B.COM
Year – II
Semester – III
Paper - V

Business Statistics

Course Coordinator
Dr. P. Nagarajan
Assistant Professor

Centre for Distance and Online Education

श्रीचन्द्रशेखरेन्द्रसरस्वतीविश्वमहाविद्यालयः

Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya

Deemed to be University u/s 3 of UGC Act 1956 - Accredited with 'A' grade by NAAC

Enathur, Kanchipuram 631561.

Sponsored and run by Sri Kanchi Kamakoti Peetam Charitable Trust

Contents

Unit I: Introduction to Statistics

Structure

Overview

Learning Objectives

- Introduction to Statistics
- Sampling Theory
- Tabulation and Presentation of Data
- Measures of Central Tendency

Overview

Statistics is a field that involves gathering, organizing, analyzing, interpreting, and presenting data to support informed decision-making in areas such as economics, business, and the social sciences. Its major functions include data collection, summarizing information, performing analyses, and drawing meaningful interpretations. The scope of statistics is wide, with applications in economics, business, social sciences, and natural sciences. At the same time, it has certain limitations, such as the risk of misinterpreting data, the inability to prove cause-and-effect relationships, and reliance on the accuracy and quality of the data used.

A statistical investigation generally proceeds through several steps: setting clear objectives, planning the survey or study design, gathering the required data, organizing and processing the information, conducting analysis, and finally presenting the results. The data used may be primary, meaning it is collected directly for the study at hand, or secondary, which refers to information taken from already available sources. Common sources of secondary data include government reports, scholarly articles, online databases, newspapers and other media, as well as publications from various organizations.

This material is designed to help students and learners gain a foundational understanding of statistics as used in business and other fields. It aligns with the prescribed textbooks and reference books.

Learning Objectives:

By the end of this module, you should be able to:

- Explain the meaning of statistics and recognize its importance
- Describe the main functions, scope, and limitations of statistical methods
- Outline the various steps involved in conducting a statistical survey
- Distinguish between primary data and secondary data

- Identify credible sources from which secondary data can be obtained

Introduction to Statistics

1.1 What is Statistics?

Statistics is a field of mathematics that deals with gathering, organizing, analyzing, interpreting, and presenting data. It is essential across many disciplines—such as economics, business, social sciences, and natural sciences—because it supports informed decisions through data-driven insights.

Noteworthy Quotes About Statistics

"There are three kinds of lies: lies, damned lies, and statistics." Often attributed to ***Benjamin Disraeli*** and popularized by ***Mark Twain***, this quote highlights the persuasive power of statistics to bolster weak arguments.

"Statistics is the grammar of science." ***Karl Pearson*** emphasized the foundational role of statistics in scientific inquiry.

"All models are wrong, but some are useful." ***George E. P. Box***, a renowned statistician, acknowledged the imperfections of models while recognizing their practical utility.

"Statistics are like bikinis. What they reveal is suggestive, but what they conceal is vital." This humorous analogy, attributed to ***Aaron Levenstein***, underscores the idea that statistics can both reveal and obscure information.

"Facts are stubborn, but statistics are more pliable", ***Mark Twain*** pointed out how statistics can be manipulated to support various narratives.

"A single death is a tragedy; a million deaths is a statistic." Attributed to ***Joseph Stalin***, this quote reflects on how large-scale tragedies can become impersonal through statistics.

"Statistical thinking will one day be as necessary a qualification for efficient citizenship as the ability to read and write." ***H.G. Wells*** emphasized the importance of statistical literacy in modern society.

Data

Data refers to a set of observations collected for one or more variables of interest. It is generally classified into two categories: **quantitative** and **qualitative**.

i) Quantitative Data:

This type of data can be measured and represented in numerical form. It helps answer questions such as “How much?”, “How many?”, or “How often?”

a) ***Discrete Data:*** Countable values (e.g., number of students).

b) ***Continuous Data:*** Measurable and can take any value within a range (e.g., height, weight).

Example	*Type*	*Explanation*
Age = 25 years	Quantitative (Discrete)	A countable value.
Height = 5.8 feet	Quantitative (Continuous)	Measurable on a continuous scale.
Number of cars = 10	Quantitative (Discrete)	You can count cars.
Temperature = 37.5°C	Quantitative (Continuous)	A measurement on a continuous scale.

ii) Qualitative Data :

Data that describes characteristics or qualities answers the question "What kind?", "Which category?", or "What type?"

a) ***Nominal Data :*** *Categories with no natural order (e.g., gender, colors).*

b) ***Ordinal Data :*** Categories with a meaningful order (e.g., rankings, education levels).

Example	*Type*	*Explanation*
Gender : Male/Female	Qualitative (Nominal)	Categories without order.
Color : Red/Blue	Qualitative (Nominal)	Descriptive categories.
Education Level : Bachelor's, Master's	Qualitative (Ordinal)	Ordered categories.
Customer Satisfaction : High, Medium, Low	Qualitative (Ordinal)	Ordered levels.

Types of statistics

The statistics discipline is broadly divided into two categories:

Descriptive Statistics:

This involves summarizing and organizing data to describe its main features. Common measures include mean, median, mode, and standard deviation, which provide insights into the data’s central tendency and variability.

Inferential Statistics:

This area of statistics focuses on drawing conclusions or making predictions about an entire population by examining a sample. It relies on probability theory to estimate population characteristics and evaluate hypotheses, allowing insights that extend beyond the observed data. Through the use of statistical techniques, researchers and analysts can identify trends, validate assumptions, and make informed, evidence-based decisions. As a result, statistics serves as a vital tool in both scholarly research and real-world problem-solving.

1.2 Importance of Statistics

Statistics is an essential tool that helps us gather, examine, and interpret data, allowing for informed decisions in many areas of everyday life. Below are some important fields where statistics is widely used, along with practical examples:

Weather Forecasting

Meteorologists utilize statistical models that analyze historical and current weather data to predict future conditions. These forecasts assist individuals in planning daily activities and help authorities prepare for severe weather events.

Healthcare and Medicine

In the medical field, statistics are crucial for analyzing patient data, understanding disease patterns, and evaluating treatment effectiveness. For instance, statistical analyses have been instrumental in identifying risk factors for diseases and assessing the efficacy of new medications.

Business and Marketing

Companies rely on statistics to analyze market trends, customer behavior, and sales performance. This information guides strategic decisions, such as product development and marketing campaigns, to enhance profitability.

Traffic Management

Urban planners and traffic engineers use statistical data to monitor traffic flow and accident rates. This analysis informs infrastructure improvements and traffic control measures to enhance road safety and efficiency.

Government and Public Policy

Governments employ statistics to inform policy decisions, allocate resources, and assess program effectiveness. For example, statistical analyses of census data help determine funding distribution and legislative representation.

1.3 What Does Statistics Help You Do? (Functions)

Simplify Complex Data:

Helps summarize large data sets using measures like averages, percentages, etc.

For example: Student Grades Summary, A teacher collects scores of 100 students in a math test. Rather than analyze every student's score individually, she calculates: Average (mean) score = 72%, Highest = 98%, Pass percentage = 85% of students passed.

Make Comparisons:

You can use statistics to compare information across different groups, time periods, or conditions.

Example:

A business reviews its sales from the first and second quarters to assess performance:

Q1 sales = $50,000

Q2 sales = $70,000

"There was a 40% rise in sales from Q1 to Q2, indicating that the second quarter performed more strongly."

Predict Future Outcomes:

Using past data trends, you can forecast future values (like sales or rainfall).

For example Meteorologists use 10 years of rainfall data to predict future patterns. If July usually gets 200 mm of rain, they forecast similar levels. "Based on previous years, July 2025 is expected to get 180–220 mm of rainfall." Similarly, businesses forecast future sales, demand, or stock prices.

Support Decision-Making:

Business managers and governments use statistics to decide on budgets, marketing, etc.

For example: A hospital finds that: 60% of patients come for outpatient services. 30% for emergency care, 10% for surgeries. "Based on usage, we will allocate 60% of the budget to outpatient services." This helps the hospital spend money wisely.

Hypothesis Testing

Researchers use statistical tests to confirm or reject assumptions.

For example: A pharmaceutical company wants to test if Drug A is more effective than the current treatment. They form two groups: Group A gets the new drug, Group B gets the old one. Using statistical tests like t-test , they check if results are significantly different. "The new drug reduced symptoms significantly ($p < 0.05$), so we accept it as more effective."

Control

In production, statistics helps in quality control through control charts and inspection plans.

For example **:** A factory tracks the diameter of metal rods it produces:

Expected diameter = 5 cm, Control limits = 4.95 to 5.05 cm. "One batch showed rods of 5.2 cm — outside limits! The machine was re calibrated."

Control charts help spot errors early and maintain product quality.

1.4 Where is Statistics Used? (Scope)

Statistics is used across various domains:

- ***Business:*** Market research, sales forecasting, inventory control
- ***Economics:*** National income estimation, employment statistics
- ***Biology and Medicine***: Drug testing, disease incidence studies
- ***Education:*** Exam result analysis, evaluation of teaching methods
- ***Social Sciences***: Demographic studies, opinion polls
- ***Government:*** Planning, budget allocation, census operations

1.5 Limitations of Statistics

While statistics is powerful, it's not perfect. Here's why:

Limitation	*Why it matters*
Not for individual cases	It works on groups or large numbers only
Misleading if misused	Wrong methods = wrong conclusions
Needs quality data	Poor data = poor insights
Doesn't show cause-effect	Just because A and B happen together doesn't mean one causes the other
Needs expert interpretation	Misreading charts or numbers can mislead decisions
Cannot study qualitative aspect	Emotions, opinions, and cultural values cannot be directly measured

1.6 What is a Statistical Enquiry?

In statistics, an *enquiry* is a way of collecting data to understand or analyze a particular situation, trend, or problem. This data helps in making decisions, forming policies, or predicting outcomes.

Real-Life Analogy :

Imagine tasting soup!

Population enquiry:

Tasting every drop of the soup pot. Very accurate, but impractical!

Sample enquiry:

Stirring the soup well and tasting a spoonful. If stirred properly, the spoonful reflects the whole pot — just like a well-chosen sample represents the whole population.

Types of enquiry

1.Population Enquiry (Complete Enumeration)

A population enquiry refers to the process of gathering information from every unit within the entire group under study. This method is also called **complete enumeration**.

Example:

Consider a national census. Once every ten years, governments attempt to record details about every individual in the country—such as age, occupation, education, and more. Since information is collected from the whole population, this process represents a population enquiry.

Advantages: High accuracy, since no one is left out, Better insights for decision-making, especially in critical areas like national planning, health care, or education.

Disadvantages:

- It is costly because it requires extensive resources.
- It takes a significant amount of time to gather information from every unit.
- It may not be feasible for very large populations or groups that are hard to access.

2. *Sample Enquiry (Sampling Method)*

A sample enquiry involves collecting data from only a part (or sample) of the population, which is expected to represent the whole.

For example: Suppose a company wants to understand customer satisfaction. Instead of asking all 10,000 customers, they select 500 randomly chosen customers and analyze their responses. The findings are then generalized to all customers. This is a sample enquiry.

Advantages: Cost-effective – fewer resources needed**.** Faster – less time required to collect and analyze data. Useful for large or inaccessible populations**.**

Disadvantages: Less accurate than population enquiry. If the sample isn't well-chosen, the results may not reflect the real situation. Possibility of sampling error**.**

Summary

Type of Enquiry	*Data Collected*	*Cost &*	*Accuracy*	*Example*

	From	*Time*		
Population Enquiry	Entire population (all units)	High	Very High	National Census
Sample Enquiry	Selected sample (some units)	Low	Moderate	Customer survey of 500 people

Objectives of Statistical Enquiry:

Understand a problem

Collect relevant data

Draw conclusions

Recommend solutions

1.7 Stages in a Statistical Survey

Let's say you're surveying mobile phone usage among students. Here's how you'd go about it:

Understand a Problem

Before solving any issue, we need to clearly identify and define the problem. A statistical enquiry starts with a specific question or area of interest.

Example:

A school observes that several students are not performing well in mathematics. The initial step is to identify the root of the issue: Is the problem related to teaching methods, student motivation, or insufficient practice? Clearly defining what needs to be investigated helps prevent misunderstandings later in the study.

Collection of Data

Once the problem is defined, we need to gather data related to it. This could be from surveys, observations, experiments, or records.

For example: The school might collect data on: Student test scores, Attendance in math classes, Homework completion rates, Teacher student ratios. This helps in building a complete picture of the issue. Make sure the data is reliable**,** accurate**,** and directly related to the problem.

Classification and Tabulation

Classification:

It is the process of organizing raw data into logical groups or categories based on common characteristics. Raw data is often messy and confusing. Classification helps

make sense of it by arranging similar items together, making it easier to analyze and draw conclusions.

For example: Imagine you surveyed 100 people about their favorite fruit. Instead of listing all the answers one by one, you group them as: Apples: 35, Bananas: 30, Mangoes: 20, Others: 15. This is classification grouping similar responses into categories.

Types of Classification:

Type	*Description*	*Example*
Qualitative	Based on qualities or attributes	Gender: Male/Female; Education: High School, Graduate
Quantitative	Based on numerical data	Age Groups: 0–18, 19–35, 36–60, 60+
Chronological	Based on time	Sales in 2020, 2021, 2022
Geographical	Based on location	Population by State or Region

What is Tabulation?

Tabulation is the process of presenting classified data in the form of a table, with rows and columns. It helps summarize large data sets for easier understanding and comparison. A table can show patterns, trends, or relationships at a glance — making the data easier to read and interpret.

For Example: Continuing with the fruit survey, the tabulated data might look like:

Fruit	Apples	Bananas	Mangoes	Others	Total
No.of People	35	30	20	15	100

Now, rather than going through all 100 names one by one, you can quickly identify which fruit is the most preferred.

Parts of a Good Table:

Part	*Description*
Title	Explains what the table is about
Rows and Columns	Divide data into categories and values
Headings	Labels for each row and column
Body	The actual data values

Footnotes	(If needed) Additional info or explanations
Source	Where the data came from

Classification vs. Tabulation

Feature	*Classification*	*Tabulation*
What it does	Groups data into meaningful categories	Presents data in table form
Purpose	Organize similar items	Make comparison and analysis easier
Example	Age groups: 0–18, 19–35, etc.	Table showing number of people in each group

Presentation

Raw numbers can be confusing. Presenting data in visual form helps to spot trends, patterns, and comparisons quickly, Communicate information clearly and effectively, Make reports more engaging and easier to understand. *A picture is worth a thousand numbers!. The few important charts are:*

Bar Charts (or Bar Graphs)

A bar chart displays data using rectangular bars to compare different categories. The height or length of each bar indicates the corresponding value.

