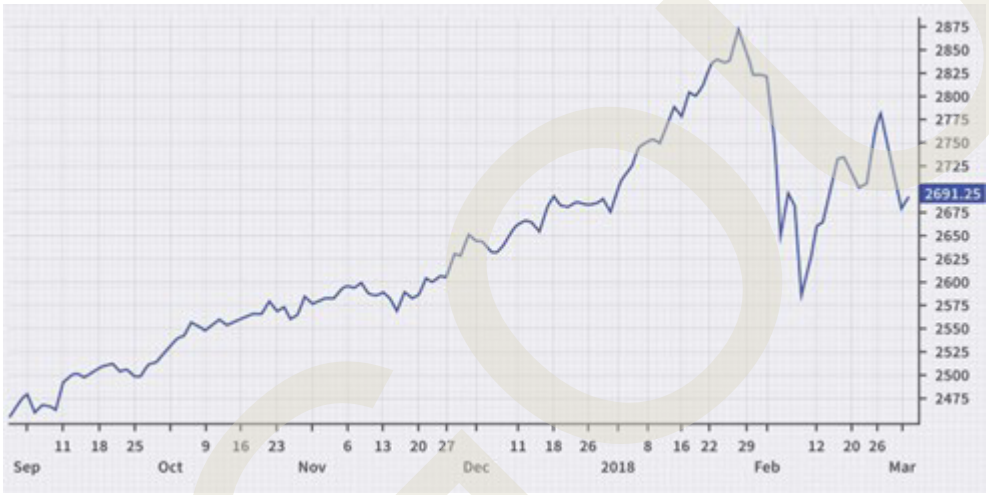


Figure 4.1.5 Irregular Component in Time Series

As you can observe from the above figure, in the context of a time series with an increasing trend, an irregular movement refers to a sudden and unexpected decrease in the observed values (see the decrease in February; figure 4.1.5). This irregular movement is often characterized by a deviation from the established upward pattern, disrupting the overall trend that had been evident over a specific period.

Irregular movements can be caused by various unforeseen factors such as economic downturns, sudden changes in consumer behavior, unexpected events, or external shocks to the market. Identifying and understanding irregular movements is crucial in time series analysis because it allows analysts and decision-makers to investigate the root causes of these disruptions and adjust strategies accordingly. It also highlights the importance of considering external factors that may influence trends in a time series, even if they were initially moving in a consistent direction.

Recap

- ◊ Time Series Analysis - arrangement of statistical data in chronological order
- ◊ Utility of time series- predict future trends and patterns, informed decision making
- ◊ Secular trend - long term movement in Time series
- ◊ Seasonal trend - identical patterns or regularity over equal interval of time
- ◊ Cyclic variations - not similar pattern over time.
- ◊ Irregular component - unpredictable twists in our data

Objective Questions

1. What is Time Series?
2. Why is Time Series important?
3. What are the components of Time Series?
4. Which component represents long term movement?
5. Which component represents short term movement?
6. What does Seasonal trend capture?
7. What is the nature of irregular movement?
8. What does Time Series analysis help predict?

Answers

1. Sequential data points observed over time.
2. To analyse trends and make predictions.
3. Secular, Seasonal, Cyclical, Irregular movements

4. Secular and Cyclical movements
5. Seasonal and Irregular movement
6. Regular, repeating patterns.
7. Unpredictability
8. Future trends and patterns.

Self-Assessment Questions

1. What is the primary focus of a time series analysis?
2. How would you define a time series in the context of data analysis?
3. What is the utility of studying time series data for businesses and analysts?
4. Can you name and briefly explain the main components of a time series?
5. In what practical scenarios can understanding time series be beneficial for decision-making?
6. Elaborate a scenario where you can observe cyclical movement.

Assignments

1. Can you provide a real-world example where a time series analysis would be beneficial, and how does it differ from other data analysis methods?
2. Discuss a scenario in which time series analysis could be applied to enhance decision-making in a business or economic context. How might this analysis contribute to better predictions or strategies?
3. Break down the components of a time series for a given dataset, highlighting the trend, seasonality, and any irregularities or random fluctuations. How does understanding these components aid in interpreting the data?
4. Imagine you have a dataset representing monthly sales figures for a retail store over several years. How would you identify and analyse the trend and seasonality in the data? How might this information be useful for the store's management?

5. Discuss potential challenges or limitations in working with time series data. How might outliers or missing data impact the accuracy of time series analysis, and what strategies can be employed to address such challenges?

Suggested Reading

1. Gupta, S. P. (1994). *Statistical Methods*, Sultan Chand & Sons, New Delhi, pp. *E10*, 1-61.
2. Gupta, C. B., & Gupta, V. (2009). *Introduction to statistical methods*. Vikas publishing House Pvt. LTD.
3. Goel, A., & Goel, A. *Mathematics and Statistics*, Taxmann Allied Services Pvt.
4. Kothari, C. R. (2013). *Quantitative Techniques*. Vikas Publishing House.

Reference

1. Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). John Wiley & Sons.
2. Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting* (3rd ed.). Springer.
3. Chatfield, C. (2004). *The analysis of time series: An introduction* (6th ed.). Chapman and Hall/CRC.
4. Cryer, J. D., & Chan, K. S. (2008). *Time series analysis: With applications in R* (2nd ed.). Springer.
5. Enders, W. (2015). *Applied econometric time series* (4th ed.). John Wiley & Sons.
6. Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.
7. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts.

Unit - 2

Measurement of Trend

Learning Outcomes

At the conclusion of this unit, the learner will be able to;

- ◊ grasp the concept of moving averages, calculate simple moving averages, and comprehend their role in smoothing time series data to identify trends
- ◊ become proficient in using the method of least squares to fit a linear trend line to time series data
- ◊ evaluate the results of moving averages and the method of least squares, making informed decisions on their suitability for different datasets

Prerequisite

Suppose you run a small business selling handmade crafts, and you meticulously track your monthly sales over several years. As you review your sales data, you notice fluctuations in the numbers, with some months showing higher sales and others experiencing a dip. This variability in sales might seem like random noise at first glance.

Here is where the measurement of trend becomes necessary: identifying the underlying pattern or trend within these sales figures. By measuring the trend, you aim to uncover the general direction in which your business is moving over time. For instance, you might observe that, despite the monthly ups and downs, there is a gradual increase in sales each year, indicating a positive trend. This insight becomes crucial for several reasons. First, it helps you make informed business decisions. If you can identify that your sales are consistently increasing, you might decide to invest more in marketing or consider expanding your product line. On the other hand, if you notice a declining trend, you may need to reassess your strategies and adjust reverse the pattern.

This Unit will take you through different methods to study the trend in time series data.

Keywords

Moving Average, Least Squares, Trend Line, Measurement of trend, Application of Time Series in Business and Management

Discussion

4.2.1 Measurement of Trend

Sometimes, we can identify the pattern or trend in a given series of data overtime, just by observing it. But that may not be possible when the data become large and complex in nature. Identifying the intricate patterns and trends visible in all types of data can be systematically done using several methods. Here, we shall discuss only two among the available methods to analyse trend in time series which are used commonly in business and economic scenarios.

They are;

I. Moving Average Method and

II. Method of Least Squares

Moving Average is an example of graphical method of finding out trend in time series and method of lease squares is an example of quantitative or mathematical method of trend analysis.

4.2.2 Moving Average Method

The moving average method is a statistical technique used to smooth out fluctuations in time series data, revealing underlying trends. The process involves calculating averages for multiple set of consecutive data points for a specific duration (like 3 years taking average of observations for year 1 to 3, year 2 to 4, year 3 to 5, and so on) from a time series data, or 'moving' averages, which helps in identifying patterns and trends.

This method smoothens the data, making trends more apparent and aiding in decision making based on a clearer representation of the underlying pattern. This method is used when trend is secular and require a high degree of accuracy. That is trend values obtained by this method are more accurate.

To put it in simple words, the moving average may be for three, four, five, six, seven years and so on. Suppose moving average is to be calculated for three years, we will take the average for first three years and will place it against the middle year of three. Similarly moving average is to be calculated for five years, we will take the average for first five years and will place it against the middle year of five. Now leaving the first

year and adding the sixth year, take the average of the five years. Place this average against the middle of those five years. This way taking the average of after leaving the first year and taking one next year i.e., if the observations are a, b, c, d, e, f and so on, then three year moving averages are (a+ b + c) /3, (b +c + d) /3, (c +d +e) /3 etc.

Five yearly moving averages are (a + b + c + d + e) /5, (b + c + d + e + f) /5 etc.

This will be much clearer once you familiarise the procedure for calculating moving averages.

4.2.2.1 Calculation of Moving Average

Generally, moving averages are calculated as 3 year moving average, 5 year moving average, etc. Let us go through the step-by-step procedure for calculating a simple moving average using a small array of data. For this example, let us say we have the following data representing monthly sales over six months:

Jan	Feb	Mar	Apr	May	June
25	30	28	35	40	32

Step 1:

Decide on the period for your moving average. For simplicity, let us use a 3 month moving average.

Step 2:

Create subsets of the data based on the chosen period. For a 3-month moving average, the subsets would be:

Jan	25	
Feb	30	(25+30+28)
Mar	28	(30+28+35)
Apr	35	(28+35+40)
May	40	(35+40+32)
June	32	

Step 3:

Find the average of each subset. For the given subsets, the averages are as follows:

Jan	25		
Feb	30	(25+30+28)	(25+30+28) /3 = 27.7
Mar	28	(30+28+35)	(30+28+35)/3 = 31
Apr	35	(28+35+40)	(28+35+40)/3 = 34.3
May	40	(35+40+32)	(35+40+32)/3 = 35.7
June	32		

Step 4: Plot these averages instead of original data to observe the underlying trend in the data

Illustration 4.2.1

Give three yearly and five yearly moving average for the following series

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Sales	1	2	3	5	6	7	9	10	11	15	20

Solution

3 Year Moving Average

Year	Sales	3 yearly moving total	3 yearly moving average
2010	1		
2011	2	(1+2+3)=6	6/3 = 2
2012	3	(2+3+5)=10	10/3 = 3.33
2013	5	(3+5+6)=14	14/3 = 4.67
2014	6	(5+6+7)=18	18/3 = 6
2015	7	(6+7+9)=22	22/3 = 7.33
2016	9	(7+9+10)=26	26/3 = 8.67
2017	10	(9+10+11)=30	30/3 = 10
2018	11	(10+11+15)=36	36/3 = 12
2019	15	(11+15+20)=46	46/3 = 15.33
2020	20		

5 Year Moving Average

Year	Sales	5 yearly moving total	5 yearly moving average
2010	1		
2011	2		
2012	3	(1+2+3+5+6) = 17	17/5 = 3.4
2013	5	(2+3+5+6+7) = 23	23/5 = 4.6
2014	6	(3+5+6+7+9) = 30	30/5 = 6
2015	7	(5+6+7+9+10) = 37	37/5 = 7.4
2016	9	(6+7+9+10+11) = 43	43/5 = 8.6
2017	10	(7+9+10+11+15) = 52	52/5 = 10.4
2018	11	(9+10+11+15+20) = 65	65/5 = 13
2019	15		
2020	20		

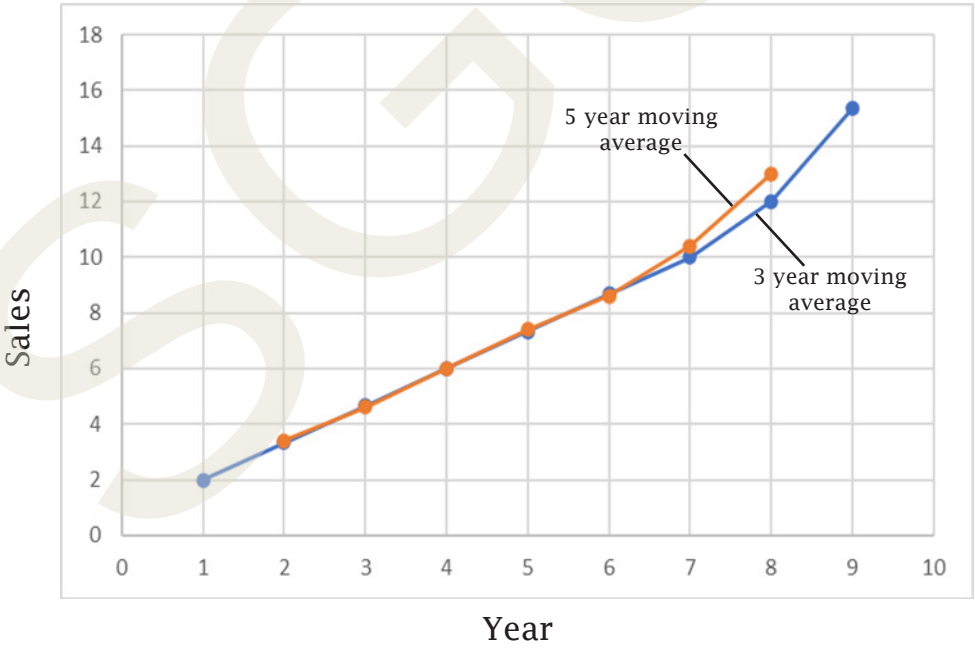


Illustration 4.2.2

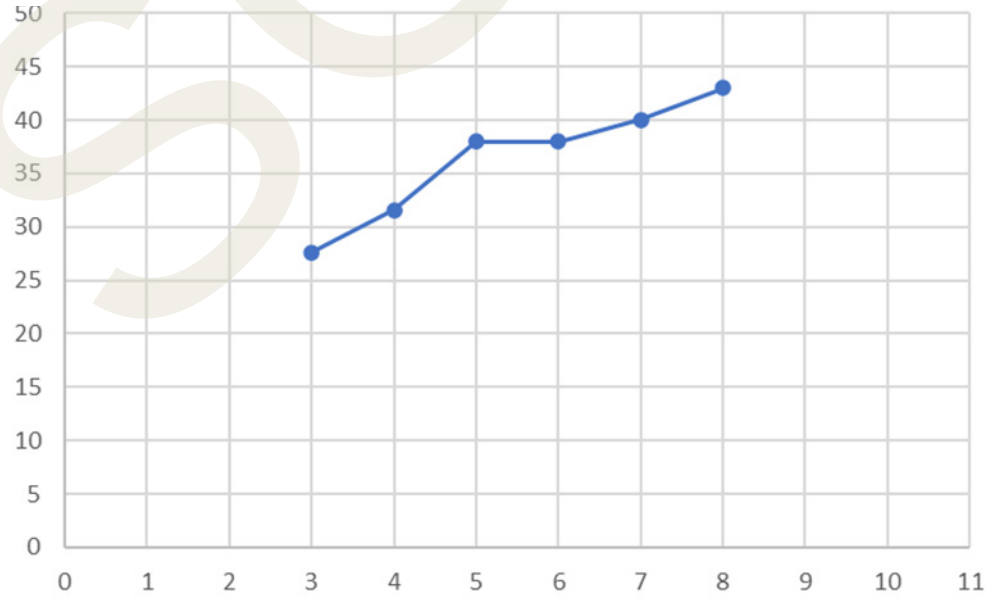
Find the underlying trend in the Sales data for 10 years using 5 year moving average.

Year	1	2	3	4	5	6	7	8	9	10
Yearly Sales	20	25	30	28	35	40	45	42	38	50

Solution

5 Year Moving Average

Year	Yearly Sales	5 yearly moving sum	5 yearly moving average
1	20		
2	25		
3	30	(20+25+30+28+35) = 138	138/5 = 27.6
4	28	(25+30+28+35+40) = 158	158/5 = 31.6
5	35	(30+28+35+40+45) = 178	178/5 = 38
6	40	(28+35+40+45+42) = 190	190/5 = 38
7	45	(35+40+45+42+38) = 200	200/5 = 40
8	42	(40+45+42+38+50) = 215	215/5 = 43
9	38		
10	50		



4.2.2.2 Method for calculating moving average for even number of years

In case, the number of years is even, for example to calculate four years moving average, calculate average of first four years. Put this value in between second and third years i.e., middle of first four years). Now, leave the first year, calculate the average of next four years and place it in the middle of these four years and so on. Then, calculate the average of the two moving averages already calculated, taking first and second, second and third etc. This average is called moving average centred. We will place the first average centred against the middle of the two moving averages. This average will be against third year of the original data. In this way we calculate averages centred for the other years. The centred moving averages will be the trend values.

Illustration 4.2.3

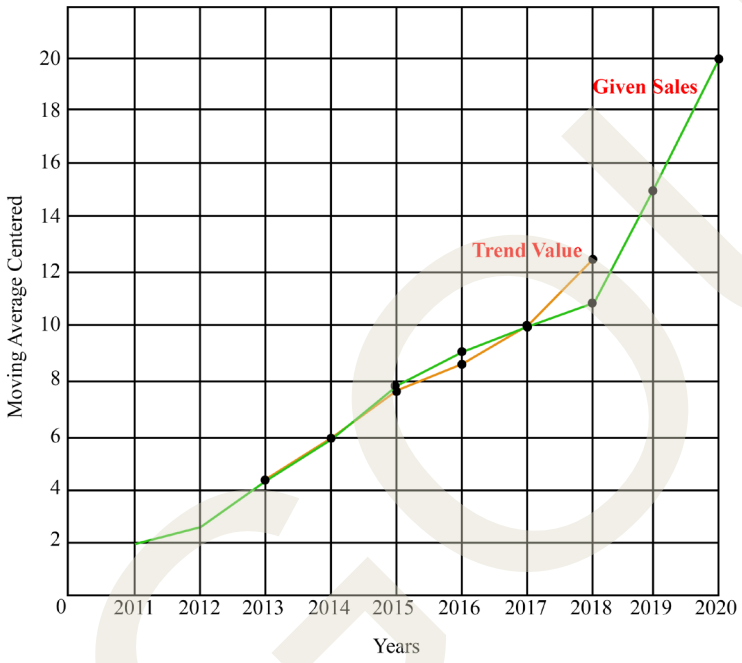
Give 4 yearly moving average for the following series. Also, plot the given values and trend values on a graph.

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Sales	2	3	5	6	7	9	10	11	15	20

Solution

Year	Sales	Total Sales of 4 years	4 yearly moving average	4 yearly moving average centered
2011	2			
2012	3			
		(2+3+5+6) = 16	16/4 = 4	
2013	5			(4+5.25)/2 = 4.625
		(3+5+6+7) = 21	21/4 = 5.25	
2014	6			(5.25+6.75)/2 = 6
		(5+6+7+9) = 27	27/4 = 6.75	
2015	7			(6.75+8)/2 = 7.375
		(6+7+9+10) = 32	32/4 = 8	
2016	9			(8+9.25/2) = 8.625
		7+9+10+11 = 37	37/4 = 9.25	
2017	10			(9.25+11.25)/2 = 10.25

		(9+10+11+15) = 45	45/4 = 11.25	
2018	11			(11.25+14)/2 = 12.625
		10+11+15+20 = 56	56/4 = 14	
2019	15			
2020	20			



4.2.3 Method of Least Squares

The method of least squares is a mathematical technique used for time series trend analysis. It aims to find the best-fitting straight line (regression line) through a set of data points by minimising the sum of the squares of the vertical distances (residuals) between the data points and the line. The method of least squares provides a systematic way to estimate the parameters of a linear trend in time series data. It is widely used for forecasting and understanding the overall trend in a dataset.

The principle of least squares, a fundamental concept in statistics, minimises the sum of squared deviations between observed values and a trend line, ensuring the best possible fit to the data.

The technique that involves obtaining a mathematical curve that optimally represents the dataset is known as curve fitting. By minimising the squared differences between actual and predicted values, least squares provide a precise method for estimating parameters in various fields, from time series analysis to regression modeling. It enhances accuracy in predicting trends, facilitating robust decision making and pattern

recognition in scientific, economic, and research endeavors.

The procedure for this method of finding trend is similar to that of regression analysis (Refer Block 3, Unit 1). Let us look in to some illustration to make it clearer.

Illustration 4.2.4

Find out a straight that fits the following data.

x	2	3	4
y	-1	-2	-3

Solution

Number of observations n = 3

x	y	x^2	xy
2	-1	4	-2
3	-2	9	-6
4	-3	16	-12
$\sum x = 9$	$\sum y = -6$	$\sum x^2 = 29$	$\sum xy = -20$

Normal equations of straight line $y = a + bx$ are

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x2$$

$$-6 = 3a + 9b \text{ ------(1)}$$

$$-20 = 9a + 29b \text{ --------(2)}$$

Equation (1) x 3 $\quad -18 = 9a + 27b$

$$-20 = 9a + 29b$$

.....

Subtracting $\quad 2 = -2b$

Therefore, $\quad b = -1$

Substituting the value of b in eqn. (1) we get

$$-6 = 3a + 9 \times -1$$

$$3a = -6 + 9$$

$$3a = 3$$

$$a = 1$$

Equation of the straight line is $y = a + bx$, we get

$$y = 1 - x$$

i.e., line of best fit is $x + y = 1$

Note:

The algebraic computation can be simplified to a great extent by shifting the origin in the time variable to a new variable x in such a way that we always get $\sum x = 0$. It is applicable if the values of time variable t are equidistant at an interval h.

If the number of items in the time series is odd, then the transformation is

$$x = \frac{\text{t - middle value}}{h}$$

If the number of items in the time series is even, then the transformation is

$$x = \frac{t - Arithmatic\ Mean\ of\ 2\ middle\ value}{h/2}$$

Since $\sum x = 0$, $a = \frac{\sum y}{n}$ and $b = \frac{\sum xy}{\sum x^2}$

Illustration 4.2.5

The sales (in tonnes) of a firm for the years 2015 to 2021 are given below. Estimate the sales for 2024 using the trend.

Years	2015	2016	2017	2018	2019	2020	2021
Sales (in tonnes)	5	6	8	6	10	10	11

Solution

$n = 7$, shift the origin t to x by $x = \frac{t-2018}{1} = t - 2018$

Year t	Sales Y	x (t - 2018)	x^2	x^y	Trend $y_c = 8+1x$
2015	5	2015-2018 = -3	9	-15	8+-3 = 5
2016	6	2016-2018 = -2	4	-12	8+-2 = 6
2017	8	2017-2018 = -1	1	-8	8+-1 = 7
2018	6	2018-2018 = 0	0	0	8+0 = 8
2019	10	2019-2018 = 1	1	10	8+1 = 9
2020	10	2020-2018 = 2	4	20	8+2 = 10
2021	11	2021 - 2018 = 3	9	33	8+3 = 11
n = 7	Σy = 56	Σx = 0	$\Sigma x^2 = 28$	Σxy = 28	56

The straight line trend of y on x is $y = a+bx$

Since Σx = 0,

$$a = \frac{\Sigma y}{n}$$

$$= \frac{56}{7}$$

$$= 8$$

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

$$= \frac{28}{28} = 1$$

The straight line is

$$y_c = 8+1x$$

Estimation of sales in the year t = 2024, the x value is 2024 - 2018 = 6

y(2024) = 8+(1x6) =14

Illustration 4.2.6

Below given are the figures of production (in thousand tons) of a sugar factory :

Year	1999	2000	2001	2002	2003	2004	2005
Production	77	88	94	85	91	98	90

Fit a straight line by the method of 'least squares' and show the trend values.

What is the monthly increase in production?

Solution

$n = 7$, shift the origin t to x by $x = \frac{t-2002}{1} = t - 2002$

Year t	Production y	x t-2002	x^2	xy	Trend $Y_c = 8 + 2x$
1999	77	-3	9	-231	2
2000	88	-2	4	-176	4
2001	94	-1	1	-94	6
2002	85	0	0	0	8
2003	91	1	1	91	10
2004	98	2	4	198	12
2005	90	3	9	270	14
n = 7	$\sum y = 56$	$\sum x = 0$	$\sum x^2 = 28$	$\sum xy = 56$	56

The straight line trend of y on x is $y = a + bx$. Since $\sum x = 0$

$$a = \frac{\sum y}{n}$$

$$= \frac{56}{7}$$

$$= 8$$

$$b = \frac{\sum xy}{\sum x^2}$$

$$= \frac{56}{28}$$

$$= 2$$

The straight line is

$$Y_c = 8 + 2x$$

From the equation the trend values increase by a constant amount 'b' units every year. Thus, the yearly increase in production is 'b' units, i.e., 2 × 1000 = 2000 tons. Hence, the monthly increase in production = 2000/ 12 = 166·67 tons

Illustration 4.2.7

Fit a straight line trend to the data given here by the method of least squares and show the trend values

Year	1990	1992	1994	1996	1998	2000
Gross exfactory values of output (Rs crore)	11	16	13	18	22	20

Solution

$n = 6$, shift the origin t to x by $x = \frac{t - Arithmatic\ Mean\ of\ 2\ middle\ value}{h/2}$

Middle values are 1994 and 1996. AM of middle values = $\frac{1994 + 1996}{2} = 1995$

$$x = \frac{t - 1995}{\frac{2}{2}} = t - 1995$$

Year (t)	Gross exfactory values of output (y)	$x = t - 1995$	x^2	xy	Trend $Y_c = 16.667 + 0.971x$
1990	11	-5	25	-55	11.82
1992	16	-3	9	-48	13.76
1994	13	-1	1	-13	15.7
1996	18	1	1	18	17.64
1998	22	3	9	66	19.58

2000	20	5	25	100	21.52
n = 6	$\sum y = 100$	$\sum x = 0$	$\sum x^2$ =70	$\sum xy$ = 68	100.002

The straight line trend of y on x is $y = a + bx$. Since $\sum x = 0$

$$a = \frac{\sum y}{n}$$

$$= \frac{100}{6}$$

$$= 16.667$$

$$b = \frac{\sum xy}{\sum x^2}$$

$$= \frac{68}{70}$$

$$= 0.971$$

The straight line is

$$Y_c = 16.667 + 0.971x$$

4.2.4 Application of Time Series in Business and Management

Time series analysis is a valuable tool in business and management, offering insights into patterns, trends, and behaviours over time. Here are several key applications of time series in the business and management context:

i. **Sales Forecasting:** Time series analysis helps businesses predict future sales by examining historical sales data. This is vital for inventory management, production planning, and ensuring that resources are allocated efficiently.

ii. **Financial Market Analysis:** Investors and financial analysts use time series to analyse stock prices, currency exchange rates, and other financial indicators. Understanding historical patterns assists in making informed decisions about investments and financial strategies.

iii. **Demand Planning:** Businesses use time series to forecast demand for their products or services. By analysing historical demand patterns, companies

can optimise inventory levels, manage supply chains more effectively, and avoid stockouts or overstock situations.

iv. **Employee Performance and Productivity:** Time series can be applied to track employee performance and productivity over time. This analysis helps in identifying trends, assessing the impact of interventions, and making data driven decisions about workforce management.

v. **Quality Control and Process Improvement:** In manufacturing and operations, time series analysis helps monitor and control product quality over time. It enables businesses to identify trends in defects or errors, allowing for adjustments to improve overall process efficiency.

vi. **Marketing Effectiveness:** Businesses use time series data to assess the effectiveness of marketing campaigns. By analysing how sales or customer engagement fluctuates over different time periods, companies can refine their marketing strategies for better results.

vii. **Customer Behaviour Analysis:** Time series is used to understand customer behaviour over time. This includes analysing purchasing patterns, identifying peak periods of customer activity, and adapting business strategies to meet changing customer preferences.

viii. **Risk Management:** In the financial sector and beyond, time series analysis is employed for risk assessment. By examining historical data, businesses can identify potential risks, model future scenarios, and implement risk mitigation strategies.

ix. **Supply Chain Management:** Time series analysis aids in optimising supply chain processes by predicting demand, identifying seasonality, and helping businesses adapt to changes in market conditions.

x. **Budgeting and Resource Allocation:** Businesses use time series to forecast future expenses and allocate resources effectively. This assists in budgeting for various departments and projects, ensuring financial stability and efficiency.

In summary, time series analysis is a versatile tool in business and management, offering actionable insights for decision makers across various domains. By examining historical patterns, businesses can make informed predictions, optimise processes, and stay agile in an ever changing business landscape.

Recap

- ◊ 3 year moving average - consecutive averaging for 3 year period.
- ◊ 5 year moving average - consecutive averaging for 3 year period.
- ◊ Principle of least squares - minimises the sum of squared deviations between observed values and a trend line
- ◊ Curve Fitting - technique of obtaining mathematical curve by the principle of least squares is called curve fitting.