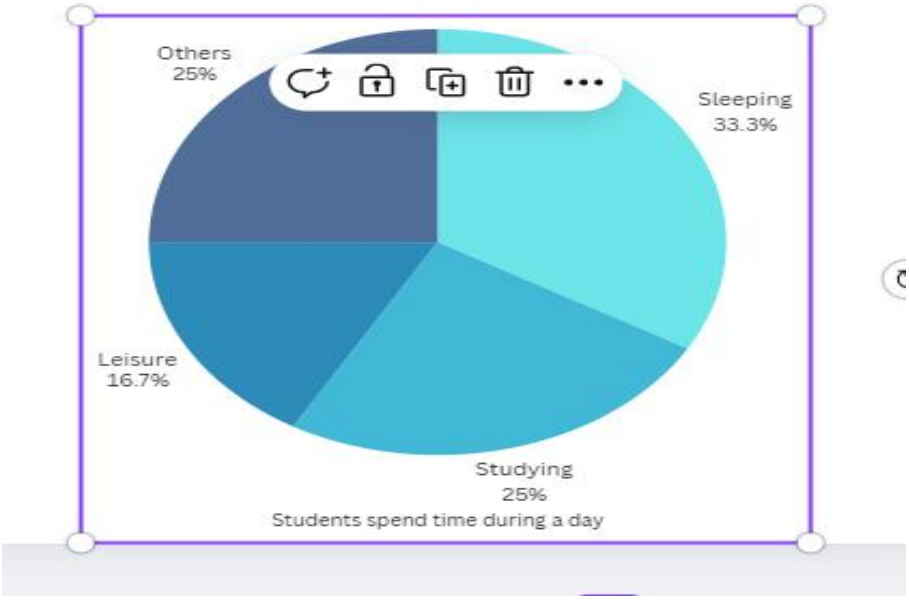
For example : A store tracks the number of items sold in a week:

Item	Pens	Pencils	Erasers	Markers
Quantity Sold	50	30	20	40

Bar chart shows each item on the x-axis, and sales (quantity) on the y-axis. Bars are drawn for each item. Best for**:** Showing distribution of discrete numerical data

Pie Diagrams (or Pie Charts)*:* A pie chart is a circular chart divided into slices. Each slice represents a part of the whole, usually in percentages.

For example: A student spends time during the day as follows: Sleeping: 8 hours, Studying: 6 hours, Leisure: 4 hours, Others: 6 hours. Convert to percentage: Sleep = 33.3%, Study = 25%, Leisure = 16.7%,Others =25% Each slice of the pie reflects these proportions.



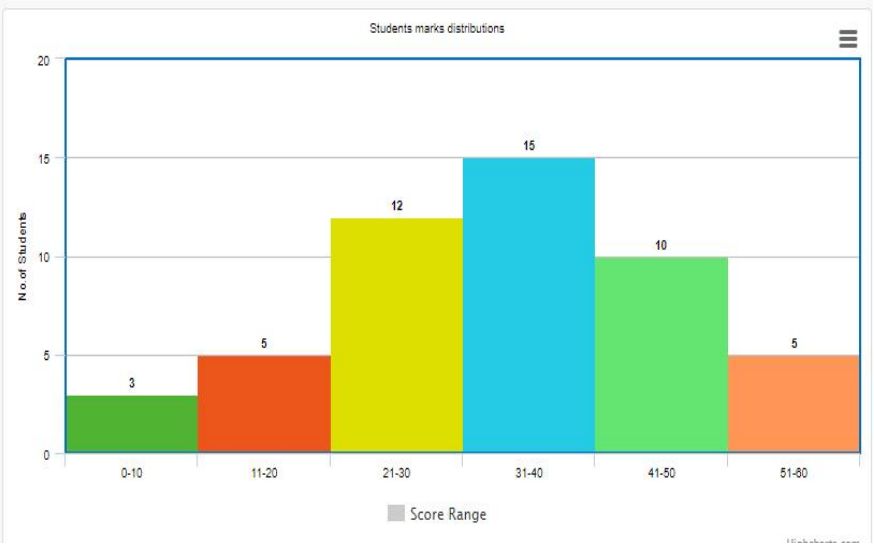
Best for Showing how a whole is divided among categories.

Histograms*:* A histogram is a graphical representation of frequency distribution**.** It looks like a bar graph, but the bars touch each other, indicating continuous data.

For example: Suppose a teacher records the scores of 50 students grouped in intervals:

Score Range	0–10	11–20	20–30	30–40	40–50	50–60
No.of Students	3	5	12	15	10	5

Plot this on a histogram with the score ranges on the x-axis and the number of students on the y-axis. Since the data is continuous, the bars touch.



Best for: Showing distribution of continuous numerical data (e.g., marks, heights, income).

Line Charts: A line graph links individual data points with straight lines. It is especially useful for illustrating how values change over time, such as monthly sales figures or variations in temperature.

Best suited for: Displaying how values change over a period of time (for example, monthly sales trends).

Summary

Chart Type	*Best For*	*Example*	*Strengths*	*Limitations*
Line Chart	Used to illustrate how values change over time	Monthly sales	Easy to see changes and patterns over time	Not ideal for comparing individual categories
Bar Chart	Comparing quantities across categories	Sales by product	Clear comparison between groups	Can get cluttered with too many bars
Pie Chart	Showing parts of a whole (percentages)	Market share	Easy to show proportions	Hard to compare similar-sized slices
Histogram	Showing data distribution	Exam score ranges	Shows frequency clearly	Only for continuous data ranges

Analysis

Apply statistical methods (mean, median, mode, standard deviation)

For example: Suppose 10 students took a math test and got the following scores:

75, 80, 88, 92, 85, 90, 88, 95, 70, 88

***Mean (Average)*:** The sum of all values divided by the number of values.

Mean=[75+80+88+92+85+90+88+95+70+88]/10=85.1

Median: The middle value when the data is ordered.

Ordered Scores: 70, 75, 80, 85, 88, 88, 88, 90, 92, 95**.**

Since there are 10 values (even number), the median is the average of the 5th and 6th scores: Median=[88+88]/2=88

Mode: The value that appears most frequently.

From the list: 88 appears three times, more than any other score.

Mode = 88

Standard Deviation (σ): Measures how much scores deviate from the mean (spread of data).

$$\sigma=\sum(xi-\bar{x})/2n=552.910\approx7.43$$

Interpretation

Summarize insights and make data-driven decisions

For Example: A retail store manager wants to understand monthly sales data to improve business decisions.

Months	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Sales in thousands	50	55	60	58	70	65	75	80	85	78	90	95

Interpretation of given Data

- Sales are generally increasing over the year, with a dip in April and June.
- Highest sales are in November and December (holiday season).
- Lowest sales are in January (post-holiday slump).
- The growth rate accelerates in the second half of the year.

Make Data-Driven Decisions

Based on the insights:

- ***Inventory Planning***: Increase stock before holiday season (Oct-Dec) to meet higher demand.
- ***Marketing Strategy:*** Launch promotions in January to boost sales during the slump.
- ***Staffing***: Hire or schedule more staff in November and December to handle busy periods.
- ***Sales Goals***: Set monthly sales targets based on historical trends to motivate staff.

Why is Interpretation Important?

- It turns raw numbers into actionable knowledge.
- Helps avoid decisions based on assumptions or guesswork.
- Improves business outcomes by aligning actions with real data.

Reporting

Reporting is the process of communicating your analysis results clearly and professionally, so that stakeholders can understand the insights and take action. It typically includes:

- Introduction/Objective
- Data Summary
- Analysis

- Key Insights
- Recommendations
- Visual Aids (charts, tables, graphs)

For example: Reporting on Student Performance

Objective: To analyze the performance of Class 10 students in the math exam and suggest improvement strategies.

Data Summary:

Student	A	B	C	D	E	F	G	H	I	J
Score	75	88	92	70	85	60	88	90	65	95

Analysis:

Mean Score 80.8, Median Score 86.5, Mode 88, Standard Deviation 11.05

Key Insights:

Most students scored between 70–95. One student scored significantly lower (60), which lowers the average. The most frequent score is 88. There is moderate variability in performance.

Recommendations:

- Provide extra support to low scorers (students below 70).
- Encourage high-performing students to mentor peers.
- Focus revision on topics where multiple students lost marks.

Visual Aids:

- Include a bar chart to show distribution of scores.
- A box plot to highlight range and outliers.
- A trend line if comparing across terms.

Final Report Format (in brief):

Section	***Content***
Introduction	Goal: Analyze Class 10 Math Exam
Data Summary	Table of scores
Analysis	Stats: Mean, Median, Mode, SD
Key Insights	What the data tells us
Recommendations	Actionable next steps
Visuals	Graphs/charts to support insights

Purpose of Reporting:

- Clarity: Makes the findings easy to understand
- Structure: Follows a logical flow (what, so what, now what)
- Actionable: Leads to decisions or improvements

1.8 Primary and Secondary Data

Primary Data

Primary data is data collected first-hand by the researcher for a specific purpose or study. It is original, collected directly from sources.

For example:: A university researcher wants to understand students' study habits and how they affect academic performance.

Methods of Collecting Primary Data:

Method	*Description*	*Example in Study Habits Survey*
Direct Interview	Face-to-face or virtual conversation to gather detailed responses.	Interviewing 30 students about how many hours they study daily.
Questionnaires	Structured set of questions, distributed physically or digitally.	Google Forms survey with questions on study time, methods used, sleep, etc.
Observations	Watching subjects in their natural setting without interference.	Observing how students use the library or study rooms.
Experiments	Controlled study to test cause and effect.	Dividing students into groups with different study methods and measuring performance after 1 month.

***Merits of Primary Data*:**

Advantage	*Explanation*
High Accuracy	Data is collected directly from the source.
Relevant	Specifically tailored to your research needs.

.

Demerits of Primary Data:

Limitation	*Explanation*
Time-Consuming	Planning, collecting, and organizing takes time.
Expensive	Resources needed for surveys, interviews, etc.

Secondary Data

Secondary data is data that has already been collected and published by someone else, usually for a different purpose — but you can reuse it for your own analysis.

For Example: You are a business student researching economic growth trends in India over the last 10 years. Instead of collecting your own data, you use: Census of India for population trends., RBI bulletins for interest rate policies., World Bank reports for GDP and inflation, Statista for visual statistics on employment. These are all secondary sources — data already available.

Sources of Secondary Data:

Category	*Examples*
Government Publications	Census of India, Economic Survey, RBI bulletins
International Agencies	UN, IMF, World Bank, WHO, UNESCO
Educational Institutions	University research papers, academic journals, thesis reports
Private Organizations	Company annual reports, market research by firms
Media & Internet	News articles, Statista, MOSPI, Wikipedia (for references)

Merits of Secondary Data:

Advantage	*Explanation*
Quick to Obtain	No need to go out and collect — it's already available.
Cost-Effective	No survey costs, no travel, no materials needed.

Demerits of Secondary Data:

Limitation	*Explanation*
May Not Fit Your Needs	Data might not exactly match your topic or target group.
Could Be Outdated/Biased	It might not be the latest, or the source may have its own agenda or errors.

Where to Find Secondary Data:

Source Type	*Examples*
Government	Census of India, MOSPI, RBI Bulletin
International Agencies	UN, IMF, World Bank, OECD
Academia	ResearchGate, Google Scholar, University repositories
Companies	Tata Group reports, Amazon white papers, Industry trend reports
Online Databases	Statista, Trading Economics, Knoema, CEIC
Media Sources	Times of India, The Hindu, Business Standard, Economic Times

Questions

1. Define statistics and list two of its definitions.
2. What are the main functions of statistics?
3. Mention any three limitations of statistics.
4. What are the key steps in a statistical survey?
5. Differentiate between primary and secondary data.
6. List four sources of secondary data.
7. What are the key differences between population and sample enquiries?
8. Why might a business prefer a sample enquiry over a population enquiry?
9. Can you think of a situation where using a sample might give misleading results?
10. What is the main purpose of classification in statistics?
11. How does tabulation help in data analysis?
12. Think of a dataset from your life (e.g., monthly expenses). How would you classify and tabulate it?
13. When would you use a bar chart instead of a pie chart?
14. Why do histogram bars touch while bar chart bars don't?
15. Think of a daily routine — how would you show it using a pie chart?

1.9 THEORY OF SAMPLING

Meaning: Simply saying sampling consists of obtaining information from a bigger group or a universe.

A social researcher has to get information about a population that has large, differentiated population spread over a large area and that too with in very little amount of time and money.

Collecting information from all members of such a big population is, therefore, always impossible. It is easy to see that part of a whole can give enough dependable information if the procedures followed in collecting the part has of been scientific.

What should be the expected features of a Sample?

- ❖ A proper sample should give a precise but exact picture of the universe from which it is taken.
- ❖ The sample should be collected by probability process. This could permit using of statistical procedures to explain and analyze the sample data.
- ❖ The sample must be as small as precision considerations permit
- ❖ It must be as economical as possible and gathered elegantly to be completed within the given time schedule.

1.10 Sampling Concepts

The following concepts are used in sampling designs

1. population
2. Stratum
3. Elements
4. Sample

1. **Population**

In statistics, a population could be defined as any collection of persons or objects or event in which somebody is interested.

In other words a population consists of the people who are related to the particular problem under investigation.

For example, if we are analyzing the relationship between the class achievements of the university students and the teaching methods then the students of any place and of any time will lie in our population. If we are analyzing the voting behavior or political participation of the Indians then all the adult citizens of India, living in India or abroad will be the elements of the population.

Population Characteristics

In research, we often discuss in terms of population characteristics. e.g. age, sex, income, place of residences, caste, occupation population, size, etc. at the same time all of these characteristics evaluated.

The characteristics which are to be measured depends upon the nature and type of problem of investigation.

Types of Universe

The universe, on the basis of characteristics, can be classified into three types.

a) Univariate population
b) Bi-variate population
c) Multivariate population

a) Univariate population

Here only one characteristic is considered, for studying at a time. Which may be age, income, sex, T.V. listening habit, etc.

b)Bi-Variate Population

It can be defined as when we are measuring two characteristics simultaneously of each member. In sociology we often want to know how characteristics are connected to each other or are associated with each other. For example, we want to know how crime going habit differs from urban people to rural people or how political participation is determined by degree of political awareness etc.

c) Multivariate Universe

In a multivariate universe we consider observations on three or more characteristics at a time. Many social factors together determine the happening of an event. e.g. a car accident is often happening not only by the mechanical factor of the car but also by the factors like, the drivers mental and physical condition, traffic volume, improper signals at crossing, pedestrians behavior etc. likewise poverty is caused by many factors like fast growing population lack of proper industrialization proportional to the growing need of the population, discriminate distribution of wealth, etc.

2. Stratum

When the population is divided into many groups on the basis of one or more characteristics, we call each group as a stratum. Stratum could also be called as a sub population. A stratum could be defined by one or more specifications which divide a population into mutually exclusive segments.

e.g. a given population could be divided into various stratums on the basis of the cinema going habit of the people viz.

a) males who go to cinema frequently,

b) males who rarely go to cinema;

c) males who go to cinema occasionally;

d) males who do not at all go to cinema.

Therefore the number of stratums will depend upon the number of characteristics included for stratification.

3. Population Element

A population element means the units that make the population. Such units might be an individual, an object, or even a small group.

4. Sample

Sample means the aggregate of objects, persons or elements, chosen from the population. It is a part or sub part of the population.

The methods given below are used to collect information about the population

- ❖ Census ;
- ❖ Sampling

Census: When all elements of the population is studied

Sampling: When a small part of the population is chosen for study.

Need of Sampling

Advantages

Helps to gather vital information quickly. Even small samples, when properly chosen.

- ❖ Help to estimates of the characteristics of the population quickly.
- ❖ The modern world is highly dynamic, hence, any study must be completed quickly, otherwise, by the time the survey is finished the situations, characteristics etc may have changed.
- ❖ It reduces costs; enumeration of total population is highly expensive than the sample studies.
- ❖ Sampling techniques increases the accuracy of data. With small sample, it becomes easy to check the accuracy of the data. Some sampling methods make it possible to measure the reliability of the sample estimates from the sample itself.
- ❖ From the administrative angle also sampling becomes easier, because it needs less staff, equipment etc.

Disadvantages

- ❖ Sampling is not possible where knowledge about each element or unit or a statistical universe is required.
- ❖ The sampling procedures should be correctly designed and followed, else, what we call as wild sample, would come up with misleading results.
- ❖ Each type of sampling has its own limitations.
- ❖ There are number of situations in which units, to be measured, are highly variable. Here a very big sample is required in order to get enough cases for achieving statistically reliable information.

- To know some population characteristics like population growth rate, population density etc. census of population at frequent intervals is more appropriate than studying by sampling.

1.11 Sampling Techniques

- Probability Sampling
- Non Probability Sampling

Probability Sampling Techniques

In probability sampling technique one can specify for each element of population, the probability of its being included in the sample. Every probability may be expressed in the form of a proportion e.g. the probability of getting a head in testing a coin is 1/2 or 1 chance in 2 trials. Thus, probability samples are characterized by the fact that the probability of selection of all units are known.

In the sample of example every element of the sample has got the same probability of being included as in random sampling method. An important quality of a probability sample is that it makes possible representative sampling plans. It also gives an estimate of the extent to which the sample characteristics or findings are expected to differ from the total population.

Major types of Probability Sampling Methods :

- Simple random sampling method,
- Stratified random sampling method

1.12 Types of Probability Sampling

1. Simple Random Sampling Method

In our daily life we use the term random is frequently used for careless, unpremeditated, casual haphazard activity or process. It means that a random samples is drawn carelessly in unplanned manner, without a definite aim or deliberate purpose. This concept is wrong. Random sampling exactly means the arranging of conditions in such a manner that every item of the whole universe from which we are to select the sample will have the same chance of being chosen as any other item.

Random sampling, hence, involves careful planning and orderly procedure.

Steps of Simple Random Sampling

- Involves listing of all the elements in the population and assigning them continuous numbers.
- Concluding upon the desired sample size.
- Using any method of sampling, a particular number of elements from the list is selected.

Advantages of Random Sampling Technique

- Highly basic, simple and easy method
- gives a representative sample.

Disadvantages

- Mostly it is difficult to find data list of all units of the population to be sampled.
- The job of numbering all units before the sample is chosen is time consuming and expensive.
- The units need not only to be numbered but also arranged in a particular order.
- The possibility of obtaining a poor or misleading sample is always there when we use random selection.

Methods of Drawing, Sample in Random Method

Lottery Method: The numbers of all the elements of the population are written on different tokens or pieces of paper of equal size shape and color. which are then shuffled completely in a box, or a container. Then tickets are then drawn randomly their numbers are noted and the corresponding individuals or objects are studied.

Tippets Numbers: It was developed by Prof L. H. C. Tippet and since then is called by his name. He developed a list of 10,400 sets of numbers randomly, each set being of four digits there numbers are written on several pages in unsystematic order.

Grid Method: Grid method is applied in selection of the areas. Imagine we have to select any number of areas from a town or any number of towns from a state for survey. Selecting method, first a map of the whole area is prepared, which is then is often divided into various blocks. A transparent plate is made equal to the size of the map which consists of several squared holes in it which carries different numbers. By random sampling method it is finalized as to which numbers are to be included in the sample.

2. Systematic Sampling Method

This method first of all a list is prepared of all the elements of the universe on the basis of a selection criterion. A list can be prepared in alphabetical order, as given in the telephone directory. After that from the list every third, every tenth every twentieth or any number in the like manner can be selected. To apply this method, preparing a list of all the elements and numbering them is a must. Secondly, the population should be homogeneous in nature. Social phenomenon is variable in nature and individuals are heterogeneous. But on their social characteristics they are homogeneous viz. we could decide to cover only the students, the professors, the slum dwellers etc. The characteristics to be selected must be relevant to the problem under study.

Advantages

- It is often used because it is simple, direct and in- expensive.
- If a list of names or items is available, systematic sampling is must be an efficient approach.

Disadvantages

- One must not use systematic sampling in case of exploring unknown areas because listing of elements is not possible
- When there is a periodic fluctuation in the characteristic under study in relation to the order in which the items appear, the methods is ineffective

Stratified Random Sampling Method

Definition: When the population is spitted up into different strata or groups and then samples are selected from each stratum by simple random sampling

procedure,we call it as stratified random sampling method. According to the nature of the problem relevant criteria are selected for stratification. Among the available stratifying criteria, cum age, sex, family income, number of years of education, occupation, religion, race, place of residence etc. On the basis of characteristics universe can be divided into various strata or stratum, Each stratum must be homogeneous from within such a division can be done on the basis of any single criterion. e.g. on the basis of age we can divide people into below 25 and above 25 groups, on the basis of education into matriculates and non matriculates etc. Stratification can also be done on the basis of a combination of any two or more criteria viz. on the basis of sex and education; we can split up the people into four groups.

- ❖ Educated women
- ❖ Un-educated women
- ❖ Education men
- ❖ Un educated men

Then elements are selected from each stratum through simple a random sampling method. An estimate is prepared for each stratum separately. These estimates are combined to provide an estimate for the entire population.

Purpose:

The aim is to increase the representatives of the sample without increasing the size of the sample on the basis of having greater knowledge of the population characteristics.

Advantages

- ❖ The population is first stratified into different groups and then the elements of the sample are selected from each group. Hence, the different groups are sure to have representation in the sample. For a random sample, there is possibility that large groups have greater representation and the smaller groups are often eliminated or under represented.
- ❖ With more homogeneous population greater accuracy could be achieved with fewer cases. This reduces time in collecting and processing of the

data when detailed study about population characteristics are wanted it is more effective.

- ❖ Compared to random samples, stratified samples are geographically more concentrated and therefore save time, money and energy.

Disadvantages

- ❖ Unless there are extreme differences between the strata, the expected proportional representation will be small. Here a random sampling could give a nearly proportional representation.
- ❖ After stratification, the sample is chosen from each stratum either by simple random sampling method or by systematic sampling method; as such the draw backs of both methods could be present.
- ❖ For application of the stratified method, we must know the characteristics of the specified universe in which the study is to be conducted. We must also realize as to which characteristics are related to the subject under investigation and therefore can be considered as relevant for stratification.
- ❖ Stratification becomes more and more difficult as the numbers of characteristics to be used for stratification are increased.

Types of Stratified Sampling

Stratified random sampling method could further be divided into two groups

- ❖ Disproportionate stratified sampling
- ❖ Proportionate stratified sampling

Disproportionate stratified sampling:

Disproportionate stratified sampling is also known as same size stratified sampling. In this method, an "equal number" of cases are chosen from each stratum irrespective of the size of the stratum in the universe. The number of cases chosen from each one is controlled to the number of pre designated in the plans. It is called "controlled sampling" as the number of cases to be selected in various strata is restricted.

Advantages

- ❖ When equal numbers of cases are taken from every stratum, comparing different strata are facilitated.
- ❖ Economy of procedure
- ❖ The controlled sample prevents the researchers from securing an unnecessary high number of schedules for most prevalent groups of universe.

Disadvantages

- ❖ It needs the weighing of results stratum by stratum, the relative frequency of every b stratum in the universe must be known or estimated to find the weights.

Proportionate stratified Sampling:

In this method cases are selected from every stratum in same proportion as they occur in the population. To apply this method we first of all we should have a list of all stratum and also need to understand their proportionate size in total population. Since the size of the stratum differ, the number of elements coming from each stratum in the sample on the basis of selection of a given percentage of people will also differ.

Advantage

- ❖ Proportional representation is definite.

Disadvantage

- ❖ The researcher could have poor judgment or less information upon which to base the stratification. The higher the number of characteristics on which we are to do our stratification, and the more are the strata the more complicated becomes the problem of acquiring proportional representation of every stratum.

Cluster Sampling

Here the stratification is performed in a manner that the groups are heterogeneous in nature than homogeneous. Here the elements are not chosen from each stratum as is done in stratified sampling, rather the elements are obtained by choosing a sample of group and not from within groups,Which means that out of many clusters or groups, one, two or more number of clusters are chosen by simple or stratified random method and the elements are analyzed.

All the members in the clusters are not going to be selected in the sample; the final selection from within the clusters is also done on simple or stratified sampling basis.

Aim: The Aim of a cluster sample is to reduce cost only and not to increase percussion.

Advantage

- ❖ In cluster sampling the cost per member is greatly reduced.
- ❖ We can take a larger sample and regain the amount of precision
- ❖ It can be used in when it is impossible to obtain sample by other methods.

Disadvantage

- ❖ It is a complicated sample design the person who is doing inquiry has to be highly talented in sampling.
- ❖ Its standard errors are higher than those of sample random sampling.

Multi-stage sampling

The method is used in choosing a sample from a very big area. As the name tells this type of sampling refers to a sampling technique which is done in different stages. Usually a multi-stage sampling is the one that uses both cluster and random sampling methods.

Eg., if we want to analyze the socio-economic background, attitudes and motivations of poor people, we can first make a list of the cities which would thus make our clusters.

From these clusters we can choose few cities. Then each city or cluster would be stratified into different slum areas. Now the cities can be called as primary sampling units and the slum areas as secondary sampling units.

Non-Probability Sampling

Non probability sampling is the one in which one cannot conclude before had the probability of each element being included in the sample.

The three forms of non-probability samples are;

- ❖ Accidental samples
- ❖ Quota samples and
- ❖ Purposive samples

Accidental Samples:

In accidental sampling the units are chosen on the basis of quick approaches. Here one chooses the sample that falls to hand easily.

E.g. suppose one researcher is investigating the political socialization and political participation in university and college students of A.U. and the sample size is 100.

He goes to the university campus and could choose the first hundred students whom he meets in the way, whether in class room, or in students general room or in field. This type of sampling is easy to do and saves time and money. But the chances of bias are also high.

Quota Sampling

In quota sampling the researchers are interested to interview a particular number of persons from every category. The needed numbers of elements from every category are determined in the office ahead of time as per to the number of elements in each category. Thus a researcher could need to contact a specified number of men and specified number of women, from various age categories from different religious or social groups etc. The aim of quota sampling is choosing selection a sample that not true replace of the population about which we want to generalize.

Advantage

- If perfectly planned and executed, a quota sample will most likely to give maximum representative sample of the population.
- In purposive sampling we pick up the cases that are considered to be typical of the population in which one is interested.
- The cases are judged to be typical on the basis of the requirement of the interviewer.
- Since choosing elements is based upon the judgment of the researcher, the purposive sampling as called judgment sample.
- The researcher trees in his sample to match the population in some of the important known characteristics.

Disadvantage

- The disadvantage with this method is that the researcher could easily make error in judging as to which cases are typical.

Purposive Sampling - "Deliberate Sampling" or "Judgment Sampling".

- When the researcher quickly chooses few units from the population, it is called as purposive sampling.
- But it should be kept in mind that the units chosen should be representative of the population.
- The names may be selected from a Telephone Directory, Automobile Registration Records (RTOs) etc.

Advantage

- Quote sampling is a stratified and purposive sampling and therefore enjoys the features of both samplings.
- If perfect controls or checks are imposed, it will give accurate results.
- It is highly useful method if there is no sample frame available.

Convenience Sampling

It is understood as unsystematic, careless, accidental or opportunistic sampling. In this method sample is selected according to the convenience of the Researcher.

The following situation this method may be used.

- ❖ Universe is not correctly defined
- ❖ Sampling units are unclear
- ❖ Complete source list is unavailable

Essential characteristics of Sampling

If the sample results are to have any significant meaning, it must possess the following essential characteristics.

- **Representativeness:** A sample must be so selected that it actually represents the population, otherwise the results obtained might be misleading.
- **Adequacy:** The sample size must be adequate else it may not represent the characteristics of the universe.
- **Independence:** All the items of the sample must be chosen independently of one another and all the items of the population must be the same chance of being selected in the sample.
- **Homogeneity:** The word homogeneity gives the meaning that there is no basic difference in the nature of the population and that of the sample. It two samples from the same population are considered, they should give more or less the same unit.

Sampling Method Adopted

- The sample size is also influenced by the type of sampling plan applied. For example, if the sample is a simple random sample, it could necessitate a bigger sample size. However, in a perfectly drawn stratified sampling plan, even a small sample could give a better result.

Respondent's Nature

- If it is expected that more number of respondents may not co-operate and send back the questionnaire, a big sample must be selected.

Determination of Sample Size

- There are more number of formulae have been devised for determining the sample size according to the availability of information.

  $n = \left(\frac{Z-\sigma}{d}\right)^2$

  Where

  n = sample size

  z = value at a specified level of confidence or desired degree of precision

  σ = standard deviation of the universe

  d= difference between population mean and sample mean.

1.15 SAMPLING AND NON SAMPLING ERRORS

- ❖ The error happen due to drawing inferences about universe on the basis of some observations (sampling), is termed 'sampling error'.
- ❖ In the complete enumeration survey since the full population is surveyed, sampling error does not exist. But the mainly arising error at the stage of ascertainment and processing of data that are termed non-sampling errors, are prevalent both in complete enumeration and sample surveys.

Sampling Errors: Even if we take utmost care in selecting a sample, the results obtained from a sample study could not be exactly equal to the true value in the population. This is because estimate is based on a part and not on the whole and samples are seldom, if ever, perfect miniature of the universe. Hence sampling gives rise to some type of errors known as sampling errors. However, the errors may be controlled. The modern sampling theory helps in designing the survey in such a way that the sampling errors could be made small.

Sampling errors are of two types:

- ❖ biased,
- ❖ un-biased

Biased Errors: These errors come from any bias in selection, estimation, etc. As an example, if in place of simple random sampling, deliberate sampling has been applied in a particular case some bias is introduced is the result and hence such errors are called sampling errors.

Un-biased Errors: These errors happens due to "chance" differences between the members of the population added in the sample and those not included. An error in statistics is the difference between the value of a statistic and that of the corresponding parameter.

- ❖ Therefore the total sampling error is made up of errors due to bias and the random sampling error.
- ❖ The bias error, forms a constant component of error that is not decreasing in large population as the number of sample increases. The other name for such error is also known as cumulative or non-compensating error. The random sampling error, at this time, decreases, on an average, as the size of sample increases. Such errors are, therefore, called as non-cumulative or compensating error.

Causes of Bias: Bias could arise due to:

- ❖ incorrect process of selection;
- ❖ Incorrect work during the collection; and
- ❖ Wrong methods of analysis

Faulty Selection: Careless selection of a 'representative' sample.

Substitution: Substitution of an item in place of one selected in random sample may lead to bias.