Objective Questions

1. Which is the most common graphical method of trend analysis?
2. Which is the quantitative method of trend analysis?
3. What is the primary purpose of using the moving average method in time series analysis?
4. In the context of time series, how does the moving average method handle seasonality effects?
5. Define the method of least squares in the context of time series analysis.
6. What is the primary advantage of using the method of least squares for trend measurement?
7. How does the choice of the number of periods impact the moving average method's ability to capture trends in time series?

Answers

1. Moving average method
2. Method of least squares
3. Smoothing out fluctuations to identify trends.

4. It reduces the impact of seasonality by averaging out short-term fluctuations.
5. It is a statistical approach to finding the best-fitting line that minimizes the sum of squared differences between observed and predicted values.
6. It provides a mathematical model to estimate and predict trends based on historical data.
7. A shorter period emphasizes recent data, capturing short-term trends, while a longer period provides a smoother representation of long-term trends.

Self-Assessment Questions

1. Explain the concept of a moving average in the context of time series analysis.
2. If you have monthly sales data for a two-year period, describe the steps involved in calculating a three-month moving average.
3. What is the fundamental principle behind the method of least squares in the context of time series analysis?
4. Given a series of data points, demonstrate how to calculate the least squares regression equation.
5. Compare and contrast the moving average method and the method of least squares

Assignments

1. Given the following monthly sales data for a small business, apply a 3-month moving average to identify the trend.

Monthly Sales	120	130	110	140	125	135	145	130	120	110

2. Utilise the linear trend equation obtained through the method of least squares for the stock prices over the past five years. Predict the stock price for the next two years based on this trend.

Year	2018	2019	2020	2021	2022
Stock Price	45	50	55	60	65

3. You have applied both the moving average method and the method of least squares to analyse a time series dataset. Now, interpret the trends you have identified. How can these trends inform decision-making for a business or inform predictions for the future?

4. Imagine you are tasked with analysing a time series of monthly stock prices. How would you use both the moving average method and the method of least squares to identify trends and potential turning points in the stock prices? What considerations would you take into account in such a financial context?

Suggested Reading

1. Gupta, S. P. (1994). *Statistical Methods*, Sultan Chand & Sons, New Delhi, pp. *E10*, 1-61.

2. Gupta, C. B., & Gupta, V. (2009). *Introduction to statistical methods*. Vikas publishing House Pvt. LTD.

3. Goel, A., & Goel, A. *Mathematics and Statistics*, Taxmann Allied Services Pvt.

4. Kothari, C. R. (2013). *Quantitative Techniques*. Vikas Publishing House.

Reference

1. Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). John Wiley & Sons.
2. Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting* (3rd ed.). Springer.
3. Chatfield, C. (2004). *The analysis of time series: An introduction* (6th ed.). Chapman and Hall/CRC.
4. Cryer, J. D., & Chan, K. S. (2008). *Time series analysis: With applications in R* (2nd ed.). Springer.
5. Enders, W. (2015). *Applied econometric time series* (4th ed.). John Wiley & Sons.
6. Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.
7. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts.

BLOCK - 05
Index Numbers

Unit - 1
Index Number

Learning Outcomes

After completing this unit, the learner will be able to:

- ◊ familiarise the meaning and characteristics of index number
- ◊ get an awareness on the limitations of index number
- ◊ identify the problems in constructing index number

Prerequisite

Index numbers are statistical tools used to measure and track changes in economic, financial, or social indicators over time. They provide a standardised way of comparing and summarising relative changes in different variables by assigning them a reference value or base period. Index numbers serve as a simplified representation of complex data, making it easier to analyse trends, compare values, and draw meaningful conclusions. They are widely used in economics, finance, and various fields to assess inflation rates, economic growth, market performance, and more. Index numbers play a crucial role in data analysis, aiding in decision making, forecasting, and understanding trends in diverse industries and sectors.

Keywords

Index number, Price Index, Quantity Index, Value Index, Inflationary Bias, Economic Indicators, Base Period, Current Year, Constant Weights

Discussion

5.1.1 Introduction

Index numbers offer a standardized and concise way to monitor changes and trends in economic, financial, or social indicators over time. By establishing a base period or reference value, they facilitate comparisons and provide valuable insights into how these variables evolve. This simplifies complex data and allows for smarter decision-making and analysis.

An index number is a device used to measure the relative change in the magnitude of a group of related variables in different situations. The situations may be two different periods of time or places and the group of related variables may be prices or quantities. It represents the general trend of diverging ratios, from which it is calculated. It is a measure of the average change in a group of related variables over two different situations. The comparison may be between categories such as persons, schools, hospitals etc. An index number also measures changes in the value of the variables such as prices of specified list of commodities, volume of production etc. In general, index numbers are used for comparison of changes in the prices or quantities of different periods of time. When comparison is made between prices of two periods, it is called index number of prices and when the comparison is made between quantities, it is called index number of quantities. The year for which the comparison is made is the base year.

5.1.2 Definition

John I. Griffin has described it as "An index number is a quantity which by reference to a base period, shows by its variation, the changes in the magnitude over a period of time." Index numbers are the numbers which express the value of a variable at any time (current period) as a percentage of the value of that variable at some reference period or base period.

Edge worth gave the definition of index numbers as- "Index Number shows by its variation the changes in a magnitude which is not susceptible of either accurate measurement in itself or of direct variation in practice".

Meaning

Index numbers are statistical tools that provide a simplified representation of complex data, making it easier to analyse trends, compare values, and draw meaningful conclusions. These numerical indicators are widely used in economics, finance, and various fields to measure changes in variables over time or across different categories.

5.1.3 Characteristics of Index Numbers

◊ **Relative Measurement**

Index numbers measure changes in relation to a base period. They make a comparison between the variable values in subsequent periods and the base period. This relative measurement makes it possible to analyse changes over time and identify trends.

◊ **Simplification**

Index numbers reduce complex data to a single measure, hence simplifying it. Index numbers give a summary of the general change in the variable rather than providing various data points. This simplification makes it easier to comprehend and evaluate the data.

◊ **Comparison and Benchmarking**:

Index numbers allow data to be compared across time periods, geographies, or industries. They serve as a standard or reference point against which changes in variables can be measured. This comparison assists in understanding the relative performance or changes of several entities or time periods.

◊ **Tracking Trends**

Index numbers are helpful instruments for tracking variable patterns. It is possible to identify patterns, directional changes, and magnitudes of the variable being monitored by analysing changes in index numbers across time. This tracking is useful for forecasting and making decisions.

◊ **Quantitative Representation**

Index numbers are stated as percentages or ratios, and they provide a quantitative depiction of variable changes. This numerical representation makes the magnitude of changes easier to understand and analyse. It enables for more precise analysis and makes comparisons easier.

◊ **Standardised Measurement**

Index numbers give a standardised measurement for evaluating variable changes. They provide a metric that may be applied uniformly across multiple sources or contexts. This standardisation provides data analysis uniformity and comparability.

◊ **Reflecting Weighted Importance**

Index numbers can incorporate weighting schemes to reflect the importance of different variables within the index. Weighting assigns greater significance to certain variables, ensuring that they have a larger impact on the overall index. This consideration of weights ensures a more accurate representation of the variable being measured.

◊ **Application in Various Fields**

Index numbers are used in a variety of sectors, including economics, finance, labour markets, demography, and more. They are used to measure inflation rates, analyse stock

market performance, assess purchasing power, and make informed decisions based on accurate data. Index numbers are useful in a variety of fields due to their adaptability.

5.1.4 Importance of Index Numbers

Earlier index numbers were used to track changes in commodity prices. However, the use of index numbers has expanded significantly in recent years. They are presently used to measure changes in physical quantities, volume of production, volume of trade, cost of products and services, and share and stock prices. Index numbers play a crucial role in various fields, including economics, finance, statistics, and social sciences. These statistical measures serve as essential tools for analysing and interpreting data, tracking changes over time, and making informed decisions.

◊ **Economic Indicators**

Index numbers are critical in measuring economic indicators such as inflation, economic growth, and employment rates. The Consumer Price Index (CPI) measures changes in the average price of goods and services, assisting policymakers in monitoring inflation rates and adjusting monetary policies. GDP and associated indexes provide information on a country's economic performance and growth.

◊ **Comparison across Time and Space**

Index numbers allow for meaningful comparisons of data across different periods, regions, or variables. They provide a standardized framework to measure changes over time or compare different entities. This comparative aspect enables researchers, policymakers, and businesses to assess trends, identify variations, and make informed decisions.

◊ **Development Assessment**

Index numbers are critical for analysing and comparing levels of development across countries or regions. The Human progress Index (HDI) measures a country's overall progress by combining factors such as life expectancy, education, and income. Policymakers can use the HDI to identify areas for improvement, prioritise development activities, and track progress over time.

◊ **Social Science Research**

Index numbers are used in social science research, notably surveys and studies aimed at quantifying subjective experiences. Researchers can analyse and compare large-scale survey data by employing indices to quantify characteristics such as satisfaction, happiness, or well-being. Index numbers make it easier to see patterns, correlations, and social trends, which improves understanding of human behaviour and societal dynamics.

◊ **Business and Marketing Insights**

In business and marketing, index numbers are used to assess consumer happiness, brand loyalty, and market competitiveness. These indices provide vital insights to

businesses, assisting them in making data-driven decisions to improve goods, optimise marketing tactics, and improve customer experiences.

5.1.5 Limitations

Index numbers are versatile statistical tools widely used in economics, finance, and various fields to simplify complex data and make meaningful comparisons. However, they have a number of drawbacks that may compromise their relevance, accuracy, and interpretation

◊ **Base Period Dependency**

Index numbers are typically calculated relative to a base period. The choice of this base period can impact the results significantly. The base period selection has a considerable bearing on the outcomes. It is difficult to compare indices calculated using various basis periods since they can provide different index values. Because of this restriction, the analysis may be biased in terms of time.

◊ **Composition Bias**

Index numbers are frequently used to indicate a defined set of things or components. If the content of this basket varies over time, the index may become biased. A Consumer Price Index (CPI), for example, may not effectively reflect the cost of living if the items in the basket do not correspond to consumer spending patterns.

◊ **Substitution Bias**

Traditional index numbers may not properly account for customer substitution behaviour. When the price of one item in the basket rises dramatically, consumers may switch to cheaper alternatives. Traditional indexes may not effectively capture this real-world customer behaviour.

◊ **Quality Changes**

Index numbers often do not account for changes in the quality of goods and services. For example, improvements in the quality of smartphones or automobiles may not be adequately reflected in an index.

◊ **Weighting Issue**

The weights assigned to different components in an index can be somewhat subjective. These weights may not always reflect the true importance or relevance of these components to the population being measured. Inaccurate weighting can skew the index's representation of reality.

◊ **Inflationary Bias**

Price indices, like the CPI, may overestimate or understate inflation for a variety of reasons, including the challenge of precisely tracking price changes and accounting for changing product quality. Decisions about economic policy may be significantly impacted by this.

◊ **Lack of Regional Variation**

Some indices may fail to account for regional variations in prices and consumption habits. This constraint is especially relevant in countries or areas with varying economic conditions and lifestyles.

◊ **Data Availability**

The precision of index numbers is strongly reliant on the availability of trustworthy data. Data restrictions can have an impact on the index's precision, especially when full data is unavailable.

◊ **Assumption of Constant Weights**

Many indices are assumed to have constant weights across time. In reality, the relevance of various components may vary and consumer preferences can shift. Constant weights may not effectively represent these dynamics.

◊ **Complexity**

Index numbers can be difficult to calculate and interpret, limiting their accessibility and utility, particularly for the general public.

5.1.6 Problems in Construction of Index Numbers

The generation of index numbers is a fundamental method for summarising and comparing data in economics, statistics, and other domains. It is not, however, without hurdles and possible problems. Changes in values over time, across categories, or between distinct data points are represented by index numbers. They are useful for identifying trends, comparing data, and making decisions. Regardless of its utility, index number construction can be a complex process, with various issues that must be carefully addressed to assure the accuracy, relevance, and dependability of the final indices. In this section, we look deeper into the issues surrounding index number production.

◊ **Selection of Base Year**

One of the most important decisions in index construction is selecting a base year or period against which other observations will be evaluated. The base year acts as a reference point, and choosing an incorrect base year might have a substantial impact on the outcomes. Using an out-of-date base year may not adequately reflect current economic conditions, while using a recent base year may miss long-term trends. The base year should preferably be current while also being reflective of average conditions.

◊ **Weighting Issues**

It is critical to assign weights to distinct index components because these weights indicate the relative relevance of each component in the overall index. Choosing adequate weights, on the other hand, can be difficult. Weights should ideally be based on current and precise data; however, this information is not always readily available.

Furthermore, customer tastes and economic conditions can shift over time, making constant weights less useful.

◊ **Data Quality**

The dependability and correctness of the data used to generate an index are critical. Data quality difficulties can originate from a variety of sources, such as sample errors, inaccuracies in reporting, or limits in data collection methods. Data that is inconsistent or biased might result in inaccurate index values and misrepresentations of economic reality.

◊ **Composition Bias**

Many index numbers indicate a predetermined set of things or components referred to as the "basket." This basket's composition is typically determined by historical consumption patterns. If the composition does not closely reflect actual consumption trends, the index may be biased. A Consumer Price Index (CPI), for example, may not effectively reflect the cost of living if the items in the basket do not correspond to what people are currently purchasing.

◊ **Substitution Bias**

Consumer substitution behaviour may not be fully accounted for by traditional index numbers. In practise, if the price of one item in the basket rises dramatically, consumers may opt for less expensive alternatives. Conventional indexes may not effectively represent this real-world consumer behaviour, causing inflation to be overstated.

◊ **Quality Changes**

The quality of goods and services can alter throughout time. For example, technological developments may result in higher product quality or durability. Many index numbers, however, do not take these quality gains into consideration. This constraint might lead to an underestimation of actual economic growth or an overestimation of inflation.

◊ **Basket Updating**

To keep an index relevant, the basket components should be updated on a regular basis to reflect current consumption patterns. The process of picking items for the basket and determining their weights, on the other hand, might be subjective and may not necessarily reflect consumers' genuine preferences. Furthermore, upgrading the basket raises the issue of guaranteeing consistency and comparability with historical data.

◊ **Inflationary Bias**

Due to a variety of reasons, price indices such as the CPI may overestimate or understate inflation. Among these factors are the difficulties in accounting for changing product quality, accurately quantifying pricing fluctuations, and recording shifts in consumption habits. Price measuring precision is critical for making educated economic policy decisions and determining pay adjustments.

◊ **Geographic Variation**

In some circumstances, index numbers may fail to account for regional variations in prices and consumption patterns. This constraint is especially relevant in countries or areas with varying economic conditions and lifestyles. Failing to consider these variations can lead to misleading or inadequate assessments of economic conditions.

◊ **Data Availability**

The availability of extensive and up-to-date data determines an index's accuracy and reliability. In some circumstances, data may be incomplete or missing, making it difficult to create a reliable index. This is especially important in developing countries or places with poor data infrastructure.

◊ **Temporal Variation**

The importance and usefulness of index components may fluctuate throughout time. For example, the importance of specific commodities and services in the consumer basket may change as a result of technological achievements or changing societal requirements. Failure to account for such temporal fluctuations may result in a less relevant index.

◊ **Assumption of Constant Weights**

Many indices assume constant weights across time, implying that the relative importance of distinct components remains constant. However, the importance of individual components can fluctuate due to changes in consumer tastes, technical improvements, or economic trends. Constant weights may not effectively reflect these dynamics.

◊ **Lack of Transparency**

Transparency during the construction process is essential for user confidence and understanding. Failing to document and communicate the methodology, data sources, and assumptions used in creating an index can hinder users' ability to assess its limitations and potential biases.

5.1.7 Types of Index Number

Index numbers are statistical tools that are used to measure changes in economic and noneconomic variables across time. They are commonly used in economics, finance, and other fields to compare the relative changes in distinct quantities, prices, or values. There are various sorts of index numbers, each with a distinct function. Here are some examples of index numbers.

◊ **Price Index**

A Price Index is a sort of index number that is used to track price changes in a certain basket of products and services across time. It measures the degree of inflation or deflation in an economy by comparing current prices to those in a base period.

The Consumer Price Index (CPI), which analyses changes in the cost of living for consumers, is one of the most well-known price indices. Price indices are critical tools for tracking inflation, adjusting salaries, and making sound economic policy decisions.

◊ **Quantity Index**

Quantity Index numbers are concerned with changes in the physical quantities or volumes of products and services, rather than their prices. These indices are often used to analyse production trends in industries such as manufacturing and agriculture. A quantity index, for example, could measure the output of a certain crop or the production of manufactured goods through time. These indexes are useful for analysing changes in physical output, which can be critical for resource allocation and economic planning.

◊ **Value Index**

A Value Index uses both price and quantity information to calculate changes in the total value of a certain set of goods and services. These indices are critical for evaluating general economic trends. For example, the Gross Domestic Product (GDP) deflator is a value index that measures changes in an economy's total price level. Value indices provide a full view of economic performance by taking into account both price and quantity changes, making them useful tools for policymakers and analysts.

Recap

- ◊ Index number - measure changes in relation to a base period
- ◊ Index number - reduce complex data to a single measure
- ◊ Basket components should be updated on a regular basis
- ◊ Price Index - a sort of index number that is used to track price changes
- ◊ Quantity Index number- changes in the physical quantities or volumes of products and services,
- ◊ Value Index - both price and quantity information to calculate changes in the total value
- ◊ Index numbers -give a standardised measurement for evaluating variable changes

Objective Questions

1. What is the primary purpose of index numbers in measurement?
2. In what form are index numbers typically expressed?
3. In which field are index numbers used to quantify subjective experiences such as happiness or well-being?
4. Which bias may occur if the composition of the basket in an index does not reflect actual consumption trends?
5. What does the term "composition bias" in index numbers refer to?
6. What bias may occur if an index does not consider regional variations in prices and consumption patterns?
7. What is the primary purpose of a Consumer Price Index (CPI)?

Answers

1. To simplify complex data
2. Percentage or ratios
3. Social science research
4. Substitution bias
5. Failing to account for consumer substitution behaviour
6. Geographic variation
7. To measure changes in the average price of consumer goods and services over time.

Self-Assessment Questions

1. What is the primary purpose of index numbers in statistics?
2. How do index numbers simplify complex data in various fields?

3. Why is the selection of a base period important in index number calculations?
4. Describe the role of index numbers in tracking trends.
5. What is the importance of reflecting weighted importance in index numbers?
6. Name some fields where index numbers are widely used and provide examples of their applications.
7. What are the limitations associated with the use of index numbers?
8. Discuss the importance of transparency in the construction of index numbers.
9. Provide examples of fields where index numbers find practical applications.
10. Explain the importance of data availability in ensuring the accuracy of index numbers

Assignments

1. Research and provide a detailed analysis of the importance of selecting an appropriate base year for index numbers
2. Examine the challenges associated with assigning weights to index components.
3. Investigate the role of data quality in the construction of index numbers and provide examples of data quality issues and their impact on index values.
4. Conduct a case study on composition bias in a specific index and identify instances where the composition of the "basket" did not accurately reflect consumer trends and the resulting implications.
5. Research and explain the concept of substitution bias in index numbers and analyse real-world scenarios where substitution behaviour was not fully accounted for.
6. Explore how quality changes in goods and services can affect index numbers and provide examples of industries where quality improvements are significant and discuss potential consequences for index calculations.

7. Evaluate the challenges and subjectivity involved in updating the components of a consumption basket in an index and discuss the importance of maintaining consistency and comparability with historical data.

Suggested Reading

1. C.B Gupta and Vijay Gupta (2004) *An Introduction to Statistical Methods.* Vikas Publishing House.
2. S.P. Gupta. *Statistical Methods*. Sultan Chand and Sons, New Delhi.
3. Frederick E Croxton, Dudley J Cowden and Sidney Klein. *Applied General Statistics*. Prentice Hall India.
4. Naval Bajpai. *Business statistics*. Pearson Educational Publications
5. Dr. S M Shukla and Dr. Sahai. *Principles of statistics*. Sahitya Bhavan Publication, Delhi

Reference

1. Naval Bajpai. *Business statistics*. Pearson Educational Publications
2. Dr. S M Shukla and Dr. Sahai. *Principles of statistics.* Sahitya Bhavan Publication, Delhi

Unit - 2

Price Index Number

Learning Outcomes

After completing this unit, the learner will be able to:

- ◊ familiarise with the concept of price index
- ◊ get an awareness on the wholesale price index number
- ◊ assess methods of construction of index numbers

Prerequisite

A price index number is a statistical measure used to track and quantify changes in the average prices of a specified basket of goods and services over time. It serves as a vital tool for assessing inflation or deflation within an economy. The calculation involves comparing the current prices of items in the basket to their prices in a chosen base period. This index provides a percentage or ratio that reflects the overall price level's variation. The Consumer Price Index (CPI) is a well-known example, tracking changes in the cost of living for consumers. Price index numbers play a crucial role in economic analysis, monetary policy, salary adjustments, and understanding purchasing power fluctuations.

Keywords

Price index number, Whole sale price index number, Consumer price index number

Discussion

The most common and widely used type of index is the price index number. Price index, a measure of relative price changes composed of a series of numbers structured in such a way that comparing the values for any two periods or places reveals the average change in prices between periods or the average difference in prices between places. Price indexes were initially developed to measure changes in the cost of living in order to calculate the pay increases required to maintain a consistent standard of living. They are still widely used to predict price changes over time and to quantify cost variations across different places or countries. Price index measures the changes in the price level of the current period on the basis of the price level of the base period.

Price index = (Current year price / Base year price) x 100

$$P_{01} = \frac{P_1}{P_0} \times 100$$

Here, P_{01} is the current year's price index based on the base year price of a given commodity, and 100 is the common denominator.

5.2.1 Types of Price Index

5.2.1.1 Wholesale Price Index Number (WPI)

Index numbers are indicators that indicate changes in commodity prices, industrial production, sales, imports and exports, cost of living, and so on during a given time period. These indicators are useful for reviewing and managing current economic positions as well as formulating plans. Some of the most important indices, such as the Wholesale Price Index (WPI), Index of Industrial Production (IIP), Consumer Price Index (CPI), and others, provide a strong indication of what is going on in the economy. WPI is a key indicator for measuring the dynamic movement of wholesale pricing. Prices are constantly shifting in a dynamic world. WPI is used as a deflator for a variety of nominal macroeconomic indicators, including GDP. WPI-based inflation estimates are also used by the government in the formation of trade, fiscal, and other economic policies. WPI is also utilised in the delivery of raw materials, machinery, and construction activities as an escalation clause. Price adjustment (escalation) provisions in long-term sales and purchase contracts are frequently used by businesses seeking effective methods of dealing with price increases. WPI is commonly regarded as a useful objective indexing method in price adjustment clauses by businesspeople, economists, statisticians, and accountants. As a result, the Wholesale Price Index represents the price of a wholesale products basket.

WPI is concerned with the price of items transferred between corporations. It does not focus on things purchased by consumers. WPI's primary goal is to monitor price fluctuations that reflect demand and supply in manufacturing, construction, and

industry. WPI assists in measuring an economy's macroeconomic and microeconomic circumstances.

5.2.1.2 Retail Price Index

These indices reflect the general changes in the retail prices of various commodities such as consumption goods, stocks and shares, bank deposits, government bonds, etc. In India, these indices are constructed by the Labour Ministry in the form of the Labour Bureau Index Number of Retail Prices—Urban Centres and Rural Centres. Consumer Price Index, commonly known as the Cost-of-Living Index is a specialised kind of retail price index and enables us to study the effect of changes in the prices of a basket of goods or commodities on the purchasing power or cost of living of a particular class or section of the people like labour class, industrial or agricultural worker, low income or middle-income class, etc. In India, cost of living index numbers are available for (i) Central Government employees, (ii) middle-class people, and (iii) the working class.

5.2.2 Methods of Construction of Price Index

The use of simple index number converts prices, costs, quantities for different time periods to comparable index values with base period. Simple index numbers allow comparison of only one item or commodity for different time periods. A decision maker faces a problem when he needs to compare multiple items. This section focuses on techniques for combining several index numbers and determining index numbers for the total (aggregate). The focus will be on constructing aggregate price index numbers. Methods for constructing price indexes can be divided into two categories: unweighted aggregate price index numbers and weighted aggregate price index numbers.

5.2.2.1 Unweighted aggregate Price Index numbers/ Simple Aggregate Price Index Numbers

The unweighted aggregate index is the simplest form of aggregate index numbers. The term unweighted index indicates the equal importance given to all the items considered for computing the index number.

In this method, the total current year prices of the commodities are divided by the total base year prices, and the quotient is multiplied by 100.

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

Here, P_{01} is the Price Index, i.e., the current year's price index on the basis of the price of the base year.

$\sum p_1$ = Total prices of the current year of various commodities.

$\sum p_0$ =Total prices of the base year.

Steps

1. To get $\sum p_1$, add the current year prices of various commodities.

2. Calculate $\sum p_0$ by adding the base year prices of several commodities.

3. To acquire the index number, divide $\sum p_1$ by $\sum p_0$ and multiply the quotient by 100.

Limitations of the method

The simple Aggregate method is highly defective, despite its ease of calculation. The following are some of the drawbacks:

1. When using this method, the items with the highest unit price significantly impact the index number.
2. The relative importance of goods is not taken into account.

Illustration 5.2.1

From the following data calculate Index Number using Simple Aggregate Method.

Commodity	A	B	C	D
Price in 1990 (₹)	160	258	250	139
Price in 1991(₹)	170	165	190	144

Solution:

Computation of price index number

Commodity	Price (in Rupees)	
	1990(p_0)	1991(p_1)
A	160	170
B	258	165
C	250	190
D	139	144
Total	$\sum p_0 = 807$	$\sum p_1 = 669$

The price index number using Simple Aggregate Method is given by:

$$P_{01} = \frac{P_1}{P_0} \times 100$$

$$= \frac{669}{807} \times 100$$

$$= 82.90$$

Illustration 5.2.2

The table given below provides the retail prices for 1998, 2002, and 2006 for five items—soap, edible oil, sugar, rice, and bread that are part of a family's shopping basket. Using this data, compute the unweighted aggregate price index numbers for 2002 and 2006, using 1998, as the base year.

Retail prices of a family's shopping basket in 1998, 2002, and 2006

Items	1998	2002	2006
Soap (1 dozen)	80	100	120
Edible oil (1 litre)	60	75	90
Sugar (1kg)	25	27	30
Rice (1Kg)	20	22	25
Bread (250gm)	15	17	20

Solution:

To compute the unweighted aggregate price index numbers, first of all we need to compute the total (aggregate) price for five items soaps, edible oil, sugar, rice, and bread, as shown in table below:

Items	1998	2002	2006
Soap (1 dozen)	80	100	120
Edible oil (1 litre)	60	75	90
Sugar (1kg)	25	27	30
Rice (1Kg)	20	22	25
Bread (250gm)	15	17	20
Total	**200**	**241**	**285**

Using 1998 as the base year, the unweighted aggregate price index numbers for 2002 can be computed as;

$$I_{2002} = \frac{\sum p_{2002}}{\sum p_{1998}} \times 100 = \frac{241}{200} \times 100 = 120.5$$

Using 1998 as the base year, the unweighted aggregate price index numbers for 2006 can be computed as;

$$I_{2006} = \frac{\sum p_{2006}}{\sum p_{1998}} \times 100 = \frac{285}{200} \times 100 = 142.5$$

The value $I_{2002} = 120.5$ indicates that the price of the items included in the family's shopping the basket has increased by 20.5% when compared to 1998. Similarly, $I_{2006} = 142.5$ indicates that price of the items included in the family's shopping basket has increased by 42.5% when compared to 1998.

5.2.2.2 Weighted Aggregate Method

In this method, appropriate weights are assigned to various commodities to reflect their relative importance in the group. The weights can be production figures, consumption figures or distribution figures. For the construction of the price index numbers, quantity weights are used, i.e., the amount of the quantity consumed, purchased or marketed. If 'w' is the weight attached to a commodity, then the price index is given by:

$$P_{01} = \frac{\sum w p_1}{\sum w p_0} \times 100$$

There are different ways to assign weights to each item in the basket. In addition, there are different ways to use weighted aggregates for calculating an index. This section will focus on a few approaches to determine the method to assign weights to different items in a basket. These methods are as below:

i. Laspeyre's Method

ii. Paasche's Method

iii. Dorbish-Bowley Method

iv. Fisher's Ideal Method

v. Marshall- Edgeworth Method

vi. Kelley's Method

i. Laspeyre's Method

In 1871, German economist Etienne Laspeyre developed this approach for constructing price indices. The method assumes that the quantities consumed in the base year and current year are the same and that weights are determined by the quantities consumed in the base year. The formula is:

$$P_{01}(L) = \frac{\sum p1q0}{\sum p0q0} \times 100$$

Steps

1. Calculate p1q0 by multiplying the current year price of each commodity by the

corresponding base year quantity.