Non response: If all the elements to be included in the sample are not covered then there will be bias even though no substitution is attempted.

Ambiguity in questions will give rise to yet another kind of bias. As an example, the question. Are you a good student? Is such that most of the students would answer 'Yes'.

Bias Due to wrong Collection of Data: Any consistent error in measurement may give rise to bias whether the measurements are carried out on a sample or on all

units of the universe. The danger of error is, however, may be greater in sampling work. Bias may happen due to improper formulation of the decision, problem or strongly defining the population etc. Bias observation will result from poorly designed questionnaire, ill trained researcher, failure of a respondents memory.

Bias in Analysis: Except bias, which arises from wrong process of selection and wrong collection of information, faulty methods of analysis could also introduce such bias. Such bias may be avoided by adopting the perfect method of analysis.

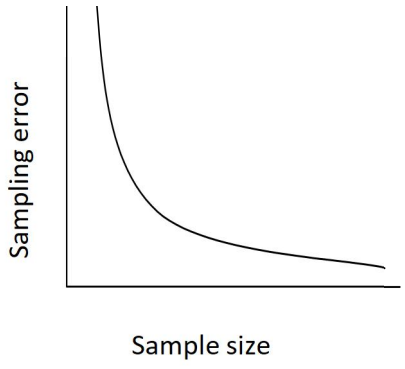
Avoidance of Bias: If the possibility of bias is still there ,fully objective conclusion could not be drawn. The first important property of any sampling procedure should, therefore, be the elimination of all sources of bias.

1.16 Method of Reducing Sampling Errors

Once the absence of bias has been confirmed, attention must be given to the random sampling errors.

Those errors must be reduced to the minimum so as to obtain the desired accuracy.

In addition to reducing errors of bias, the easiest way of increasing the accuracy of a sample is to increase the sample size. The sampling error normally decreases with increase in sample size, and in many situations the decrease is inversely proportional to the square root of the sample size.



From this diagram it is easy to see that though the reduction in sampling error is substantial for initial increases in sample size, it is marginal after a certain stage. In other words, considerably greater effort is required after a particular stage to decrease the sampling error this is the initial instance.

From this angle it can be said that there is a strong case for resorting to a sample survey to give estimates within permissible margins of error rather than taking a complete enumeration survey.

1.17 Non Sampling Errors

Regarding to non-sampling errors they may be more in case of complete enumeration survey than in a sample survey. When a complete enumeration of units in the population is needed, one could expect that it will give rise to date free from errors. However, really it is not so. For example, it is not easy to completely reduce errors of observation or. Likewise, in the processing of data, tabulation errors can be committed, affecting the final result. Errors coming in this manner are called as non-sampling errors. Non-sampling error will occur at each stage of planning and execution of census or survey. Those errors may arise due to a number of causes such as defective methods of data collection, and tabulation, wrong definition, incomplete coverage etc. In particular, non-sampling error will arise from one or more of the following factors:

Data specification could be inadequate and inconsistent with respect to the objectives of the study.

- Inaccurate or inappropriate ways of interview, observation or measurement with inadequate on ambiguous schedules.
- Non availability of trained and experienced investigators.
- Absence of inadequate inspection and supervision of primacy staff. Errors occurs because of non-response.
- Errors happening in data processing operations.
- Errors committed during presentation and printing of tabulated results.

Controlling of Non Sampling Errors:

In a few situations the non-sampling errors can be large and deserve greater attention than sampling errors. But, in general, sampling error reduces with increase in size of the sample, non-sampling error likely to increase with the sample size.

In case of complete enumeration non-sampling errors and in case of sample surveys both sampling and non- sampling errors needs to be controlled and reduced at a level at which their presence do not vitiate the use of final result.

Reliability of Samples:

The reliability of samples may be tested in the following ways.

More samples of the same size must be taken from the same population and their results be compared. If the results are similar, the sample may be reliable.

If the measurements of the population are known, then they must be compared with the measurements of the sample. If similarity of measurements happens the sample will be reliable.

1.18 Tabulation and Presentation of Data Basic Charts and Their Suitability

Introduction

In the world of data analysis, collecting raw data is only the beginning. For data to be meaningful and actionable, it needs to be organized and presented in a structured manner. This process begins with **tabulation**, which arranges data systematically into rows and columns for ease of interpretation. Once the data is tabulated, it can be further simplified and visualized through **charts and diagrams**, making complex datasets accessible even to non-technical audiences.

The **presentation of data** is not merely about creating attractive visuals – it is about selecting appropriate methods of representation based on the nature of data, the target audience, and the objectives of analysis. Effective data presentation can highlight trends, reveal patterns, and aid decision-making.

This self-learning material provides a comprehensive understanding of how to transform raw data into informative tables and clear visualizations. We will explore the types of data tables, various chart formats like bar graphs, pie charts, line graphs, histograms, and the specific contexts in which each is most suitable.

Why This Topic is Important:

- Helps in organizing and summarizing data effectively.
- Facilitates easy comparison and interpretation of complex data.
- Enhances clarity in statistical reports and presentations.
- Forms the foundation for deeper statistical analysis.

By the end of this unit, you will be equipped with the fundamental skills to present data meaningfully—whether in academic work, research, or business settings.

Learning Objectives

After completing this unit, you will be able to:

- Understand the purpose and importance of data tabulation in statistical analysis.
- Identify various forms of tables and recognize their components.
- Differentiate between types of charts such as bar charts, pie charts, line graphs, and histograms.
- Select the most suitable chart or diagram based on the type of data and the objective of the presentation.
- Construct simple tables and charts using real-life business data.
- Interpret tabular and graphical data effectively for decision-making in commerce and business.

1. 19 Tabulation of Data

Tabulation is the process of arranging data in a systematic and logical order using rows and columns. It helps in simplifying complex data and makes it easier to compare, analyse, and draw conclusions. In business and commerce, tabulated data plays a key role in reports, records, and analysis of financial and marketing trends.

1.20 Objectives of Tabulation

The main objectives of tabulation are:

- **Simplification:** Tabulation helps to condense large data sets into a format that is easy to read and understand.
- **Comparison:** It enables comparison between different variables or over periods.
- **Clarity:** A table presents facts in a clear and organized manner without any confusion.
- **Saving Time and Space:** Tabulated data occupies less space and takes less time to interpret compared to descriptive data.
- **Basis for Graphs:** Tables act as a foundation for creating charts and diagrams.

1.21 Parts of a Statistical Table

A good statistical table includes the following parts:

1. **Table Number:** Every table should be numbered for easy reference.

2. **Title:** The title should be clear and mention what the table is about. Example:

"Monthly Sales of XYZ Ltd. for 2024."

3. **Caption (Column Headings):** The top row contains headings for each column.

4. **Stub (Row Headings):** The first column which contains headings for rows.

5. **Body:** This is the main area where numerical data is presented.

6. **Footnote:** Any additional explanation or unit of measurement is given below the

table.

7. **Source:** If the data is taken from some report or agency, the source should be

mentioned.

Example of a Simple Table:

Month	Sales (in Rs. '000)
January	450
February	520
March	490

1.22 Uses of Tables

Tables are widely used in business, economics, statistics, and various social sciences. Their uses include:

- Presenting large amounts of data compactly.
- Highlighting relationships and trends in the data.
- Serving as a base for making statistical calculations.
- Providing ready-made data for reports and presentations.
- Assisting managers in making business decisions based on past performance.

1.23 Limitations of Tables

Though tables are very useful, they have certain limitations:

- Not as visually effective as charts and graphs.
- Complex tables can be confusing for common users.
- Requires careful reading and interpretation.
- Sometimes, too much information in a table can overwhelm the reader.

1.24 Types of Tables

Based on complexity and purpose, statistical tables are broadly classified as:

1. **Simple Table:** Shows data about one characteristic only.

Example: Sales of a product in a year.

2. **Complex Table:** Shows data about two or more characteristics.

Example: Sales of different products over several years.

3. **Frequency Table:** Lists each item or group and shows how often it occurs.

Example: Frequency of customers visiting a store per day.

4. **Two-way Table:** Displays data according to two variables at once.

Example: Age and Gender-wise distribution of employees.

Example of Two-way Table:

Age Group	Male	Female	Total
18-25	12	18	30
26-35	15	20	35
36-45	10	5	15

Tables like this help in presenting demographic or business data with multiple categories.

1.25 Frequency Table and Tally Table

A **frequency table** shows how often each value in a dataset occurs. It is mostly used in statistical and business surveys.

A **tally table** uses tally marks (grouped in sets of five) to count the frequency of each item or class.

Example: Number of Students Scoring in Ranges

Marks Range	Tally Marks	Frequency
0-10	||	2
11-20	||||	4
21-30	𝍸 ||	7
31-40	𝍸	5
41-50	𝍸 ||||	9

This form of tabulation is useful for summarizing raw survey data and is easy to construct during manual data collection.

1.26 Presentation of Data

After data has been collected and arranged in a table, the next important step is to present the data in a visual format using diagrams or charts. A diagrammatic representation helps the reader to grasp the data easily and quickly. Presentation of data means converting raw and tabulated information into diagrams such as bar charts, pie charts, line graphs, and others. These visual tools are especially useful in business settings, as they convey numerical information clearly in reports, meetings, and advertisements.

Importance of Presenting Data Visually

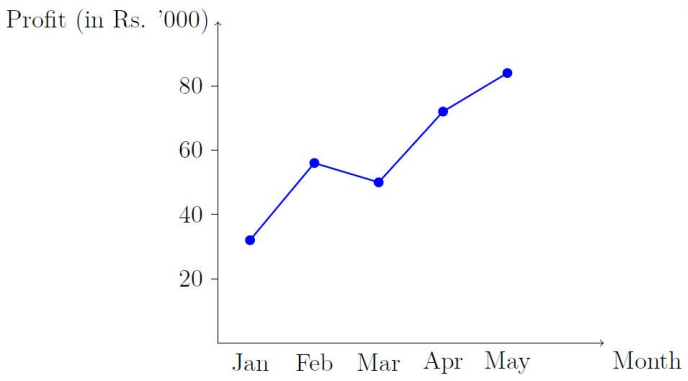
- **Easy Understanding:** Visuals communicate facts much faster than numerical tables. Even laypersons can understand the message.
- **Quick Comparison:** Charts help in comparing different sets of data easily.

- **Attractive Format:** Diagrams make the data interesting and attention-catching.
- **Spotting Trends:** Graphs can help in identifying upward or downward trends over time.
- **Useful in Business:** Business managers and investors use visual presentations to make decisions.

Example: Consider the monthly profit of a company in the first five months of a year.

Month	Profit (in Rs. '000)
January	40
February	55
March	50
April	65
May	75

If we convert this into a line graph, the increasing trend in profit becomes visually clear and easier to interpret than reading values from the table.

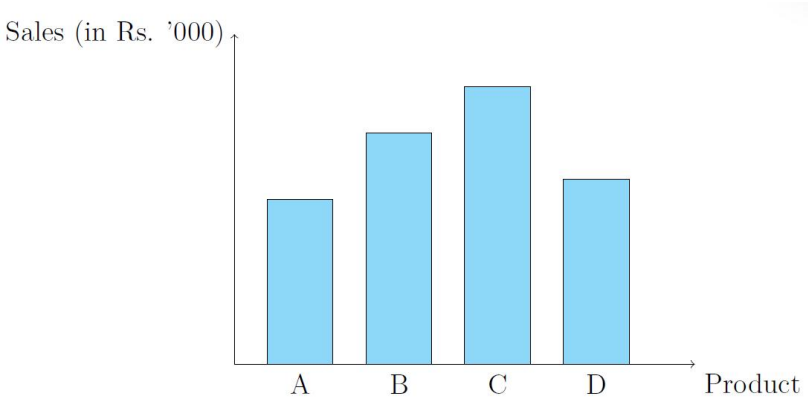


1.27 Need for Diagrammatic Representation

- **To Convey Information Quickly:** In meetings or reports, people may not have time to read through detailed data tables.

- **To Communicate with Non-Technical Audiences:** Not everyone understands numbers well, but diagrams speak a universal language.
- **To Support Decision-Making:** Managers and entrepreneurs use graphs to decide on sales targets, budgeting, and investments.
- **To Present Summary of Data:** Diagrams provide a summary of key findings in a single picture.
- **To Highlight Key Messages:** Important points stand out better in a visual format than in long rows of numbers.

Example: In an annual report of a retail company, showing sales trends using a bar chart makes the changes more visible than giving just raw figures.



This bar chart shows that Product C has the highest sales. The vertical axis is properly labeled with units (in Rs. ' 000), and the bars are equal in width and correctly proportioned.

1.28 General Rules for Constructing Diagrams

While diagrams are useful, they must follow certain principles to ensure they are meaningful and accurate:

1. **Title:** Every diagram should have a suitable title explaining what it represents.
2. **Scale:** A proper scale must be chosen so that the entire data fits well within the diagram area.
3. **Neatness and Simplicity:** Avoid too much clutter. Keep diagrams simple and easy to follow.
4. **Correct Labeling:** Axes, categories, and data points should be labeled clearly.

5. **Proportion:** All elements, like bars or pie segments should be proportionate to the data.

6. **Source and Units:** Mention the source of data and units used (like Rs., kg)

7. **Avoid Misleading Graphics:** Do not use 3D effects or incorrect proportions that distort the facts.

Important Tip: In exams and practical work, always use a pencil and ruler for drawing neat diagrams. In digital reports, use tools like Excel or MS-Paint etc.

1.29 Basic Charts and Their Suitability

Charts and diagrams help to present data in a simple, attractive, and meaningful way. Each type of chart has a specific purpose and should be chosen based on the type of data and the objective of presentation.

Let us understand the most commonly used charts and where they are suitable.

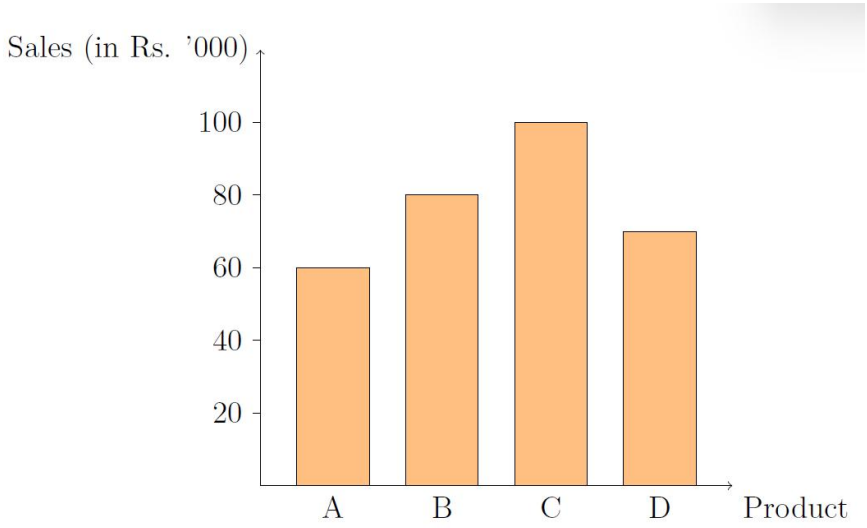
Bar Chart

A bar chart represents data with the help of rectangular bars. The height (or length) of the bar corresponds to the magnitude of the data. Bars can be drawn vertically or horizontally and should have equal width and equal spacing between them.