2. Calculate p0q0 by multiplying the base year price of each commodity by its base year quantity.
3. To calculate the price index, divide p1q0 by p0q0 and multiply the quotient by 100.

Advantages

The main advantages of the method are:

1. It is simple to understand and compute.
2. It is based on fixed weights since the quantity of the base year is used as the weight of the item in both years.
3. It does not necessitate quantities consumed in the current year.

Disadvantages

Following are the main disadvantages of the method:

a. The concept that the same quantities are utilised in the base year and current year is inappropriate.

b. It does not allow the use of averages such as geometric mean, median, etc.

c. This method does not satisfy the ideal index number test.

d. It does not use current year quantities, even though they are accessible.

ii. Paasche's Method

Paasche, a German mathematician, developed this method in 1874 to improve Laspeyre's method. He used the current year's quantities as the weights of the items in this method. As a result, Paasche provides the following formula:

$$P_{01}(P) = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

Steps

1. To get $\sum p_1 q_1$, multiply the current year prices of various commodities by their respective current year quantities.
2. Calculate $\sum p_0 q_1$ by multiplying the base-year prices of various commodities by their respective current year quantities.
3. Multiply the quotient by 100 after dividing $\sum p_1 q_1$ by $\sum p_0 q_1$.

Advantages

a. It is easy to calculate and understand.

b. It is based on constant weights for both the current and base years.

c. The base year quantity is not required.

d. It meets the unit test for the adequacy of an index number formula.

Disadvantages

a. The method assumes that consumption quantities are the same in the base year and the current year.

b. Each time an index number is constructed, weights must be determined.

c. The time-reversal, factor reversal, and circular tests required for an ideal index number are not met.

d. It completely ignores the actual quantities consumed in the base year.

iii. Dorbish-Bowley Price Index

The Dorbish-Bowley Price Index is a composite index that combines the Laspeyres and Paasche price indices. It is calculated as the arithmetic mean of the Laspeyres and Paasche price indices. The formula for the Dorbish-Bowley Price Index is given by:

Dorbish-Bowley Price Index = (Laspeyres Price Index + Paasche Price Index) / 2

To understand the formula, let's first define the Laspeyres and Paasche price indices:

Laspeyres Price Index: The Laspeyres Price Index measures the change in the cost of a fixed basket of goods and services from the base period to the current period. It uses the quantities from the base period as weights.

Laspeyres Price Index = $(\Sigma p_1q_0 / \Sigma p_0q_0) \times 100$

Where: p1 = Price of the commodity in the current period p0 = Price of the commodity in the base period q0 = Quantity of the commodity in the base period

Paasche Price Index: The Paasche Price Index measures the change in the cost of a fixed basket of goods and services from the base period to the current period, but it uses the quantities from the current period as weights.

Paasche Price Index = $(\Sigma p_1q_1 / \Sigma p_0q_1) \times 100$

Where: p_1 = Price of the commodity in the current period p_0 = Price of the commodity in the base period q_1 = Quantity of the commodity in the current period

Now, the Dorbish-Bowley Price Index is calculated by taking the arithmetic mean of the Laspeyres and Paasche price indices:

Dorbish-Bowley Price Index = $$P_{01}^{DB} = \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}\right] \times 100$$

The Dorbish-Bowley Price Index is considered a better measure than either the Laspeyres or Paasche index alone because it takes into account both the base period and current period quantities as weights. It provides a compromise between the two indices and is often used when there is a significant difference between the Laspeyres and Paasche price indices due to changes in consumption patterns or substitution effects.

iv. Fisher's Ideal Index Number

The method for constructing index numbers proposed by Prof. Irving Fisher is fisher's ideal index. It is the geometric mean of the index numbers of Laspeyre's and Paasche.

The formula is:

$$P_{01}(F) = \sqrt{\frac{\Sigma p1q0}{\Sigma p0q0} \times \frac{\Sigma p1q1}{\Sigma p0q1}} \times 100$$

Characteristics

a. It is based on the Geometric Mean, which is thought to be the ideal average for building index numbers.

b. It satisfies time and factor reversal tests.

c. It removes the bias associated with Laspeyre's and Paasche's methods by considering both the current year's and base year's weights.

Merits

a. It is created using the Geometric Mean, the most appropriate average for an index number.

b. It uses all available data, including p_1, p_0, q_1 and q_0.

c. It is ideal since it meets time reversal, factor reversal, and unit tests.

d. It reflects the impact of the current year and the base year.

Demerits

a. Calculation is quite complex.

b. Each time an index is constructed, it requires data on the current price and quantity.

c. It's a combination of Laspeyre's and Paasche's methods.

v. Marshall-Edgeworth Index (MEI)

The Marshall-Edgeworth Index (MEI) is a price index that seeks to measure changes in the prices of a fixed basket of goods and services over time.

Taking the arithmetic cross of the quantities in the base year and the current year as weights i.e., w = (q0 + q1)/2, we obtain the Marshall-Edgeworth (M.E.) formula given by:

$$P_{01}(ME) = \frac{\Sigma p1(q0+q1)/2}{\Sigma p0(q0+q1)/2} \text{ x } 100$$

$$= \frac{\Sigma p1(q0+q1)}{\Sigma p0(q0+q1)} \text{ x } 100$$

$$= \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

Therefore,

$$P_{01}(ME) = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

Illustration 5.2.3

From the following data calculate price index numbers for 1980 with 1970 as base by (i) Laspeyre's method, (ii) Paasche's method, (iii) Marshall- Edgeworth method, and (iv) Fisher's ideal method.

Commodities	1970		1980	
	Price	Quantity	Price	Quantity
A	20	8	40	6
B	50	10	60	5
C	40	15	50	15
D	20	20	20	25

a. It is stated that Marshall- Edgeworth index number is a good approximation to Fisher's ideal index number. Verify this for the data in Part(a).

Solution

Calculations for Price Indices in various methods

Commo-dities	1970		1980					
	p_0	q_0	p_1	q_1	p_0q_0	p_0q_1	p_1q_0	p_1q_1
A	20	8	40	6	160	120	320	240
B	50	10	60	5	500	250	600	300
C	40	15	50	15	600	600	750	750
D	20	20	20	25	400	500	400	500
					Σp_0q_0= 1660	Σp_0q_1= 1470	Σp_1q_0= 2070	Σp_1q_1= 1790

i. Laspeyre's Price Index

$$P_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{2070}{1660} \times 100 = 1.24699 \times 100 = 124.699$$

ii. Paasche's Price Index

$$P_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{1790}{1470} \times 100 = 1.2177 \times 100 = 121.77$$

iii. Marshall- Edgeworth Price Index

$$P_{01}^{ME} = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100 = \frac{2070+1790}{1660+1470} \times 100 = \frac{3860}{3130} \times 100 = 123.32$$

iv. Fisher's Price Index

$$P_{01}^{F} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 = \sqrt{\frac{2070}{1660} \times \frac{1790}{1470}} \times 100$$

$$= \sqrt{1.24699 \times 1.2177} \times 100 = \sqrt{1.51846} \times 100$$

$$= 1.23226 \times 100 = 123.23$$

Aliter:

$$P_{01}^{F} = \sqrt{P01La \times P01Pa} = \sqrt{124.699 \times 121.77} = \sqrt{15184.597} = 123.23$$

(b) Since $P_{01}^{ME} = 123.32$ and $P_{01}^{F} = 123.23$, are approximately equal, Marshall-Edgeworth index number is a good approximation to Fisher's ideal index number.

vi. Kelly's Price Index

Kelly's Price Index or Fixed Weights Index. This formula, named after Truman L. Kelly, requires the weights to be fixed for all periods and is also sometimes known as aggregative index with fixed weights and is given by the formula:

$$P_{01}(K) = \frac{\sum p_1 q}{\sum p_0 q} \times 100$$

where the weights are the quantities (q) which may refer to some period (not necessarily the base year or the current year) and are kept constant for all periods. The average (A.M. or G.M.) of the quantities consumed of two, three or more years may be used as weights. Kelly's fixed base index has a distinct advantage over Laspeyre's index because unlike Laspeyre's index the change in the base year does not necessitate a corresponding change in the weights which can be kept constant until new data become available to revise the index. As such, currently this index is finding great favour and becoming quite popular. The Labour Bureau wholesale price index in U.S.A. is based on this method.

Illustration 5.2.4

Calculate the weighted price index from the following data;

Materials required	Unit	Quantity required	Price (Rs.)	
			1963	1973
Cement	100 lb	500 lb.	5.0	8.0
Timber	c.ft.	2,000 c.ft.	9.5	14.2
Steel sheets	cwt.	50 cwt.	34.0	42.20
Bricks	per'000	20,000	12.0	24.0

Solution

Since the quantities (weights) required of different materials are fixed for both the base and current years, we will use Kelly's formula for finding out price index. Further, for cement unit is 100 lbs. and the quantity required is 500 lbs. Hence, the quantity consumed per unit for cement is 500/100 = 5. Similarly, the quantity consumed per unit for bricks is 20,000/1,000 = 20.

Computation of Kelly's Index Number

Materials required	Unit	Quantity required	q	Price (Rs.)		$q p_0$	$q p_1$
				1963	1973		
				p_0	p_1		
Cement	100 lb	500 lb.	5	5.0	8.0	25	40
Timber	c.ft.	2,000 c.ft.	2000	9.5	14.2	19,000	28,400
Steel sheets	cwt.	50 cwt.	50	34.0	42.20	1,700	2,100
Bricks	per'000	20,000	20	12.0	24.0	240	480
						$\sum qp_0$=20,965	$\sum qp_1$=31,020

Kelly's Price Index is given by: $P_{01}^{K} = \frac{\sum q\, p1}{\sum q\, p0} \times 100 = \frac{31020}{20965} \times 100 = 147.96$

5.2.3 Cost of Living Index Number

The wholesale price index numbers measure the changes in the general level of prices and they fail to reflect the effect of the increase or decrease of prices on the cost of living of different classes or groups of people in a society. Cost of living index numbers, also termed as 'Consumer Price Index Numbers, or 'Retail Price Index Numbers' are designed to measure the effects of changes in the prices of a basket of goods and services on the purchasing power of a particular section or class of the society during any given (current) period based on some fixed (base) period. They reflect upon the average increase in the cost of the commodities consumed by a class of people so that they can maintain the same standard of living in the current year as in the base year. Due to the wide variations in the tastes, customs and fashions of different sections or classes of people, their consumption patterns of various commodities also differ widely from class to class or group to group (like poor, lower income group, high income group, labour class, industrial workers, agricultural workers) and even within the same class or group from region to region (rural, urban, plane, hills, etc.). Accordingly, the price movements affect these people (belonging to different class or group or region) differently. Hence, to study the effect of rise or fall in the prices of various commodities consumed by a particular group or class of people on their cost of living, the 'cost of living' Index Numbers are constructed separately for different classes of people or groups or sections of the society and also for different geographical areas like town, city, rural area, urban area, hilly area and so on.

It should be clearly understood that the cost of living index numbers measure the changes in the cost of living or purchasing power of a particular class of people due to the movements (rise or fall) in the retail prices only. They do not measure the changes in the cost of living as a consequence of changes in the living standards. The cost of living index numbers should not be interpreted as a measure of 'Standard of Living'. Cost of living index numbers are based on (retail) prices and price is a factor which affects the purchasing power of the class of the people. But price of the commodities or consumer goods is only one of the various factors on which the standard of living of people depends, some other factors being family size, its age and sex-wise composition, its income and occupation, place, region, etc., none of which is taken into account while computing the cost of living index number. Accordingly, the Sixth International Conference of Labour Statisticians held under the auspices of the International Labour Organisation (I.L.O.) in 1949 recommended the replacement of term 'Cost of Living' index by a more appropriate term "Consumer Price Index' or 'Retail Price Index'.

5.2.3.1 Uses of Consumer Price Index Number

a. The cost-of-living index number is used to adjust dearness allowance in order to maintain the same level of living as in the base date.

b. It is used to determine wage policy, taxation policy, and a variety of other economic issues.

c. It is used to measure money's purchasing power. Money's purchasing power varies inversely with the cost-of-living index.

d. It is employed in income and value deflation. Real income is calculated by dividing real income received during a period by the period's cost of living index.

5.2.3.2 The main steps which are required for construction of CPI are described below

(l) Decision about the class of people for whom the index is meant

It is essential to determine clearly which group of people the index is intended for, such as industrial workers, instructors, officers, and so on. The index's scope must be clearly stated. For example, when we talk to teachers, we are referring to primary teachers, middle class teachers, etc. or to all the teachers taken together. Along with the class of people it is also necessary to decide the geographical area covered by the index.

(2) Conducting, family budget enquiry

After defining the scope of the index, the following stage is to perform a family budget survey of the population group for which the index is to be built. The goal of a family budget investigation is to establish how much an average family in the index spends on various items of consumption. The amounts of commodities consumed and their pricing are taken into account when conducting such an investigation. The consumption pattern can thus be easily ascertained. It is necessary that the family budget enquiry amongst the class of people to whom the index series is applicable should be conducted during the base period.

(3) Deciding on the items

The items on which the money is spent are classified into certain well–accepted groups. One of the most preferred and frequently used classification is –

(i) Food

(ii) Clothing

(iii) Fuel and lighting

(iv) House Rent

(v) Miscellaneous

(4) Obtaining price quotations

The collection of retail pricing is an important but time-consuming and complex task. This is because such prices might vary from place to place, shop to shop, and person to person. Price quotes should be sought from the areas where the persons in question live or shop. Some of the principles that should be followed while collecting retail price data for the purposes of building cost of living indices are given below.:

a. The retail prices should relate to a fixed list of items and for each item., the quality should be fixed by means of suitable specifications.

b. Retail prices should be those actually charged to consumers for cash sales.

c. Discount should be taken into account if it is automatically given to all customers.

d. In a period of price control or rationing, where illegal prices are charged openly, such prices should be taken into account along with the controlled prices.

(5) Working on CPI

After collecting quotations from all the retail outlets, an average price for each of the index goods must be calculated. Such averages are calculated initially for the index's base period and then monthly if the index is maintained on a monthly basis. The method used to average the quotations should produce impartial estimates of average prices paid by the group as a whole. This, of course, will be determined by the method used to select retail shops as well as the scope of the index. Prices or their relatives must be weighted in order to be converted into index numbers. The need for weighting arises because the relative importance of various items for different classes of people is not the same. For this reason, the cost-of-living index is always a weighted index.

5.2.3.3 Construction of Cost of Living Index Numbers

As already pointed out, the relative importance of different items of consumption is different for different classes or groups of people and even within the same class from region to region. Accordingly, the cost of living indices are obtained as weighted indices, by taking into consideration the relative importance of the commodities which is decided on the basis of the amount spent on various items. The cost of living index numbers are constructed by the following methods :

i. Aggregate Expenditure Method or Weighted Aggregate Method

ii. Family Budget Method or Method of Weighted Relatives

Aggregate Expenditure Method

In this method, the quantities consumed in the base year are used as weights. Thus in the usual notations:

$$\text{Cost of Living Index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

$$= \frac{\text{Total expenditure in current year}}{\text{Total expenditure in base year}} \times 100,$$

The total expenditure in current year is obtained with base year quantities as weights. This formula is nothing but Laspeyre's price index.

Family Budget Method or Method of Weighted Relatives

In this method the cost of living index is obtained on taking the weighted average of price-relatives, the weights being the values of the quantities consumed in the base year. Thus, if we write

$$I = \text{Price Relative} = \frac{p_1}{p_0} \times 100 \qquad \text{and} \qquad W = p_0\, q_0, \text{ then}$$

$$\text{Cost of Living Index} = \frac{\sum WI}{\sum W}$$

Substituting the values of W and I, we get

$$\text{Cost of Living Index} = \frac{\sum p_0\, q_0 \left(\frac{p_1}{p_0} \times 100\right)}{\sum p_0\, q_0} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

which is same as the result obtained under aggregate expenditure method.

Thus we see that the cost of living index numbers obtained by both the methods are same.

Illustration 5.2.5

Calculate the Cost of Living Index Number from the following data :

Items	Price		Weights
	Base Year	Current Year	
Food	30	47	4
Fuel	8	12	1
Clothing	14	18	3
House Rent	22	15	2
Miscellaneous	25	30	1

Solution:

Items	Price		Weights(w)	Price Relatives $I=\frac{P1}{P0} \times 100$	WI
	Base Year(p0)	Current Year(p1)			
Food	30	47	4	156.67	626.67
Fuel	8	12	1	150.00	150.00
Clothing	14	18	3	128.57	385.71
House Rent	22	15	2	68.18	136.36
Miscellaneous	25	30	1	120.00	120.00
			ΣW=11		ΣWI=1418.74

$$\text{Cost of Living Index Number} = \frac{\Sigma WI}{\Sigma W} = \frac{1418.74}{11} = 128.98$$

Illustration 5.2.6

In calculating a certain cost of living index no. the following weights were used : Food 15, Clothing 3, Rent 4, Fuel and Light 2, Miscellaneous 1. Calculate the index for a date when the average percentage increases in prices of items in the various groups over the base period were 32, 54, 47, 78 and 58 respectively. Suppose a business executive was earning Rs. 2,050 in the base period. What should be his salary in the current period if his standard of living is to remain the same ?

Solution

The current index number for each item is obtained on adding 100 to the percentage increase in price.

Group (1)	Average % increase in Price (2)	Group Index (I) (3) =100+(2)	Weight (W)	WI
Food	32	132	15	1980
Clothing	54	154	3	462
Rent	47	147	4	588
Fuel and Light	78	178	2	356
Miscellaneous	58	158	1	158
			ΣW = 25	ΣWI = 3544

Cost of Living Index = $\frac{\Sigma WI}{\Sigma W} = \frac{3544}{25} = 141.76$

This implies that if a person was getting Rs. 100 in the base year, then in order to fully compensate the business executive for rise in prices, his salary in the current period should be Rs. 141·76. Hence, if a business executive was earning Rs. 2,050 in the base period, his salary in the current period should be:

$$\frac{141.76}{100} \times 2050 = \text{Rs. } 2906.08$$

To enable him to maintain the same standard of living with reference to price rise, other factors remain constant.

5.2.4 Fixed Base Index Number

When index numbers for a number of years are computed serially on the basis of a fixed base year's data, it is a case of fixed base index numbers.

Such index numbers are otherwise called the price relatives. The base year thus fixed may be the remote most past year, any middle year, any recent past year, average of some years, or average of all the years given. The formula for computing such an index number is

$$P_{R1} = \frac{P1}{P0} \times 100$$

Where, P_{R1} = Price relative of the current year

P_1 = Price of the current year.

P_0 = Price of the fixed base year.

Illustration 5.2.7

(Fixed Base)

From the following data relating to the average prices of a commodity compute the index numbers for each of the seven years taking 2010 as the base year.

Year	2010	2011	2012	2013	2014	2015	2016
Average Prices (₹)	50	60	55	65	75	80	70

Solution

Calculation of fixed base index taking 2010 as base year.

Year	Prices (₹.)	Indices
2010	50	$\frac{50}{50} \times 100 = 100$
2011	60	$\frac{60}{50} \times 100 = 120$
2012	55	$\frac{55}{50} \times 100 = 110$
2013	65	$\frac{65}{50} \times 100 = 130$
2014	75	$\frac{75}{50} \times 100 = 150$
2015	80	$\frac{80}{50} \times 100 = 160$
2016	70	$\frac{70}{50} \times 100 = 140$

5.2.5 Chain Index Numbers

Under this method, firstly we express the figures for each year as a percentage of the preceding year. These are known as Link Relatives. We then need to chain them together by successive multiplication to form a chain index. Thus, unlike fixed base methods, in this method, the base year changes every year. Hence, for the year 2001, it will be 2000, for 2002 it will be 2001, and so on.

Steps in the construction of Chain Index Numbers

Calculate the link relatives by expressing the figures as the percentage of the preceding year. Link Relatives of current year = (price of current year/ price of previous year) × 100

Calculate the chain index by applying the following formula:

Chain Index = (Current year relative ×Previous year link relative) / 100

Illustration 5.2.8 (Chain Base)

From the following data of the prices of certain goods, construct chain base index numbers.

Year	2010	2011	2012	2013	2014	2015	2016
Prices (₹)	75	50	65	60	72	70	69

Solution

Construction of chain base index

Year	Price in ₹	Link relatives	Chain indices 2010=100
2010	75	100.00	100.00
2011	50	$\frac{50}{75}\times100 = 66.67$	$\frac{66.67\times100}{100} = 66.67$
2012	65	$\frac{65}{50}\times100 = 130.00$	$\frac{130\times66.67}{100} = 86.67$
2013	60	$\frac{60}{65}\times100 = 92.30$	$\frac{92.30\times86.67}{100} = 80.00$
2014	72	$\frac{72}{60}\times100 = 120.00$	$\frac{120\times80}{100} = 96.00$
2015	70	$\frac{70}{72}\times100 = 97.22$	$\frac{97.22\times96}{100} = 93.33$
2016	69	$\frac{69}{70}\times100 = 98.57$	$\frac{98.57\times93.33}{100} = 92.00$

Difference between fixed base and chain base methods

Fixed Base	*Chain Base*
Base year is fixed	Base period is not fixed
Base period is arbitrarily chosen	Any period immediately preceding the current year is taken as base.
It is easily understood by a common man	It is not easily understood by a common man
Calculation is easy	Calculation is not easy
It is very rigid not permitting the accommodation of new items	It is quite flexible and so it easily permits the inclusion of some items
It does not permit frequent alteration of the weights of different items	It permits frequent adjustments of the weight of different items
It does not facilitate comparison between two adjacent period	It facilitates comparison between two adjacent periods
It is heavily affected by seasonal variation	It is least affected by seasonal variations
It does not create any problem in calculating the indices of the subsequent period when data of any year is missing	If data f any year is missing, it creates problems in the calculation of index numbers
It is suitable for long period	It is suitable for short period
An error in the calculation of any year does not vitiate the calculation of the balance of years	An error in the calculation of any one year will vitiate the calculation of the balance year

Recap

- ◊ Price index number - a statistical measure used to track and quantify changes in the average prices
- ◊ Price index number - serves as a vital tool for assessing inflation or deflation within an economy.
- ◊ Index provides a percentage or ratio that reflects the overall price level's variation.
- ◊ Wholesale Price Index Number- is a key indicator for measuring the dynamic movement of wholesale pricing
- ◊ Retail Price Index - measures changes in a sector's or community's price level
- ◊ Consumer price index - measures changes in the cost of goods and services purchased by households.

Objective Questions

1. Which index is used to track changes in the cost of living for consumers?
2. What is the formula for calculating a price index number?
3. Which index is primarily concerned with tracking price fluctuations in manufacturing and industry?
4. What does WPI stand for?
5. Which type of index assigns weights to various items explicitly based on quantities or values?
6. What is the primary purpose of a Consumer Price Index (CPI)?
7. Which method of constructing index numbers uses quantities of various commodities consumed in the base year as weights?

Answers

1. Consumer Price Index (CPI)
2. (Current year price / Base year price) x 100
3. Wholesale Price Index (WPI)
4. Wholesale Price Index (WPI)
5. Weighted Index Numbers
6. Assessing inflation or deflation
7. Aggregate Expenditure Method

Self-Assessment Questions

1. How do price index numbers impact economic analysis, monetary policy, and salary adjustments?
2. What are the key components of the Consumer Price Index (CPI) basket of goods?
3. What is the difference between Simple Aggregate Method and Simple Average Relatives Method in constructing index numbers?
4. When is the Simple Aggregate Method considered less reliable, and what is its limitation?.
5. What are weighted index numbers, and why are they used in constructing price indices?
6. List the types of weighted index numbers, and briefly describe each one.
7. Compare and contrast Laspeyre's and Paasche's methods for constructing weighted index numbers.
8. What is Fisher's Ideal Method, and why is it considered an improvement over Laspeyre's and Paasche's methods?
9. Explain the importance of the Marshall-Edgeworth Index (MEI) in constructing price indices and its relationship with Fisher's ideal index.

Assignments

1. Write a brief essay explaining the concept of price index numbers, their purpose, and how they are used to measure changes in the cost of living.
2. Describe the Wholesale Price Index (WPI), its significance in economic analysis, and provide examples of how it is used by businesses and policymakers.
3. Explain the Retail Price Index, its role in measuring changes in consumer prices, and how it differs from the Wholesale Price Index.
4. Discuss the Simple Aggregate Method for calculating price index numbers, including its formula and limitations. Provide an example of its application.
5. Explain the Simple Average Relatives Method for constructing price index numbers, its advantages, and disadvantages. Provide a practical example.
6. Define Weighted Index Numbers and discuss why they are used in price index calculations. Provide examples of different weighting methods and their significance.
7. Describe the Consumer Price Index (CPI), its purpose in measuring changes in the cost of living, and how it is calculated. Explain the uses of CPI in economic analysis.

Suggested Reading

1. C.B Gupta and Vijay Gupta (2004) *An Introduction to Statistical Methods*. Vikas Publishing House.
2. S.P. Gupta. *Statistical Methods*. Sultan Chand and Sons, New Delhi.
3. Frederick E Croxton, Dudley J Cowden and Sidney Klein. *Applied General Statistics.* Prentice Hall India.

Reference

1. Naval Bajpai. *Business statistics*. Pearson Educational Publications
2. Dr. S M Shukla and Dr. Sahai. *Principles of statistics.* Sahitya Bhavan Publication, Delhi

Unit - 3
Quantity and Value Index Number

Learning Outcomes

After completing this unit, the learner will be able to:

- ◊ familiarise with the concept of quantity index
- ◊ identify the value index
- ◊ differentiate between the time reversal and factor reversal tests
- ◊ assess the importance of the unit test, time reversal test, and factor reversal test in checking the consistency of index numbers

Prerequisite

From the previous units, we have learned that an index number is a statistical measure used to represent changes in a set of data relative to a specific reference point. Therefore, it finds wide applications in economics, finance, and various other fields for tracking changes in different quantities over time. When calculating the index number, various methods, such as Laspeyres', Paasche's, and Fisher's index numbers, come into play. However, the critical question arises: How can we ensure that the output from these price indices is unbiased, accurate, and consistent?

To address this concern and check the consistency of index numbers, we employ tests like the unit test, time reversal test, and factor reversal tests. In this unit, our focus shifts to familiarising ourselves with the tools used to verify the consistency of index numbers.

Keywords

Quantity index, Value index, Unit test, Time reversal test, Factor reversal test

5.3.1 Quantity index number

Quantity index number is a classification of index numbers which measures the changes in the quantity or volume of a particular variable within a specified timeframe. A base period is selected as reference and the current or subsequent periods are compared to this base period. They are useful in studying the level of physical output in an economy.

Quantity index number $= \frac{q_1}{q_0} \times 100$

Where,

q_1 is the quantity of the current year,

q_0 is the quantity of the base year

The resulting index number indicates how the quantity of the variable has changed relative to the base period. If the index number is greater than 100, it suggests an increase in quantity compared to the base period, while an index number less than 100 indicates a decrease.

For example, if you want to study how smartphone production has changed from the year 2020 to 2022. Suppose 2020 is the base year with 100000 smartphones produced, giving it an index of 100. In 2021, when 120000 smartphones were produced, the index is 120, indicating a 20% increase from the base year. Conversely, in 2022, with 90000 smartphones produced, the index is 90, signaling a 10% decrease from the base year. These index numbers simplify comparisons, where 100 represents the base year, values above 100 signify growth, and values below 100 indicate a decline. This method facilitates the analysis of trends in various data sets, such as economic indicators like the Consumer Price Index, enabling researchers to monitor changes effectively.

Similar to price index, quantity index can also be calculated using the following methods:

a. Laspeyre's quantity index- In this approach, we assign the base price as the weight, and we solely consider the price of the base year, excluding the prices of the current year.

$$q_{01}(L) = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$$

b. Paashe's quantity index- In this context, the weight is determined by using the current year's price (p_1) for the commodity.

$$q_{01}(P) = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$$

c. Fisher's quantity index- the formula to compute Fisher's quantity index number is:

$$q_{01}(F) = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$$

5.3.2 Value Index Number

Value index numbers are not common like price index numbers and quantity index numbers. Value index number measures the changes in the level of value of items consumed during the year under study with reference to the level of value of item consumed in the base year. The value of an item is the price multiplied by quantity.

$$\text{Value index number } (V_{01}) = \frac{\textit{Total value of items consumed in current year}}{\textit{Total value of items consumed in base year}} \times 100$$

$$= \frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100$$

Illustration 5.3.1

From the following data, calculate Laspeyre's quantity index number.