Uses:

- To compare quantities across different categories
- Suitable for discrete data
- Commonly used in business for showing product sales, customer responses, etc.

Example: Monthly sales (in Rs.’000) of four products A, B, C, and D are 60, 80, 100, and 70, respectively.



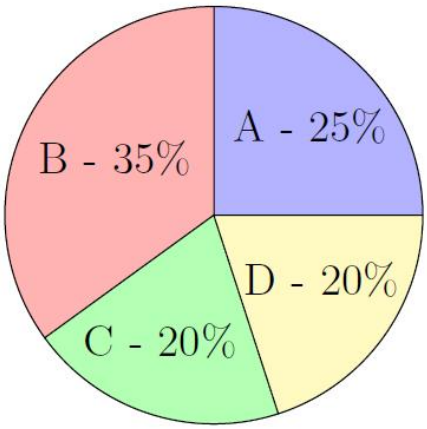
Pie Chart

A pie chart is a circular diagram divided into sectors. Each sector shows a proportion of the total data. The full circle represents 100%.

Uses:

- Suitable for showing percentage or part-whole relationships
- Common in financial reports, budgets, market share distribution, etc.

Example: Market share distribution of four companies: A - 25%, B - 35%, C - 20%, D - 20%.



Line Graph

A line graph is used to display data that changes over time. Points representing values are plotted on a graph and connected by straight lines.

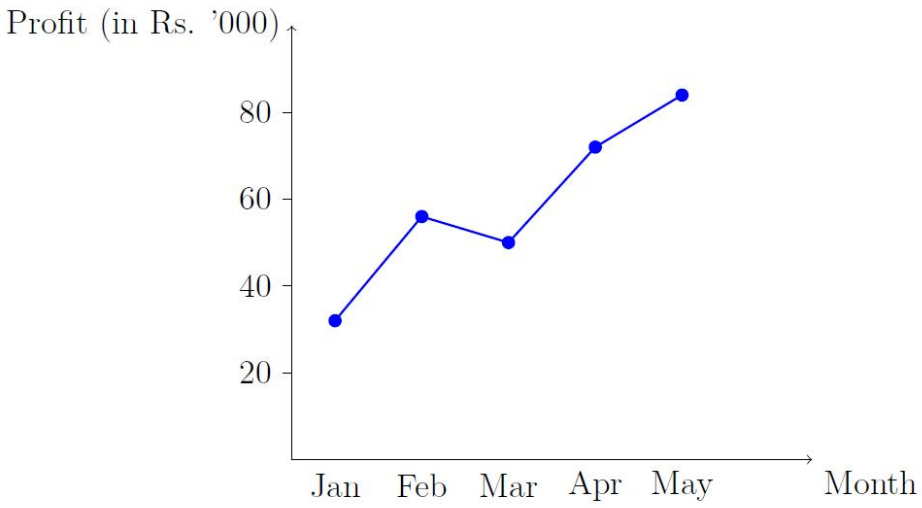
Uses:

- Best for time-series data like profits over months, temperatures over days, sales over quarters, etc.
- Useful to show trends and patterns

Example: Consider the monthly profit of a company in the first five months of a year.

Month	Profit (in Rs. '000)
January	40
February	55
March	50
April	65
May	75

If we convert this into a line graph, the increasing trend in profit becomes visually clear and easier to interpret than reading values from the table.



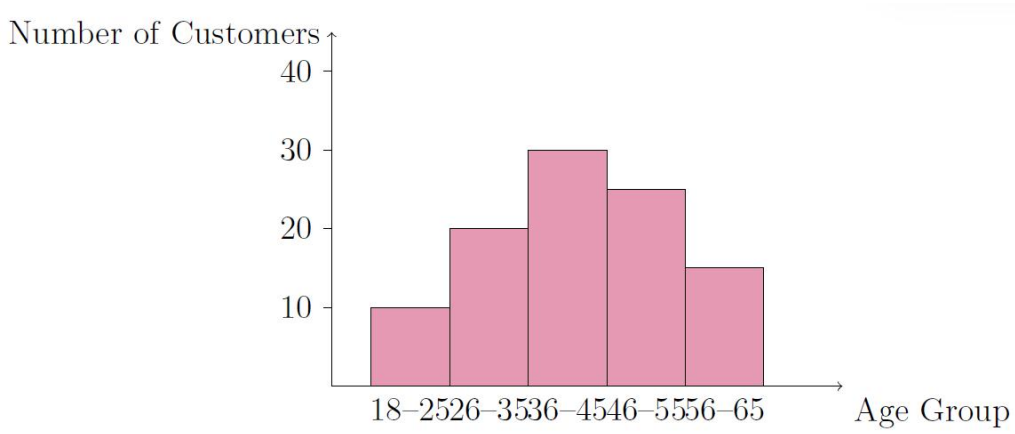
Histogram

A histogram looks similar to a bar chart but is used for continuous data. The bars are joined together without gaps.

Uses:

- Shows frequency distribution for grouped continuous data
- Common in business to analyse customer age groups, salary levels, etc.

Example: Number of customers grouped by their age range.



Note: In a histogram, since the data is continuous, there are no gaps between the bars.

Frequency Polygon

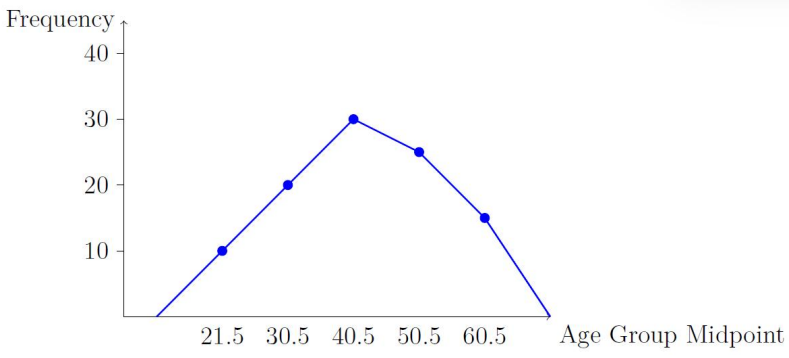
A frequency polygon is a line graph obtained by joining the midpoints of the tops of the bars in a histogram. It is used to understand the shape and spread of the data.

Uses:

- Represents the frequency distribution in a clear, continuous line
- Helps in comparing two or more frequency distributions

Example:

Use midpoints of the age group to draw histogram



a frequency polygon.

The frequency polygon is constructed using the midpoints of class intervals. Additional points at the beginning and end are included at frequency zero to close the shape.

Ogive Curves (Cumulative Frequency Curves)

An Ogive is a graph used to show cumulative frequencies. It helps us determine how many values fall below or above a particular class boundary.

There are two types of Ogive curves:

1. **Less Than Ogive:** Plots cumulative frequency against upper class boundaries.
2. **Greater Than Ogive:** Plots cumulative frequency against lower class boundaries.

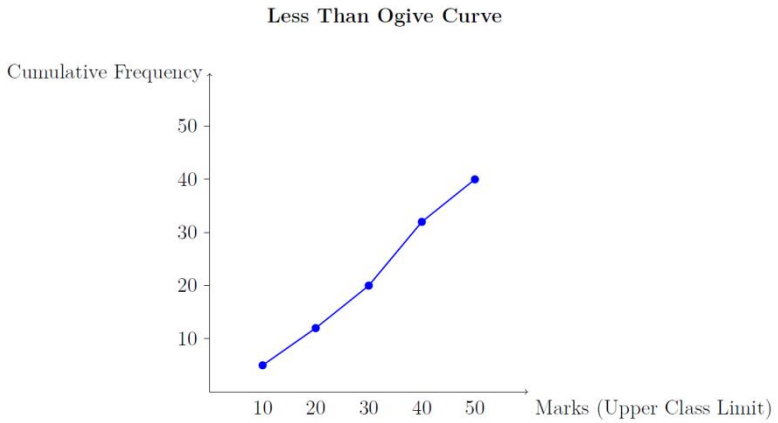
Less Than Ogive Curve

The Less Than Ogive is constructed by plotting cumulative frequency against the **upper boundary** of each class.

Example: The following table shows the marks obtained by 40 students.

Marks Range	Frequency	Cumulative Frequency (Less Than)
0-10	4	4
10-20	6	10

20-30	10	20
30-40	12	32
40-50	8	40

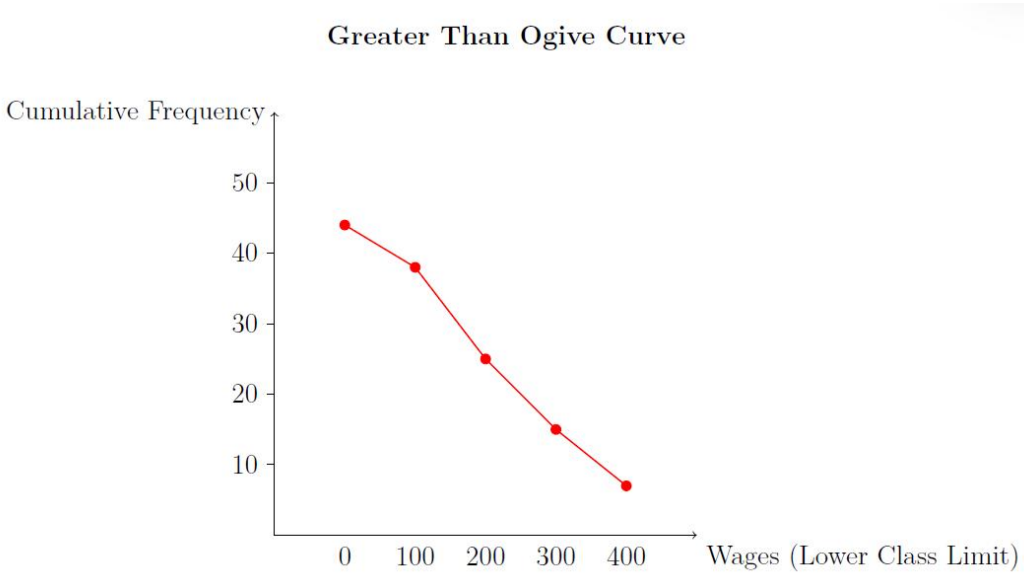


Greater Than Ogive Curve

The Greater Than Ogive is constructed by plotting cumulative frequencies against the **lower boundary** of each class.

Example: The following frequency distribution shows daily wages of 35 workers.

Wages Range (Rs.)	Frequency	Cumulative Frequency (Greater Than)
0-100	5	35
100-200	8	30
200-300	10	22
300-400	7	12
400-500	5	5



Intersection of Both Curves: If we draw both ogives on the same graph, the point where they intersect gives the median of the data.

Summary Table

Chart Type	Suitable for
Bar Chart	Comparing values across categories such as product sales, expenses, survey responses (discrete data).
Pie Chart	Showing proportionate distribution of a whole such as budget allocations or market share.
Line Graph	Representing trends over time such as monthly profits, stock prices, or temperature changes.
Histogram	Showing frequency distribution of continuous data like ages, income groups, or time intervals.
Frequency Polygon	Displaying the shape of a frequency distribution; helps compare distributions using a line graph.
Ogive Curve (less than / greater than)	Analysing cumulative frequencies; useful for determining medians, percentiles, and growth trends.

Check Your Progress

Answer the following questions to assess your understanding of the concepts discussed in this unit.

Part A: Objective Type Questions (Choose the Correct Answer)

1. Which of the following is not a part of a statistical table?

(a) Stub

(b) Caption

(c) Legend

(d) Title

2. A pie chart is most suitable for:

(a) Comparing trends over time

(b) Showing part-to-whole relationships

(c) Representing frequency distribution

(d) Comparing two variables

3. Which chart is best for showing the cumulative frequency?

(a) Histogram

(b) Bar chart

(c) Ogive

(d) Pie chart

4. In a frequency polygon, we plot:

(a) Frequencies on both axes

(b) Midpoints vs Frequencies

(c) Class width vs Class limits

(d) Time vs Cumulative frequency

Part B: Short Answer Questions

1. Define tabulation. What are the main objectives of tabulating data?

2. What is the difference between a bar chart and a histogram?

3. List the components of a well-structured statistical table.

4. Mention any three uses of diagrammatic data presentation in business.

5. What is a frequency polygon? How is it constructed?

Part C: Draw and Interpret

1. Using the following data, draw a bar chart:

Sales of Four Products (in Rs. ’000): A – 50, B – 70, C – 90, D – 60

2. Draw a frequency polygon for the following frequency distribution:

Class	Frequency
0-10	4
10-20	6
20-30	10
30-40	7
40-50	3

3. The following cumulative frequencies are given. Draw a less than Ogive and find the approximate median:

Marks (Less than)	Cumulative Frequency
10	5

20	15
30	27
40	35
50	40

Answers to Check Your Progress

Part A: Objective Type Questions

1. (c) Legend

2. (b) Showing part-to-whole relationships

3. (c) Ogive

4. (b) Midpoints vs Frequencies

Part B: Short Answer Questions

1. **Definition of Tabulation:** Tabulation is the process of organizing data into rows and columns to make it easier to read, compare, and analyze.

Objectives:

• To simplify large data sets

• To enable easy comparison

• To help in further statistical processing

2. **Difference between Bar Chart and Histogram:**

• Bar charts are used for discrete data and have spaces between bars.

• Histograms are used for continuous data and have no gaps between bars.

3. **Components of a Statistical Table:**

• Table number

• Title

• Captions (column headings)

• Stubs (row headings)

• Body (data)

• Footnote (if any)

• Source

4. **Uses of Diagrams in Business:**

• Easy to interpret for reports and presentations

• Useful in comparing business data

• Help in decision-making by showing trends

5. **Frequency Polygon:** It is a line graph that is formed by joining the midpoints of the tops of histogram bars. It gives a clear picture of the distribution shape.

Part C: Suggested Approach to Drawing

1. **Bar Chart for Product Sales:**

• Draw X-axis as Product (A, B, C, D)

• Y-axis as Sales (scale appropriately)

• Use bars of equal width with heights 50, 70, 90, and 60 units respectively

2. **Frequency Polygon:**

• Calculate midpoints of each class (e.g., for 0 – 10 it’ s 5, for 10 – 20 it’ s 15, etc.)

• Plot midpoints on X-axis and frequency on Y-axis

• Join the points with straight lines; extend to zero frequency at ends

3. **Less Than O give:**

• Plot cumulative frequency against upper class limits

• Join points with a smooth curve

• To find median: draw horizontal line from 50% of total frequency and drop vertically to X-axis.