Commodity	Price		Quantity	
	2022	2023	2022	2023
A	20	22	16	14
B	16	22	12	14
C	18	25	12	10
D	20	26	8	12

Solution

Construction of quantity index number

Commodity	Price		Quantity		q_1p_0	q_0p_0
	p_0	p_1	q_0	q_1		
A	20	22	16	14	280	320
B	16	22	12	14	224	192
C	18	25	12	10	180	216
D	20	26	8	12	240	160
					$\sum q_1p_0$	$\sum q_0p_0$
					924	**888**

Laspeyre's index q_{01} (L) = $\frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$

$= \frac{924}{888} \times 100 = 104.054$

Illustration 5.3.2

From the following data, calculate Paashe's quantity index number.

Commodity	Price		Quantity	
	2022	**2023**	**2022**	**2023**
A	22	26	16	14
B	18	26	12	16
C	20	25	14	10
D	29	33	8	12

Solution

Construction of quantity index number

Commodity	Price		Quantity		q_1p_1	q_0p_1
	p_0	p_1	q_0	q_1		
A	22	26	16	14	364	416
B	18	26	12	16	416	312
C	20	25	14	10	250	350
D	29	33	8	12	396	264
					$\sum q_1p_1$	$\sum q_0p_1$
					1426	**1342**

Paashe's quantity index $q_{01}(P) = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$

$= \frac{1426}{1342} \times 100$

$= 106.259$

Illustration 5.3.3

From the following data, calculate Fisher's quantity index number.

Commodity	Price		Quantity	
	2022	**2023**	**2022**	**2023**
A	20	22	6	4
B	20	26	2	6
C	25	27	4	10
D	29	33	8	2

Solution

Construction of quantity index number

Commodity	Price		Quantity		q_1p_1	q_0p_1	q_0p_0	q_1p_0
	p_0	p_1	q_0	q_1				
A	20	22	6	4	88	132	120	80
B	20	26	2	6	156	52	40	120
C	25	27	4	10	270	108	100	250
D	29	33	8	2	66	264	232	58
					$\sum q_1p_1$	$\sum q_0p_1$	$\sum q_0p_0$	$\sum q_1p_0$
					580	**556**	**492**	**508**

Fisher's quantity index $q_{01}(F) = \sqrt{\frac{\sum q_1p_0}{\sum q_0p_0} \times \frac{\sum q_1p_1}{\sum q_0p_1}} \times 100$

$= \sqrt{\frac{508}{492} \times \frac{580}{556}} \times 100$

$= \sqrt{1.0325 \times 1.0431} \times 100$

$= \sqrt{1.077} \times 100$

$= 103.778$

Illustration 5.3.4

From the following data, calculate value index number.

Commodity	2022		2023	
	Price	Expenditure	Price	Expenditure
A	5	30	15	75
B	9	18	20	60
C	9	54	10	80
D	6	30	12	48

Solution

To compute the quantity of commodities A, B, C, and D, we have to divide expenditure

by price of each commodity.

Quantity of commodity A in 2022 = $\frac{30}{5} = 6$

Quantity of commodity A in 2023 = $\frac{75}{15} = 5$

Construction of value index number

Commodity	Price		Quantity		p_1q_1	p_0q_0
	p_0	q_0	p_1	q_1		
A	5	6	15	5	75	30
B	9	2	20	3	60	18
C	9	6	10	8	80	54
D	6	5	12	4	48	30
					∑p1q1	∑p0q0
					263	**132**

Value index number = $\frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100$

$= \frac{263}{132} \times 100 = 199.242$

5.3.3 Tests of Index Number

Consistency is an important property of index numbers, which are used to measure changes in the relative value of a set of items over time or across different groups. There are several mathematical tests and criteria that can be used to assess the consistency of index numbers. These tests help ensure that the index numbers accurately reflect the underlying data and are suitable for making meaningful comparisons. Here are some of the key mathematical tests of consistency for index numbers:

i. Unit test

ii. Time reversal test

iii. Factor reversal test

5.3.3.1 Unit Tests

The unit test is a mathematical test of consistency of index number that ensures that the index number is independent of the units in which prices and quantities are

expressed. This means that the index number should not change if the prices and quantities are expressed in different units, such as kilograms or pounds, or rupees or dollars. This makes the index number more reliable as a measure of changes in prices and quantities.

5.3.3.2 Time Reversal Tests

This test has been put forth by Prof. Irving Fisher, who proposes that a formula of index number should be such that it turns the value of the index number to its reciprocal when the time subscripts of the formula are reversed. This method is a device to determine if a method will work both ways in time 'backward and forward'.

The time reversal test is a mathematical test used to check the consistency of an index number when comparing two periods in opposite directions. In other words, it examines whether the index value remains the same when you reverse the direction of comparison between two time periods. If the time reversal test is satisfied, it means that the index is consistent when comparing Period, A to Period B and Period B to Period A. The test helps to ensure that the index number accurately reflects the relative changes between the two periods, regardless of the order in which they are compared.

For example, if the price of a commodity has increased to ₹20 per kilogram in 2022, as compared to ₹10 per kilogram in 2021, we would say that the price in 2022 is 200 per cent of the price in 2021 and the price in 2021 is 50 per cent of the 2022 price. Now these two figures are reciprocal of one another and their product (2.00 x 0.50) is equal to unity. If the method does not work both ways, i.e., if the index number for two years secured by the same method but with basis reversed are not reciprocal of each other there is an inherent bias in the method.

Algebraically, the test may be expressed as

$$P_{01} \times P_{10} = 1$$

Where P_{01} stands for index for the current year on the base year omitting the factor 100, (i.e., for price change in current year as compared with base year) and P_{10} stands for index for the base year on the current year without the factor 100 (i.e., for the price change in base year compared with current year).

According to Fisher, "the test is that the formula for calculating an index number should be such that it will give the same ratio between one point of comparison and the other, no matter which of the two is taken as base," or putting it another way, "the index number reckoned forward should be the reciprocal of that reckoned backward."

That Fisher's ideal index satisfies the 'Time Reversal Test' can also be seen from the following illustration:

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}}$$

Changing current year to base

$$P_{10} = \sqrt{\frac{\sum P_0 q_0}{\sum P_1 q_0} \times \frac{\sum P_0 q_1}{\sum P_1 q_1}}$$

Time Reversal test is: $P_{01} \times P_{10} = 1$

$$\text{Now } P_{01} \times P_{10} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0} \times \frac{\sum P_0 q_1}{\sum P_1 q_1}}$$

$$= 1$$

Thus, we see that the indices prepared according to Fisher's ideal formula satisfy the Time Reversal Test.

5.3.3.3 Factor Reversal Test

This test has also been put forth by Prof. Irving Fisher, who proposes that a formula of index number should be such that it permits the interchange of the price, and the quantity factors without giving inconsistent result i.e., the two results multiplied together should give the true ratio in as much as the product of price and quantity is the value of a thing.

According to Fisher "Just as our formula should permit the interchange of the two times without giving inconsistent results so it ought to permit interchanging the prices and quantities without giving inconsistent result, i.e., the two results multiplied together should give the true ratio"

In simple words the test is satisfied if the product of the price index and the quantity index is equal to the ratio of the aggregate value (quantity x price) in the current year to the aggregate value in the base year.

Algebraically: $p_{01} \times q_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_0}$

Where P_{01} stands for the price change for the current year over the base year, q_{01} stands for the quantity change for the current year over the base year, $\sum P_1 q_1$ stands for the total value in the current year, and $\sum P_0 q_0$ stands for the total value in the base year.

That Fisher's ideal index satisfies this test can be seen from the following example:

Commodity	2021		2022		p_0q_0	p_1q_0	p_0q_1	p_1q_1
	p_0	q_0	p_1	q_1				
A	5	12	7	15	60	84	75	105
B	7	10	9	12	70	90	84	108
C	9	8	10	9	72	80	81	90
D	10	5	13	6	50	65	60	78
E	11	3	15	4	33	45	44	60
					285	**364**	**344**	**441**

Factor Reversal Test is satisfied, if

$$p_{01} \times q_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_0}$$

Where p_{01} stands for the price change for the current year over the base year, and q_{10} stands for the changes for the current year over the base year.

Now according to Fisher's ideal index number formula:

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}}$$

and $$q_{01} = \sqrt{\frac{\sum P_0 q_1}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_1 q_0}}$$

Hence,

$$P_{01} \times q_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum P_0 q_1}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_1 q_0}}$$

As per the aforementioned table, we get

$$P_{01} \times q_{01} = \sqrt{\frac{364}{285} \times \frac{441}{344} \times \frac{344}{285} \times \frac{441}{364}}$$

$$= \sqrt{\frac{441}{285} \times \frac{441}{285}} = \frac{441}{285}$$

Now $\frac{\sum P_1 q_1}{\sum P_0 q_0}$ is also equal to $\frac{441}{285}$

Thus, it is proved that Fisher's ideal formula for index number satisfies the 'Factor Reversal Test.'

Illustration 5.3.5

From the following data calculate Fisher's ideal index number and see whether it satisfies both Time Reversal and Factor Reversal Tests.

Commodity	2020		2022	
	Price	Expenditure	Price	Expenditure
M	3	21	5	40
N	5	30	8	24

O	7	56	9	45
P	6	18	7	28
Q	10	20	12	24

Solution

We are given price and expenditure of each commodity. As we are not given quantity, it should be obtained by dividing expenditure by price of each quantity.

Thus, the quantity of commodity M in 2020 = $\frac{21}{3} = 7$

and the quantity of commodity M in 2022 = $\frac{40}{5} = 8$

Calculation of Fisher's ideal index number

Commodity	2020		2022		p_0q_0	p_1q_0	p_0q_1	p_1q_1
	p_0	q_0	p_1	q_1				
M	3	7	5	8	21	35	24	40
N	5	6	8	3	30	48	15	24
O	7	8	9	5	56	72	35	45
P	6	3	7	4	18	21	42	28
Q	10	2	12	2	20	24	20	24
					145	**200**	**136**	**161**

Fisher's ideal index number = $\sqrt{\frac{\sum P_1q_0}{\sum P_0q_0} \times \frac{\sum P_1q_1}{\sum P_0q_1}} \times 100$

$= \sqrt{\frac{200}{145} \times \frac{161}{136}} \times 100$

$= \sqrt{1.379 \times 1.183} \times 100$

$= 127.72$

Time Reversal Test is satisfied if, $= P_{01} \times P_{10} = 1$

Now $P_{01} \times P_{10} = \sqrt{\frac{\sum P_1q_0}{\sum P_0q_0} \times \frac{\sum P_1q_1}{\sum P_0q_1} \times \frac{\sum P_0q_0}{\sum P_1q_0} \times \frac{\sum P_0q_1}{\sum P_1q_1}}$

$$= \sqrt{\frac{200}{145} \times \frac{161}{136} \times \frac{145}{200} \times \frac{136}{161}}$$

$= 1$

Since the answer is 1, it is evident that Fisher's ideal index number satisfies the Time Reversal Test.

Factor Reversal Test is satisfied, if $P_{01} \times q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$

i.e., $P_{01} \times q_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_1}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_1 q_0}} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$

$$= \sqrt{\frac{200}{145} \times \frac{161}{136} \times \frac{136}{145} \times \frac{161}{200}}$$

$$= \sqrt{\frac{161}{145} \times \frac{161}{145}} = \frac{161}{145}$$

Now $\frac{\sum p_1 q_1}{\sum p_0 q_0}$ is also equal to $\frac{161}{145}$

Thus, it is proved that Fisher's ideal formula for index number satisfies the 'Factor Reversal Test.'

Recap

- ◊ Quantity index number- Change in quantity over a period
- ◊ Value index number- Change in value over a period
- ◊ Value = price × quantity
- ◊ Quantity index number = $\frac{q_1}{q_0} \times 100$
- ◊ Value index number = $\frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100$
- ◊ $q_{01}(L) = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$
- ◊ $q_{01}(P) = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$
- ◊ $q_{01}(F) = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$
- ◊ Unit test - Checking whether the index reflects changes accurately when measured in the same units as the original data.
- ◊ Time Reversal Test - Ensuring it is not sensitive to the direction of time.

- ◊ Factor Reversal Test- Evaluates how well an index responds to changes in the weights or factors used in its construction.
- ◊ Ideal index number - Fisher's index number is said to be an ideal index number because it satisfies both time and factor Reversal test.

Objective Questions

1. What is the primary purpose of a quantity index in economic analysis?
2. What is the primary purpose of a value index in economic analysis?
3. In Quantity index formula, what does q_0 represent?
4. How is value calculated in value index number?
5. What is the purpose of the unit test in index numbers?
6. What is the primary purpose of the time reversal test in index number construction?
7. What does the factor reversal test examine?
8. If an index passes both time reversal and factor reversal tests, what does it imply?

Answers

1. To measure changes in the quantity of goods and services
2. To measure changes in both price and quantity simultaneously
3. The base period quantities of goods and services
4. Value = price x quantity
5. To check accuracy for a single unit
6. To check for time-related bias
7. Quantity-related bias
8. It is less biased and more reliable.

Self-Assessment Questions

1. What does quantity index number measures?
2. What are the three methods of quantity index numbers?
3. What do you mean by value index number?
4. Compare and contrast the unit test, time reversal test, and factor reversal test in terms of their objectives and applications in index number construction.
5. Why is Fisher's index number considered an ideal index number?
6. What is the importance of the time reversal test?
7. What is the importance of the factor reversal test?
8. What is unit test?
9. Why should we check the consistency of an index number?
10. How can we check the consistency of an index number?

Assignments

1. Compute quantity index number for the year 2023 with 2020 as the base year using:

   a. Laspeyre's method

   b. Paasche's method

   c. Fisher's method

Commodity	Quantity		Value	
	2020	2023	2020	2023
A	100	150	500	9000
B	80	100	320	300
C	60	72	150	360
D	30	33	360	297

(Answer: q_{01} (L)- 129.744, q_{01} (P)- 131.019, q_{01} (F)- 130.395)

2. The price and quantity demanded for three products are given below:

Product	Price (Year 1)	Price (Year 2)	Quantity (Year 1)	Quantity (Year 2)
A	10	12	100	110
B	20	18	50	55
C	5	6	80	85

Calculate the Fisher price index for these two years and check whether it satisfies the factor reversal test.