Let Us Sum Up

In this unit, we have learned how to organize and present data effectively using tabulation and diagrams. These are the key takeaways:

- **Tabulation** is the process of arranging data in rows and columns. It helps in summarizing large amounts of data in a compact and readable form.
- A good statistical table has important components such as table number, title, captions (column headings), stubs (row headings), the body (data), footnotes, and source.
- Tabulated data can be further presented using various diagrams and charts for easy interpretation and communication.
- **Bar Charts** are used for comparing quantities among different categories or products.
- **Pie Charts** are circular diagrams used to show parts of a whole, commonly in percentage terms.
- **Line Graphs** are best for showing trends or changes over time, such as profit or sales over months.
- **Histograms** are used for representing frequency distributions of continuous data. Unlike bar charts, the bars in histograms are joined.
- **Frequency Polygons** are line graphs created by joining the midpoints of the histogram bars. They give a clearer idea of the shape of the distribution.
- **Ogive Curves** (Less Than and Greater Than) are used for plotting cumulative frequencies. They help in identifying the median and understanding data spread.
- Proper construction of diagrams requires neatness, appropriate scales, correct labeling, and accuracy to avoid misrepresentation of facts.

By mastering tabulation and diagrammatic presentation, you can effectively summarize, compare and communicate business data, which is essential in today's competitive commerce environment.

1.30 Measures of Central Tendency

Overview

Measures of central tendency are statistical tools used to identify the central or typical value in a data set. These measures, including the mean, median, mode, geometric mean, harmonic mean, and weighted arithmetic mean, help summarize large datasets with a single representative value. Understanding these measures is essential for data analysis, interpretation, and decision-making in fields such as economics, psychology, and social sciences.

Learning Objectives

By the end of this part, students will be able to:

- Define and explain the concept of central tendency.
- Identify the characteristics of a typical average.
- Compute the mean, median, mode, geometric mean, harmonic mean, and weighted arithmetic mean.

- Compare the suitability of different measures for various types of data.
- Apply these measures to real-world datasets for analysis.

1.31 Introduction to Measures of Central Tendency

Central tendency refers to the statistical measure that represents the center point or typical value of a data set. The primary measures are:

- **Mean**: The arithmetic average of all data points.
- **Median**: The middle value in an ordered data set.
- **Mode**: The most frequently occurring value in a data set.
- **Geometric Mean**: The nth root of the product of all values, used for multiplicative datasets.
- **Harmonic Mean**: The reciprocal of the arithmetic mean of reciprocals, useful for rates and ratios.
- **Weighted Arithmetic Mean**: An average where each value is assigned a specific weight.

These measures provide insights into the distribution and trends within data.

1.32 Characteristics of a Typical Average

A good measure of central tendency should possess the following characteristics:

1. **Representative**: It should closely reflect the entire data set.
1. **Uniqueness**: It should provide a single, unambiguous value.
2. **Simplicity**: It should be easy to understand and compute.
3. **Stability**: It should not fluctuate significantly with minor changes in data.
4. **Applicability**: It should be suitable for the type of data (e.g., mean for interval data, mode for nominal data).

1.33 Computation of Mean

The **arithmetic mean** is calculated as:

$$Mean = \overline{X} = \frac{\sum x}{n}$$

Example 1: For the dataset 5,7,9,11,13:

Solution:

$$\text{Mean} = \frac{\sum x}{n} = \frac{5+7+9+11+13}{5} = 9$$

Discrete Data

In discrete data, values are individual numbers (not grouped).

Formula:

$$Mean = \overline{X} = \frac{\sum fx}{N}, where N = \sum f$$

Example 2

Obtain Mean for the following data

Marks (x): 2, 4, 6, 8

Frequency (f): 3, 5, 2, 4

Solution:

x	2	4	6	8
f	3	5	2	4
fx	6	20	12	24

$\Sigma f = N = 3 + 5 + 2 + 4 = 14$

$\Sigma(fx) = (2\times3) + (4\times5) + (6\times2) + (8\times4) = 6 + 20 + 12 + 32 = 70$

Mean = $\overline{X} = \frac{\sum fx}{N} = 70 / 14 = 5$

Example 3

Obtain Mean for the following data

Books (x): 1, 2, 3, 4

Students (f): 2, 3, 5, 4

Solution

X	1	2	3	4
f	2	3	5	4
fx	2	6	15	16

$\Sigma f = 2 + 3 + 5 + 4 = 14$

$\Sigma(fx) = (1\times2) + (2\times3) + (3\times5) + (4\times4) = 2 + 6 + 15 + 16 = 39$

Mean = $\overline{X} = \frac{\sum fx}{N} = 39 / 14 \approx 2.79$

Continuous Data

In continuous data, values are in class intervals.

Steps:

1. Find the midpoint (m) of each class: m = (lower limit + upper limit) / 2

2. Use the formula: $\overline{X} = \frac{\sum fm}{N}$

Example 4

Obtain Mean for the following data

Class Intervals: 0–10, 10–20, 20–30, 30–40

Frequency (f): 5, 8, 12, 5

Solution:

Class interval	0-10	10-20	20-30	30-40
f	5	8	12	5
midpoint(m)	5	15	25	35
fm	25	120	300	175

$\Sigma f = N = 5+8+12+5 = 30$

$\Sigma(f \cdot m) = (5\times5) + (8\times15) + (12\times25) + (5\times35) = 620$

Mean = $\overline{X} = \frac{\sum fm}{N} = 620 / 30 \approx 20.67$

Example 5

Obtain Mean for the following data

Class Intervals: 5–15, 15–25, 25–35, 35–45

Frequency (f): 4, 6, 10, 5

Solution

Class interval	5-15	15-25	25-35	35-45
F	4	6	10	5
Midpoint(m)	10	20	30	40
Fm	40	120	300	200

Midpoints (m): 10, 20, 30, 40

$\Sigma f = N = 4+6+10+5 = 25$

$\Sigma(fm) = (4\times10) + (6\times20) + (10\times30) + (5\times40) = 660$

Mean $= \overline{X} = \frac{\sum fm}{N}$ =660 / 25 = 26.4

1.34 Computation of Median

The **median** is the middle value in an ordered data set. For an odd number of observations, it is the central value. For an even number, it is the average of the two middle values.

Example:

- Odd dataset 3,5,7: Median = 5
- Even dataset 2,4,6,8: Median = $\frac{4+6}{2} = 5$

Discrete Data

In discrete data, the median is the middle value when data is arranged in ascending order.

Steps to find the Median:

1. Arrange the data in ascending order (if not already).

2. Find the cumulative frequency.

3. Use the formula:

If N is odd, Median = value of (N + 1)/2 th item

If N is even, Median = value of N/2 th item

Example 6

Obtain median for the following data

Marks (x): 2, 4, 6, 8

Frequency (f): 3, 5, 2, 4

X	2	4	6	8
f	3	5	2	4
cf	3	8	10	14

N = 14 → Even → Median = average of 7th and 8th item → Both lie in 4 marks class

So, Median = 4

Continuous Data

In continuous data, the data is grouped in class intervals.

Steps to find the Median:

1. Calculate cumulative frequency (CF).

2. Find N = total frequency.

3. Find N/2 and locate the class containing it (Median Class).

4. Use the formula:

$$\text{Median} = L + \frac{\frac{N}{2} - CF}{f} * h$$

Where:

L = lower boundary of median class

CF = cumulative frequency before median class

f = frequency of median class

h = class width

Example 7

Obtain Median for the following data

Class Interval: 0–10, 10–20, 20–30, 30–40

Frequency (f): 5, 8, 12, 5

Solution:

Class interval	0-10	10-20	20-30	30-40
Frequency	5	8	12	5
Cumulative frequency	5	13	25	30

N = 30 → N/2 = 15 → 15 lies in 20–30 class (Median class)

L = 20, CF = 13, f = 12, h = 10

$$\text{Median} = L + \frac{\frac{N}{2} - CF}{f} * h$$

$$= 20 + \frac{\frac{39}{2} - 13}{12} * 10$$

$$= 20 + 1.67$$

$$= 21.67$$

So, Median ≈ 21.67

1.35 Computation of Mode

The **mode** is the most frequently occurring value in a data set. A data set may have one mode (uni modal), multiple modes (multi-modal), or no mode if all values are unique. It is a measure of central tendency used in statistics to represent the most common or typical value in a distribution. The mode can be calculated for individual, discrete, and continuous series.

Example: For2,3,3,5,7, the mode is 3.

Individual Data

In an individual series, mode is simply the value that occurs most often. If no value repeats, the data set is said to have no mode. If multiple values repeat with the same highest frequency, the series is multi modal.

Example 8

Find mode for the following data: 5, 7, 7, 8, 10, 7, 9

Solution:

Here, 7 occurs three times, more than any other value. Hence, Mode = 7

Discrete Data

In a discrete series, each value is associated with a frequency. The mode is the value that has the highest frequency.

Example 9

Find the mode for the data:

x: 1, 2, 3, 4, 5

f: 2, 4, 6, 3, 1

Solution:

Here, the highest frequency is 6 corresponding to x = 3. Hence, Mode = 3

Continuous Data

In a continuous series, mode is calculated using a formula. The modal class is the class interval with the highest frequency. The mode is then calculated using the following formula:

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

Where:

L = Lower boundary of the modal class

f_1 = Frequency of the modal class

f_0 = Frequency of the class preceding the modal class

f_2 = Frequency of the class succeeding the modal class

h = Class width

Example 10

Calculate mode for the data given below

Class Interval: 0–10, 10–20, 20–30, 30–40, 40–50

Frequency: 5, 8, 12, 7, 3

Solution:

Modal class = 20–30 (highest frequency = 12)

L = 20, $f_1 = 12$, $f_0 = 8$, $f_2 = 7$, h = 10

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

$$= 20 + \left[\frac{(12 - 8)}{2*12 - 8 - 7}\right] \times 10$$

$$= 20 + (4 / 9) \times 10 \approx 24.44$$

Important Notes

- Mode is useful in describing categorical or nominal data.

- Unlike mean, mode is not affected by extreme values.

- Mode may not exist or may not be unique. A distribution can be uni-modal, bimodal, or multi-modal.

1.36 Computation of Geometric Mean

Definition and Formula

The Geometric Mean of a set of n positive values is the nth root of their product. It is denoted by G.M.

Formula:

If $x_1, x_2, ..., x_n$ are n positive values, then:

$$G.M. = (x_1 * x_2 * \ldots * x_n)^{(1/n)}$$

In logarithmic terms:

log G.M. $= (1/n) * \Sigma \log x_i$

Then, G.M. = Antilog [(Σ log x) / n]

Individual Data

Geometric Mean can be calculated for discrete data if all values are positive. This applies to both ungrouped and grouped data.

Example 11

Compute Geometric mean for the data: 2, 4, 8

Solution:

G.M. $= (x_1 * x_2 * \ldots * x_n)^{(1/n)} = (2 * 4 * 8)^{(1/3)} = (64)^{\wedge}(1/3) = 4$

Important Notes

- Geometric Mean cannot be used if any value is zero or negative.

1.37 Computation of Harmonic Mean

The **harmonic mean** is suitable for rates and is calculated as:

$$\text{Harmonic Mean} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Example12:

Compute Harmonic mean for 1,2,4:

Solution

$$\text{Harmonic Mean} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

$$= \frac{3}{\frac{1}{1}+\frac{1}{2}+\frac{1}{4}} = \frac{3}{1.75} \approx 1.71$$

Individual Data

Harmonic Mean can be calculated for discrete data (either ungrouped or grouped), as long as all values are non-zero.

Example 13

Determine Harmonic mean for the data: 2, 4, 6

Solution:

$$\text{Harmonic Mean} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

$$= \frac{3}{\frac{1}{2}+\frac{1}{4}+\frac{1}{6}} = \frac{3}{0.9167} \approx 3.27$$

Important Notes

H.M. is suitable for averaging rates like speed, time, or other reciprocal-based values.

It is not defined if any value is zero, as division by zero is undefined.

H.M. is always less than or equal to the arithmetic mean.

1.38 Computation of Weighted Arithmetic Mean

The **weighted mean** accounts for the importance of each value and is calculated as:

$$\text{Weighted Mean} = \frac{\sum_{i=1}^{n} (w_i \times x_i)}{\sum_{i=1}^{n} w_i}$$

Example 14

Compute weighted Arithmetic mean for the Values 5,10 with weights 2.3:

Solution:

$$\text{Weighted Mean} = \frac{\sum_{i=1}^{n}(w_i \times x_i)}{\sum_{i=1}^{n} w_i}$$

$$\text{Weighted Mean} = \frac{(5\times 2)+(10\times 3)}{2+3} = 8$$

1.39 Conclusion

Measures of central tendency are fundamental to data analysis, each serving unique purposes depending on the data set. While the mean is widely used, the median and mode offer robustness against outliers. The geometric and harmonic means are specialized for multiplicative and rate data, respectively. Mastery of these measures enables accurate data interpretation and informed decision-making.

Summary

In summary, statistics serve as a foundational element in various sectors, enabling data-driven decisions that impact our daily lives. From forecasting the weather to shaping public policy, the application of statistical analysis is integral part to understanding and improving the world around us.

Unit II: Measures of Dispersion

Structure

Introduction

Learning Objectives

- Properties of Good Measures of dispersion
- Absolute versus Relative Measures of dispersion
- Common Measures of dispersion
  - Range
  - Quartile Deviation
  - Mean Deviation
  - Standard Deviation
  - Co-efficient of Variation
- Measures of Skewness

Introduction

We understand how to summarize the data into a single representative value. However, that value does not reflect the variability in the data. This lesson will look at measures that attempt to quantify the variability of data. It is evident that averages attempt to describe only one feature of a distribution, namely the representative size of the values. To better grasp it, you should also be aware of the value distribution.

Dispersion refers to how far values in a distribution deviate from the distribution's mean.

Learning Objectives

This topic serves the following objects.

- Compute the Range
- Compute Quartile deviation, (for un-grouped and grouped data)
- To determine the reliability of an average
- To compare the changeability of different distributions
- To control the variability

2.1 Properties of Good Measures of Dispersion

A good measure of dispersion should have the following properties:

- Simple and strict definition
- Easy computation.
- The model should be based on all values, capable of additional algebraic treatment, have sampling stability, and be unaffected by extreme items.

2.2 Absolute measures of dispersion versus relative measures of dispersion

Measures of dispersion might be absolute or relative. Absolute measures of dispersion are expressed using the same statistical unit as the original data, such as rupees, kilograms, or tonnes. A relative measure of dispersion is defined as the ratio of one measure of dispersion to an appropriate average. It is also known as a coefficient of dispersion, because the term "coefficient" refers to a pure number that is independent of the unit of measurement.

2.3 Common Measures of Dispersion

2.3.1 Range

The difference between a distribution's largest (L) and smallest (S) values is its range (R). Consequently, $R = L - S$

Higher dispersion is implied by higher range values, and vice versa.

The coefficient of range, which is the appropriate relative measure, is provided by

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

Example1: Find the value of the range.

$$20,\ 30,\ 40,\ 50,\ 200$$

Solution:

Here, $L = 200$, $S = 20$ and $R = L - S = 200 - 20 = 180$,

Therefore Range$(R) = 180$

- Find the Range if the value 200 is not present in the data set?

If 200 is not in data set, New data set is 20, 30, 40, 50.

Here, L = 50, S = 20 and R = 50 - 20 = 30, Therefore Range(R) = 30

- If 50 is changed as 150, what will be the range of the above data?

New data set if 50 is replaced by 150 is 20, 30, 40, 150, 200

Here, L = 200, S = 20 and

R = L – S = 200 – 20 =180. Therefore Range(R) = 180

Merits:

- Range is easy to understand
- Easy to calculate

Demerits:

- No sampling stability
- Does not depend upon all the values

Computation of range:

(a) Discrete series

Example 2: The following are the marks obtained by six students in statistics. Find the range and coefficient of range.

Sr.No.	1	2	3	4	5	6
Marks	30	35	40	80	70	62

Solution:

Range = L - S, L = 80, S=30

Range = 80-30 =50Marks

Coefficient of range $=\frac{L-S}{L+S} = \frac{80-30}{80+30} = \frac{50}{110} = 0.4545$

(b) Continuous series

The range in a continuous distribution is the difference between the highest and lowest classes' midpoints.

Example 3:

Find the range and coefficient of range from the following data

Weight in lbs	80-90	90-100	100-110	110-120	120-130
No.of persons	4	8	12	14	7

Solution:

Range = L - S, but L = 125, the mid-point of highest class interval

and S=85, the mid-point of the lowest class interval.

Range = 125-85=40 lbs

Coefficient of range $= \frac{L-S}{L+S} = \frac{125-85}{125+85} = \frac{40}{210} = 0.1904$

Steps for calculating range

Determine the highest and lowest data in the series by sorting the data in ascending order.

Determine how the highest and lowest statistics differ from each other.

2.3.2 Quartile Deviation

The usefulness of range as a dispersion measure can be diminished if a distribution contains even one exceptionally high or low value. Therefore, you could require a measure that is not significantly impacted by the outliers. In this case, the quartiles and median values can be obtained by splitting the total data into four equal sections, each of which contains 25% of the values.

The difference between the third and first quartiles is known as the inter-quartile range, which is a measure of dispersion. Semi-inter-quartile range, also known as quartile deviation, is half of the inter-quartile range. Symbolically it is defined as;

$Q.D = (Q_3 - Q_1)/ 2$

Where Q1 and Q3 are the first and third quartiles of the data.

What do you mean by quartiles?

Another method of breaking down data is to use quartiles. One-fourth of the total population or group is represented by each quartile. One appealing aspect of the quartile deviation is that the range "median + Q.D" contains almost half of the data. Another absolute indicator of dispersion is the quartile deviation. The coefficient of quartile deviation, also known as the semi-inter-quartile range, is a relative measure. The relationship defines it as coefficient of quartile deviation = $(Q_3\text{-}Q_1)/(Q_3+Q_1)$

In a frequency distribution, the quartile deviation, abbreviated Q, is half the scale distance between the 75th and 25th percentiles.

The 25th percentile, or Q1, is the first quartile on the scoring scale, with 25% of the scores falling below it.

The 75th percentile, or Q3, is the third quartile on the scoring scale, with 75% of the scores falling below it.

To find Quartile deviation, we must first compute the Q3 and Q1. Thus:

Q.D. is therefore also called Semi-Inter Quartile Range.

Q.D. is absolute measure of dispersion.

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

The corresponding relative measure, called the coefficient of Quartile Deviation, is given by

$$Coeff.QD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Merits:

- Simple to understand
- Easy to calculate
- It is not changed by extreme values
- It is especially useful to measure the variation of a distribution with open end classes.

Demerits:

- Not all observations are incorporated.
- No extra algebraic treatment is possible.
- Results may be affected by sampling errors.

● The first and last 25% of items are overlooked.

Computation of Quartile Deviation: (*un-grouped data)*

Example 4: Calculate Range and Quartile Deviation of the following observations:

20, 25, 29, 30, 35, 39, 41, 48, 51, 60 and 70

Solution:

For Quartile Deviation, we need to calculate values of Q3 and Q1.

Here n =11.

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

$$Coeff.QD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

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

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

Q_1 (25 th percentile) $= Size\ of \left(\frac{n+1}{4}\right) th\ item$

$= \text{Size of } (11+1)/4$

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

So, $Q_1 = 29$

Q_3 (75 percentile) $= Size\ of\ 3\left(\frac{n+1}{4}\right) th\ item$

$= \text{Size of } [3(11+1)] / 4$

$= \text{Size of the } 9^{t} \text{ item}$

So ,$Q_3 = 51$

$$QD = \frac{Q_3 - Q_1}{2} = \frac{51 - 29}{2} = 11$$

$$Coeff.QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{51-29}{51+29} = \frac{22}{80}$$

$$= 0.275$$

Computation of Quartile Deviation: (*grouped data)*

Example 5: The following distribution of marks scored by a class of 40 students, Compute the quartile deviation and coefficient of Q.D.

Class	:	0-10	10-20	20-30	30-40	40-50
Frequency	:	5	8	16	7	4

Solution:

$Q_1 = Size\ of \left(\frac{N}{4}\right) th\ item$ $\qquad$ $Q_1 = L_1 + \left(\frac{\frac{N}{4} - c.f}{f}\right) X\ i$

$Q_3 = Size\ of\ 3\left(\frac{N}{4}\right) th\ item$ $\qquad$ $Q_3 = L_1 + \left(\frac{\frac{3N}{4} - c.f}{f}\right) X\ i$

Class	Frequency f	Cumulative Frequency(C.f)
0-10	5	5
10-20	8	13
20-30	16	29
30-40	7	36
40-50	4	40
	$\sum f = 40$	

$Q_1 = Size\ of \left(\frac{N}{4}\right) th\ item$

$Q_1 = Size\ of \left(\frac{40}{4}\right) th\ item$

$Q_1 = Size\ of\ 10th\ item$

Therefore Quartile class is10-20

Where, $L_1 = 10, c.f = 5, f = 8, i = 10 \; and \; N = 40$

$$Q_1 = L_1 + \left(\frac{\frac{N}{4} - c.f}{f}\right) X \; i$$

$$Q_1 = 10 + \left(\frac{10 - 5}{8}\right) X \; 10$$

$$Q_1 = 16.25$$

$Q_3 = Size \; of \; 3\left(\frac{N}{4}\right) th \; item$

$Q_3 = Size \; of \; 3\left(\frac{40}{4}\right) th \; item$

$Q_3 = Size \; of \; 30th \; item$

Therefore Quartile class is 30-40.

Where, $L_1 = 30, \; c.f = 29, \; f = 7, \; i = 10 \; and \; N = 40$

$$Q_3 = L_1 + \left(\frac{\frac{3N}{4} - c.f}{f}\right) X \; i$$

$$Q_3 = 30 + \left(\frac{30 - 29}{7}\right) X \; 10$$

$$Q_3 = 30 + \left(\frac{1}{7}\right) X \; 10$$

$$Q_3 = 31.4285$$

$$QD = \frac{Q_3 - Q_1}{2} = \frac{31.4285 - 16.25}{2} = 7.58925$$

$$Coeff.QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{31.4285 - 16.25}{31.4285 + 16.25} = \frac{15.1785}{47.6785} = 0.3183$$

LET US SUMUP

This chapter addressed the notion of dispersion and related materials. We explained

how to compute range and how it varies from lowest to highest value. We then dealt with quartile deviation and used formulate to work out the case.

Formulas:

1. Range(R) is the difference between the largest(L) and the smallest value(S) in a distribution. Thus, R = L – S
2. The corresponding relative measure, called the coefficient of range, is given by

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

Computation of Quartile Deviation: (*un grouped data)*

3. $QD = \frac{Q_3 - Q_1}{2}$
4. $Coeff.QD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

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

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

Computation of Quartile Deviation: (*grouped data)*

$Q_1 = Size\ of \left(\frac{N}{4}\right) th\ item$ $\qquad$ $Q_1 = L_1 + \left(\frac{\frac{N}{4} - c.f}{f}\right) X\ i$

$Q_3 = Size\ of\ 3\left(\frac{N}{4}\right) th\ item$ $\qquad$ $Q_3 = L_1 + \left(\frac{\frac{3N}{4} - c.f}{f}\right) X\ i$

Exercise:

1. Compute the range, quartile deviation, coefficient of quartile deviation from the following ungrouped data:
   a. 30,35,36,39,42,46,38,34,35
   b. 52,50,56,68,65,62,57,70

2. Calculate the Quartile deviation of the following scores:
   a)

Class Interval	Frequency
40-44	3
35-39	4
30-34	6
25-29	12
20-24	7
15-19	5
10-14	1
	N=38

b)

Class Interval	Frequency
50-59	6
40-49	3
30-39	5
20-29	8
10-19	4
0-9	2
	N=28

2.3.3 Mean Deviation (MD)

Mean Deviation, also called average deviation, is a measure of dispersion that shows how much the data values differ, on average, from the mean of the data set. It is calculated by taking the average of the absolute differences between each observation and the mean. This provides a clear idea of how spread out the data is around the central value.

A major advantage of mean deviation is its easy interpretation. Since it is expressed in the same units as the original data, it becomes simple for analysts, researchers, and decision-makers to understand the level of variation without worrying about changes in scale.

Mean Deviation is widely used as a statistical tool to understand the degree of spread in a dataset relative to its central value. Like other statistical measures, it comes with its own strengths and limitations.

Merits of Mean Deviation

Mean Deviation is considered useful because it addresses some drawbacks found in other dispersion measures. Its key merits include:

- Straightforward to compute
- Easy to understand and interpret
- Uses absolute deviations, avoiding complications of squared values
- Less influenced by extreme observations compared to variance and standard deviation
- Helpful for quick comparisons of variability between datasets

Demerits of Mean Deviation

Mean Deviation has limited use because it cannot be easily used for further algebraic calculations, which reduces its applicability in advanced statistical analysis. Some additional drawbacks include:

- It is not strictly defined, as it can be computed using the mean, median, or mode, leading to inconsistency.
- It is seldom used in sociological research, as this measure is not considered suitable for analyzing such types of data.

Note: Negative and positive signs are ignored because we take the absolute value. This can lead to inaccuracies in the result.

Formula to Calculate Mean Deviation

MD about mean $= \frac{\sum D}{n}$ $where\ D = |x - \overline{x}|$ (For Individual Observation)

MD about median $= \frac{\sum D}{n}$ $where\ D = |x - A|$, $A\ is\ the\ median$(For Individual Observation)

MD about mean $= \frac{\sum fD}{\sum f}$ $where\ D = |x - \overline{x}|$ (For grouped discrete data)

MD about median $= \frac{\sum fD}{\sum f}$ $where\ D = |x - A|$ A is the median (For grouped discrete data)

MD about mean $= \frac{\sum fD}{\sum f}$ $where\ D = |m - \overline{x}|$, m is the midpoint of the class interval

(For grouped continuous data)

MD about median $= \frac{\sum fD}{\sum f}$ $where\ D = |m - A|$, m is the midpoint of the class interval & A is the median

(For grouped continuous data)

Coefficient of Mean Deviation (CMD)

The coefficient of mean deviation (CMD) is the relative measure of dispersion corresponding to mean deviation and it is given by

Coefficient of MD $= \frac{MD(Mean\ or\ Median)}{Mean\ or\ Median}$

Problems:

Example1. Calculate the mean deviation about the mean for the data 3, 5, 7,9,11.

Solution:

First, find the mean

Mean $= \frac{\sum x}{n} = \frac{3+5+7+9+11}{5} = \frac{35}{5} = 7$

MD about mean $= \frac{\sum D}{n}$ $where\ D = |x - \overline{x}|$

$= \frac{|3-7|+|5-7|+|7-7|+|9-7|+|11-7|}{5} = \frac{12}{5} = 2.4$

Example 2: Calculate the mean deviation about the median for the data: 2, 4, 6, 8, 10, 12

Solution:

First find the median

Arrange in ascending order: 2,4,6,8,10,12

Number of observations = 6 (even)

Median=Average of 3rd and 4th terms $=\frac{6+8}{2}=7$

MD about median $=\frac{\sum D}{n}$ $where\ D=|x-A|\ ,A\ is\ the\ median$

$$=\frac{|2-7|+|4-7|+|6-7|+|8-7|+|10-7|+|12-7|}{6}=\frac{18}{6}=3$$

Example 3: The weights (in kilograms) of 10 children admitted to a hospital on a certain day are: 7, 4, 10, 9, 15, 12, 7, 9, 9, and 18.Calculate the mean deviation about the mean and about the median, and also determine their respective coefficients of mean deviation.

Solution:

$$\text{Mean}=\frac{\sum x}{n}=\frac{7+4+10+9+15+12+7+9+9+18}{10}=\frac{100}{10}=10$$

MD about mean $=\frac{\sum D}{n}$ $where\ D=|x-\bar{x}|$

$$=\frac{|7-10|+|4-10|+|10-10|+|9-10|+|15-10|+|12-10|+|7-10|+|9-10|+|9-10|+|18-10|}{10}=\frac{30}{10}=3$$

To find median Ascending order 4, 7, 7, 9, 9,9,10, 12, 15, 18

Median=Average of 5th and 6th terms $=\frac{9+9}{2}=9$

MD about median $=\frac{\sum D}{n}$ $where\ D=|x-A|$

$$=\frac{|4-9|+|7-9|+|7-9|+|9-9|+|9-9|+|9-9|+|10-9|+|12-9|+|15-9|+|18-9|}{10}=\frac{28}{10}=2.8$$

Coefficient of MD $=\frac{MD}{mean}=\frac{3}{10}=0.3$

Example 4: Calculate the mean deviation about the mean for the given data

x	12	9	6	18	10
f	7	3	8	1	2

Solution:

x	f	f.x	$\|x-\bar{x}\|=\|x-9.381\|$	$f.\|x-\bar{x}\|$
12	7	84	2.619	18.33

9	3	27	0.381	1.143
6	8	48	3.381	27.048
18	1	18	8.619	8.619
10	2	20	0.619	1.238
Total	**21**	**197**		**56.378**

Mean $= \frac{\sum fx}{\sum f} = \frac{197}{21} = 9.381$

MD about mean $= \frac{\sum fD}{\sum f}$ $where\ D = |x - \bar{x}|$

$= \frac{56.378}{21} = 2.684$

Example 5: Calculate the mean deviation about the mean for the given data

x	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**
f	**7**	**12**	**18**	**25**	**16**	**14**	**8**

Solution:

x	**f**	**m**	**f.m**	$\|m - \bar{x}\| = \|m - 35.5\|$	$f.\|m - \bar{x}\|$
0-10	7	5	35	30.5	213.5
10-20	12	15	180	20.5	246
20-30	18	25	450	10.5	189
30-40	25	35	875	0.5	12.5
40-50	16	45	720	9.5	152
50-60	14	55	770	19.5	273
60-70	8	65	520	29.5	236
Total	**100**		**3550**		**1322**

Mean $= \frac{\sum fm}{\sum f} = \frac{3550}{100} = 35.5$

MD about mean $= \frac{\sum fD}{\sum f}$ $where\ D = |m - \bar{x}|$, m is the midpoint of the class interval

$$=\frac{1322}{100} = 13.22$$

Example 6: Calculate the mean deviation about the median for the given data

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

Solution:

x	f	cf	$\|x - A\|= \|x - 12\|$	$f.\|x - \bar{x}\|$
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	**48**			**36**

Median= Average of $\frac{n}{2}$ & $\frac{n}{2}+1$ *th item* = Avg of 24th &25th item=12

MD about median $= \frac{\sum fD}{\sum f}$ *where* $D = |x - A|$

$$=\frac{36}{48} = 0.75$$

Example 7: Calculate the mean deviation about the median for the given data

x	15-25	25-35	35-45	45-55	55-65
f	12	6	9	4	2

Solution:

x	f	cf	m	$\|m - A\|= \|m - 32.5\|$	$f.\|m - A\|$
15-25	12	12	20	12.5	150

25-35	6	18	30	2.5	15
35-45	9	27	40	7.5	67.5
45-55	4	31	50	17.5	70
55-65	2	33	60	27.5	55
Total	**33**				**357.5**

Median class = Size of $\frac{n+1}{2}^{th} item$ = Size of17^{th} item =25-35

The cf value that is nearest to 17 is 18.