(Answer P_{01} (F) =107.464)

3. The price and quantity demanded for three commodities are given below:

Product	Price 2021	Price 2022	Quantity 2021	Quantity 2022
X	11	15	75	107
Y	18	22	65	56
Z	7	13	81	95

Calculate the Fisher price index for these two years and check whether it satisfies the time reversal test.

(Answer P_{01} (F) =141.84)

4. From the following data calculate Fisher's ideal index number and see whether it satisfies both time reversal and factor reversal tests.

Commodity	2021		2022	
	Price	Expenditure	Price	Expenditure
D	12	24	15	45
E	9	36	12	60

F	13	65	16	32
G	8	16	14	98
H	10	20	12	72

(Answer P_{01} (F) =134.002)

Suggested Reading

1. Box and Tiao. *Time Series Analysis*, Forecasting, and Control, Holden Day.
2. C.B Gupta and Vijay Gupta (2004) *An Introduction to Statistical Methods*. Vikas Publishing House.
3. S.P. Gupta. *Statistical Methods*. Sultan Chand and Sons, New Delhi.
4. Frederick E Croxton, Dudley J Cowden and Sidney Klein. *Applied General Statistics*. Prentice Hall India.
5. Naval Bajpai. *Business statistics*. Pearson Educational Publications
6. Dr. S M Shukla and Dr. Sahai. *Principles of statistics*. Sahitya Bhavan Publication, Delhi

Reference

1. Naval Bajpai. *Business statistics*. Pearson Educational Publications
2. Dr. S M Shukla and Dr. Sahai. *Principles of statistics*. Sahitya Bhavan Publication, Delhi

SREENARAYANAGURU OPEN UNIVERSITY

QP CODE: ……

Reg. No : .....
Name : …….....

Model Question Paper Set- I
SECOND SEMESTER BACHELOR OF BUSINESS ADMINISTRATION (BBA) EXAMINATION
DISCIPLINE CORE - 4- B21BB04DC- BUSINESS STATISTICS
(CBCS - UG)
2023-24 - Admission Onwards

Time: 3 Hours Max Marks: 70

Section A

(Answer any 10, each carry 1 mark)

(10x1=10 marks)

1. Which is the field of study that scientifically handles data?
2. What is the term used for collecting data from every unit in the population?
3. What is the term used to refer to data that cannot be measured or counted in the form of numbers?
4. Which measure of dispersion is calculated as the difference between the largest and smallest values in a data set?
5. Describe the relationship between mean, median, and mode in a positively skewed distribution.
6. What is the range of values for Spearman's rank correlation coefficient**?**
7. What does partial regression focus on in a multiple regression analysis?
8. What is the moving average method used for in time series data analysis?

9. At what level does the Wholesale Price Index (WPI) measure price changes?

10. In Laspeyre's method of price index construction, the weights are based on the quantities of which year?

11. Which index number satisfies the time reversal and factor reversal tests?

12. What is the tendency of a time series to increase or decrease over a long period called?

13. Identify the dependent variable in a study investigating the effect of water on plant growth.

14. What is the correlation called if the two variables tend to move together in the same direction?

15. How is the arithmetic mean of grouped data calculated?

Section B

(Answer any 5 each carry 2 marks)

(5x2=10 marks)

16. Define business statistics.

17. Distinguish between nominal and ordinal data with an example for each.

18. What is the significance of accurate data collection in research?

19. Briefly explain the term "arithmetic mean".

20. What is the purpose of the scatter diagram method in correlation analysis?

21. What does the acronym 'z = f(x, y)' represent in the context of regression?

22. What does forecasting mean in the context of time series analysis?

23. What do you mean by price index number.

24. What is an independent variable.

25. What is meant by bi-variate analysis?

Section C

(Answer any 4 each carry 5 marks)

(4x5=20 marks)

26. Explain the difference between primary and secondary data collection methods with examples.

27. From the following table, find Arithmetic mean.

Class Intervals	Frequencies
0-10	4
10-20	10
20-30	15
30-40	13
40-50	8
50-60	20
60-70	14

28. Explain the concept of quartile deviation and its application in describing the spread of a data set.

29. A company manufactures three types of products: X, Y, and Z. The revenue generated by each product and the corresponding weights (based on the company's strategic importance) are given in the following table:

Product	Revenue (in millions $)	Weight
X	30	0.4
Y	20	0.3
Z	45	0.2

Calculate the weighted mean revenue for the company.

30. Calculate Karl Pearson's coefficient of correlation between expenditure on advertising and sales from the data given below.

Advertising expenses ('000 Rs.) : 39 65 62 90 82 75 25 98 36 78

Sales (lakh Rs.) : 47 53 58 86 62 68 60 91 51 84

31. Explain the steps involved in the construction of the Consumer Price Index (CPI).

32. Differentiate between seasonal and cyclic components of a time series with examples.

33. The table shows sales of a manufacturing company in all the 12 months of 2006. Compute a three-month moving average for this time series.

Months	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Sales (million Rs.)	20	19	20	24	25	21	22	23	29	30	32	28

Section D

(Answer any 2 each carry 15 marks)

(2x15=30 marks)

34. Discuss the applications of measures of central tendency (mean, median, mode) in business decision-making processes. Provide specific examples of how each measure can be used in market analysis, employee compensation, and inventory management

35. Explain the methods of Data Collection.

36. From the following data calculate Price Index Numbers for 1980 with 1970 as base by (i) Laspeyre's method, (ii) Paasche's method, (iii) Marshall-Edgeworth method, and (iv) Fisher's ideal method.

Commodities	1970		1980	
	Price	Quantity	Price	Quantity
A	20	8	40	6
B	50	10	60	5
C	40	15	50	15
D	20	20	20	25

37. From the following data, obtain the two regression equations :

Sales	91	97	108	121	67	124	51	73	111	57
Purchases	71	75	69	97	70	91	39	61	80	47

SREENARAYANAGURU OPEN UNIVERSITY

QP CODE: …… Reg. No : .....

Name : …….....

Model Question Paper Set- II
SECOND SEMESTER BACHELOR OF BUSINESS ADMINISTRATION (BBA) EXAMINATION
DISCIPLINE CORE - 4- B21BB04DC- BUSINESS STATISTICS
(CBCS - UG)
2023-24 - Admission Onwards

Time: 3 Hours Max Marks: 70

Section A

(Answer any 10, each carry 1 mark)

(10x1=10 marks)

1. Who is known as the "Father of Statistics"?
2. What is the name of the technique that states that every unit in the population must have a known, non-zero probability of being selected in the sample?
3. What is the name given to the average calculated from a set of observations where one number is chosen based on its position to represent the complete set?
4. What is the main advantage of using standard deviation as a measure of dispersion?
5. What is the meaning of 'correlation coefficient (r) is equal to -1?
6. What is the purpose of time series analysis?
7. What does the method of least squares aim to find?

8. What do index numbers measure in relation to a base period?

9. Which is the mathematical tool that is used to describe the degree to which one variable is linearly related to the other?

10. Which component of a time series represents random fluctuations or unexpected events?

11. What kind of line does the relationship between variables follow when using linear regression?

12. What is the correlation called if the amount of change in one variable does not bear a constant ratio to the amount of change in the other?

13. What does correlation measure?

14. What is the formula for calculating the range in a data set?

15. If the standard deviation of a data set is zero, what can be inferred about the data?

Section B

(Answer any 5 each carry 2 marks)

(5x2=10 marks)

1. What is a questionnaire?
2. What is the purpose of statistical investigation?
3. Explain the concept of stratified sampling.
4. What is the mode of the following series: 3, 5, 7, 7, 10?
5. What is meant by “dispersion” in statistics?
6. What is the purpose of the method of least squares in time series analysis?
7. What do you mean by an index number.
8. State the formula for calculating the price index using the simple aggregate method.
9. Briefly explain univariate analysis.
10. State the empirical relation between mean, median, and mode for a moderately asymmetrical frequency distribution.

Section C

(Answer any 4 each carry 5 marks)

(4x5=20 marks)

16. Discuss the types of regression analysis and the difference between correlation and regression.
17. Discuss the different types of correlation with examples. Explain the significance of understanding the type of correlation in real-life situations.
18. What is Correlation? Explain the different types of Correlation.
19. Explain the difference between a simple arithmetic mean and a weighted arithmetic mean with examples.
20. Calculate the median for the following data set: 12, 15, 14, 10, 18, 20, 16.
21. Analyse the role and significance of business statistics in driving effective marketing strategies for an organization.
22. Calculate Spearman's coefficient of rank correlation for the following data of scores in psychological tests (x) and arithmetical ability (y) of 10 children.

Child	A	B	C	D	E	F	G	H	I	J
X	105	104	102	101	100	99	98	96	93	92
Y	101	103	100	98	95	96	104	92	97	94

23. The following table lists the number of units manufactured by a company in 12 years. Fit a straight-line trend by the method of least squares and estimate the sales in 2010.

Years	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Production (number of units)	1050	1010	1100	1150	1200	1250	1300	1220	1180	1330	1400	1250

Section D

(Answer any 2 each carry 15 marks)

(2x15=30 marks)

24. Discuss the various methods of constructing weighted price index numbers, highlighting their advantages and disadvantages.
25. Discuss the four major components of a time series and their significance in understanding and analyzing time series data.
26. The marks of 70 students in Mathematics exams are as follows.

Marks Scored:	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
No of students	4	7	8	5	8	7	11	9	8	3

Find out the mean, median and mode of the above marks scored by the students and also interpret the performance of the students based on the results found out.

27. A panel of judges A and B graded seven debaters and independently awarded the following marks :

Debaters	1	2	3	4	5	6	7
Marks by A	40	34	28	30	44	38	31
Marks by B	32	39	26	30	38	34	28

An eighth debater was awarded 36 marks by Judge A while Judge B was not present. If Judge B was also present, how many marks would you expect him to award to the eighth debater. Assuming the same degree of relationship exists in judgment.

സർവ്വകലാശാലാഗീതം

വിദ്യയാൽ സ്വതന്ത്രരാകണം
വിശ്വപൗരരായി മാറണം
ഗ്രഹപ്രസാദമായ് വിളങ്ങണം
ഗുരുപ്രകാശമേ നയിക്കണേ

കൂരിരുട്ടിൽ നിന്നു ഞങ്ങളെ
സൂര്യവീഥിയിൽ തെളിക്കണം
സ്നേഹദീപ്തിയായ് വിളങ്ങണം
നീതിവൈജയന്തി പാറണം

ശാസ്ത്രവ്യാപ്തിയെന്നുമേകണം
ജാതിഭേദമാകെ മാറണം
ബോധരശ്മിയിൽ തിളങ്ങുവാൻ
ജ്ഞാനകേന്ദ്രമേ ജ്വലിക്കണേ

കുരീപ്പുഴ ശ്രീകുമാർ

SREENARAYANAGURU OPEN UNIVERSITY

Regional Centres

Kozhikode
Govt. Arts and Science College
Meenchantha, Kozhikode,
Kerala, Pin: 673002
Ph: 04952920228
email: rckdirector@sgou.ac.in

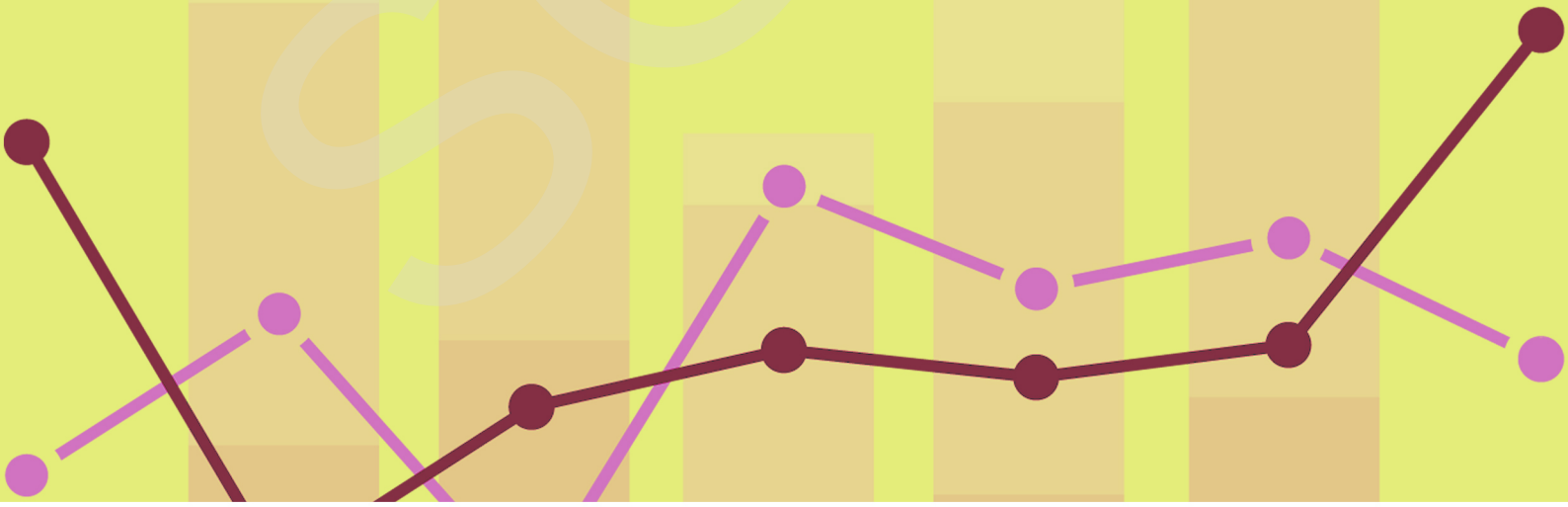
Thalassery
Govt. Brennen College
Dharmadam, Thalassery,
Kannur, Pin: 670106
Ph: 04902990494
email: rctdirector@sgou.ac.in

Tripunithura
Govt. College
Tripunithura, Ernakulam,
Kerala, Pin: 682301
Ph: 04842927436
email: rcedirector@sgou.ac.in

Pattambi
Sree Neelakanta Govt. Sanskrit College
Pattambi, Palakkad,
Kerala, Pin: 679303
Ph: 04662912009
email: rcpdirector@sgou.ac.in

BUSINESS STATISTICS

COURSE CODE: B21BB04DC



ISBN 978-81-970547-8-5

Sreenarayanaguru Open University
Kollam, Kerala Pin- 691601, email: info@sgou.ac.in, www.sgou.ac.in Ph: +91 474 2966841