Thus, median class is 25 - 35.

$$Median = L + \left(\frac{\frac{N}{2} - cf}{f}\right) i = 25 + \left(\frac{\frac{33}{2} - 12}{6}\right) 10 = 32.5$$

MD about median $= \frac{\sum fD}{\sum f}$ $where\ D = |m - A|$, m is the midpoint of the class interval

$$= \frac{357.5}{33} = 10.833$$

2.3.4 Standard Deviation (SD)

In statistics, standard deviation is regarded as one of the most effective measures of dispersion. Dispersion refers to how much individual values differ from a central value—usually the arithmetic mean. Because of its accuracy and sensitivity, standard deviation is widely used and often preferred over other dispersion measures.

Standard deviation is computed as the square root of the variance of a dataset. This makes it a more precise indicator of spread than mean deviation, as it considers the squared differences between each observation and the mean.

A major advantage of standard deviation is its ability to reflect the overall variability of data in a detailed manner. Since larger deviations are squared during calculation, the measure becomes more sensitive to extreme values or outliers, making it a powerful tool for identifying unusual data points.

Limitations of Standard Deviation

- **Highly sensitive to outliers:**
  Compared to mean deviation, standard deviation gives more weight to extreme values. In skewed datasets or when outliers are not relevant, SD may give an exaggerated impression of variability, leading to incorrect interpretations.

- **Less intuitive to understand:**
  Because standard deviation is the square root of variance, the resulting value does not directly match the units of the original data. This sometimes makes interpretation more challenging, especially for non-statisticians.

Uses of Standard Deviation

Standard deviation is commonly used when:

- A more reliable and accurate measure of variability is needed, especially when the distribution is normal or nearly normal.
- Additional statistical techniques—such as correlation, regression, and tests of significance—need to be computed, as these methods rely on the standard deviation.

Mean Deviation vs. Standard Deviation

Both mean deviation and standard deviation measure the spread of data, but they differ in several ways:

Mean Deviation	Standard Deviation
Can be calculated using the mean, median, or mode.	Calculated only using the mean.
Uses absolute deviations.	Uses the square of deviations.
Less commonly used.	Most widely used measure of variability.
More suitable when the dataset contains many outliers (mean absolute deviation).	More suitable when outliers are few or minimal.

2.3.5 Variance

Variance is a measure that indicates how far the values in a dataset are spread out from the mean. It helps determine whether the observations are closely clustered around the average or widely scattered.

Formula to Find SD

For Individual Observation

$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}, where\ \bar{x}\ is\ the\ mean\ and\ n\ is\ the\ number\ of\ items.$

Assumed Mean Method

$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}, d = x - A, A\ is\ the\ assumed\ mean\ and\ n\ is\ the\ number\ of\ items$

For Grouped Data (Discrete)

$\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{\sum f}}$

Assumed Mean Method

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

Where $d = x - A, A\ is\ the\ assumed\ mean\ and\ n\ is\ the\ number\ of\ item$

For Grouped Data (Continuous)

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

where $d = \frac{m-A}{i}$ $m - midpoint\ of\ the\ class\ interval\ and\ 'i'is\ the\ width\ of\ the\ class\ interval$

Variance $= \sigma^2$

Example 8 : Consider a scenario where a researcher is studying the daily temperature fluctuations in a desert over a week. The recorded temperatures (in degrees Celsius) are as follows: 30, 35, 28,32,40,29 and 33.Find the standard deviation.

Solution:

$$\text{Mean} = \frac{\sum x}{n} = \frac{30+35+28+32+40+29+33}{7} = \frac{227}{7} = 32.4$$

$$\sigma(SD) = \sqrt{\frac{\sum (x-\bar{x})^2}{7}} = \sqrt{\frac{(30-32.4)^2+(35-32.4)^2+(28-32.4)^2+(32-32.4)^2+(40-32.4)^2+(29-32.4)^2+(33-32.4)^2}{7}}$$

$$= \sqrt{\frac{5.76+6.76+19.36+.16+57.76+11.56+.36}{7}} = \sqrt{\frac{101.72}{7}} = 3.812$$

By Assumed Mean Method

X	**d=x-A=x-32**	d^2
30	-2	4
35	3	9
28	-4	16
32	0	0
40	8	64
29	-3	9
33	1	1
	3	**103**

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{103}{7} - \left(\frac{3}{7}\right)^2} = 3.811$$

Example 9: The heights of 6 students in a class (in cm) are as follows:
150,160,155,165,158,162. Calculate the standard deviation of the heights.

Solution:

X	d=x-A =x-150	d^2
150	0	0
160	10	100
155	5	25
165	15	225
158	8	64
162	12	144
	50	**558**

$$\sigma = \sqrt{\frac{558}{6} - \left(\frac{50}{6}\right)^2} = \sqrt{23.555} = 4.853$$

Example 10: Calculate Standard deviation and variance from the following:

Marks	10	20	30	40	50	60
No. of Students	8	12	20	10	7	3

Solution:

x	f	fx	$(x-\bar{x})$= x-30.833	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
10	8	80	-20.833	434.0138	3472.1104
20	12	240	-10.833	117.3538	1408.2456
30	20	600	-0.833	0.69388	13.8776
40	10	400	9.167	84.033	840.33
50	7	350	19.167	367.3738	2571.6166

60	3	180	29.167	850.7138	2552.1414
	60	**1850**			**10858.3216**

$\text{Mean} = \frac{\sum fx}{\sum f} = \frac{1850}{60} = 30.833$

$\sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}} = \sqrt{\frac{10858.3216}{60}} = 13.4525$

Variance=180.9720

Example 11: Calculate Standard deviation and variance from the following:

Marks	**3.5**	**4.5**	**5.5**	**6.5**	**7.5**	**8.5**	**9.5**
No. of Students	**3**	**7**	**22**	**60**	**85**	**32**	**8**

Solution:

x	**f**	$\boldsymbol{d = (x - A)}$ $\boldsymbol{=(x - 6.5)}$	$\boldsymbol{fd}$	$\boldsymbol{fd^2}$
3.5	3	-3	-9	27
4.5	7	-2	-14	28
5.5	22	-1	-22	22
6.5	60	0	0	0
7.5	85	1	85	85
8.5	32	2	64	128
9.5	8	3	24	72
	217	**0**	**128**	**362**

$\sigma = \sqrt{\frac{362}{217} - \left(\frac{128}{217}\right)^2} = \sqrt{\frac{62170}{47089}} = \sqrt{1.3202} = 1.148$

Variance= 1.3202

Example 12: Calculate Standard deviation and variance from the following:

Marks	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**

No. of Students	**8**	**12**	**17**	**14**	**9**	**7**	**4**

Solution:

x	**f**	**m**	$\boldsymbol{d=\frac{(m-A)}{i}=\frac{(m-35)}{10}}$	$\boldsymbol{fd}$	$\boldsymbol{fd^2}$
0-10	8	5	-3	-24	72
10-20	12	15	-2	-24	48
20-30	17	25	-1	-17	17
30-40	14	35	0	0	0
40-50	9	45	1	9	9
50-60	7	55	2	14	28
60-70	4	65	3	12	36
Total	**71**			**-30**	**210**

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

$$= \sqrt{\frac{210}{71} - \left(\frac{-30}{71}\right)^2} \text{ (i)}$$

$$= \sqrt{\frac{14010}{5041}}(10) = 16.67$$

Variance$= \sigma^2 = 277.88$

Example 13: Calculate Standard deviation and variance from the following:

Marks	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**
No. of Students	**5**	**12**	**30**	**45**	**50**	**37**	**21**

Solution:

x	**f**	**m**	$\boldsymbol{d=\frac{(m-A)}{i}=\frac{(m-35)}{10}}$	$\boldsymbol{fd}$	$\boldsymbol{fd^2}$
0-10	5	5	-3	-15	45
10-20	12	15	-2	-24	48
20-30	30	25	-1	-30	30

30-40	45	35	0	0	0
40-50	50	45	1	50	50
50-60	37	55	2	74	148
60-70	21	65	3	63	189
	200			**118**	**510**

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2} \times i = \sqrt{\frac{510}{200} - \left(\frac{118}{200}\right)^2}\,(10) = \sqrt{\frac{88076}{40000}}\,(10) = 14.83$$

Variance$= \sigma^2 = 219.9289$

2.3.6 Co-efficient of Variation (CV)

The coefficient of variation (CV) is the relative measure corresponding to the standard deviation. Developed by Karl Pearson, it is used to compare the degree of variation or dispersion between two different data sets. Since standard deviation alone cannot be used for comparison when the datasets have different units or scales, the CV provides a better alternative.

The coefficient of variation is defined as the ratio of the standard deviation to the mean of the dataset. It is usually expressed as a percentage**,** making it more appropriate for comparing the consistency or variability of multiple data series.

Formula:

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

Advantages

- Useful for comparing two or more datasets, even when their units or measurements differ.
- Effective when datasets have very different means**,** allowing a fair comparison of variability.

Disadvantages

- When the mean is close to zero, the CV becomes highly unstable and very sensitive to even small changes in the mean.

- Minor variations in the mean can lead to large changes in the CV value.
- Cannot be applied to logarithmic values.
- Cannot be used to determine confidence intervals for the mean.

Note: The group for which the CV is less is more stable or more consistent.

Example 14: From the prices of shares X and Y given below, state which share is more stable in value?

X	**55**	**54**	**52**	**53**	**56**	**58**	**52**	**50**	**51**	**49**
Y	**108**	**107**	**105**	**105**	**106**	**107**	**104**	**103**	**104**	**101**

Solution:

x	**d_1=X-50**	**d_1^2**	**y**	**d_2=y-100**	**d_2^2**
55	5	25	108	8	64
54	4	16	107	7	49
52	2	4	105	5	25
53	3	9	105	5	25
56	6	36	106	6	36
58	8	64	107	7	49
52	2	4	104	4	16
50	0	0	103	3	9
51	1	1	104	4	16
49	-1	1	101	1	1
530	**30**	**160**	**1050**	**50**	**290**

Mean of x $= \bar{x} = \frac{\sum x}{n} = \frac{530}{10} = 53$

$\sigma_x = \sqrt{\frac{\sum {d_1}^2}{n} - \left(\frac{\sum d_1}{n}\right)^2} = \sqrt{\frac{160}{10} - \left(\frac{30}{10}\right)^2} = \sqrt{\frac{700}{100}} = 2.6457$

Mean of y $= \bar{y} = \frac{\sum y}{n} = \frac{1050}{10} = 105$

$\sigma_y = \sqrt{\frac{\sum {d_2}^2}{n} - \left(\frac{\sum d_2}{n}\right)^2} = \sqrt{\frac{290}{10} - \left(\frac{50}{10}\right)^2} = \sqrt{\frac{400}{100}} = 2$

$$CV\ of\ x = \frac{\sigma_x}{\bar{x}} \times 100 = \frac{2.6457}{53} \times 100 = 4.99$$

$$CV\ of\ y = \frac{\sigma_y}{\bar{y}} \times 100 = \frac{2}{105} \times 100 = 1.90$$

Since CV of y is less than CV of x, y is more stable.

Example 15: The scores of two batsmen A and B in ten innings during a certain season are given below:

A	**32**	**28**	**47**	**63**	**71**	**39**	**10**	**60**	**96**	**14**
B	**19**	**31**	**48**	**53**	**67**	**90**	**10**	**62**	**40**	**80**

Find, which of the two batsmen better scorer is and who is more consistent?

Solution:

x	**d_1=X-46**	**d_1^2**	**y**	**d_2=y-50**	**d_2^2**
32	-14	196	19	-31	961
28	-18	324	31	-19	361
47	1	1	48	-2	4
63	17	289	53	3	9
71	25	625	67	17	289
39	-7	49	90	40	1600
10	-36	1296	10	-40	1600
60	14	196	62	12	144
96	50	2500	40	-10	100
14	-32	1024	80	30	900
460	**0**	**6500**	**500**	**0**	**5968**

Mean of x $= \bar{x} = \frac{\sum x}{n} = \frac{460}{10} = 46$

$\sigma_x = \sqrt{\frac{\sum {d_1}^2}{n} - \left(\frac{\sum d_1}{n}\right)^2} = \sqrt{\frac{6500}{10} - 0} = \sqrt{650} = 25.49$

Mean of y $= \bar{y} = \frac{\sum y}{n} = \frac{500}{10} = 50$

$$\sigma_y = \sqrt{\frac{\sum {d_2}^2}{n} - \left(\frac{\sum d_2}{n}\right)^2} = \sqrt{\frac{5968}{10} - 0} = \sqrt{596.8} = 24.43$$

$$CV\ of A = \frac{\sigma_x}{\bar{x}} \times 100 = \frac{25.49}{46} \times 100 = 55.41$$

$$CV\ of\ B = \frac{\sigma_y}{\bar{y}} \times 100 = \frac{24.43}{50} \times 100 = 48.86$$

Since mean of B is greater than A, B is the better scorer. Since CV of B is less than CV of A, B is more stable.

2.3.7 Skewness

In addition to measures of central tendency and dispersion, we also need to have an idea about the shape of the distribution. Measure of skewness gives the direction and the magnitude of the lack of symmetry. Lack of symmetry is called skewness for a frequency distribution. If the distribution is not symmetric, the frequencies will not be uniformly distributed about the centre of the distribution.

Difference between Variance and Skewness

- Variance tells us about the amount of variability while skewness gives the direction of variability.
- In business and economic series, measures of variation have greater practical application than measures of skewness. However, in medical and life science field measures of skewness have greater practical applications than the variance.

Types of Skewness

- **Positive Skewness (Right Skew):** The tail of the distribution is longer on the right side, indicating more extreme values on the higher end.
- **Negative Skewness (Left Skew):** The tail of the distribution is longer on the left side, indicating more extreme values on the lower end.
- **Zero Skewness:** The distribution is symmetrical.

Karl Pearson's Coefficient of Skewness

This method is most frequently used for measuring skewness. The formula for measuring coefficient of skewness is given by

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

The value of this coefficient would be zero in a symmetrical distribution. If mean is greater than mode, coefficient of skewness would be positive otherwise negative. The value of the Karl Pearson's coefficient of skewness usually lies between ±1 for moderately skewed distribution.

Example 16: Calculate Karl – Pearson's coefficient of skewness for the following data:

25, 15, 23, 40, 27, 25, 23, 25, 20

x	**d=x-A =X-25**	d^2
25	0	0
15	-10	100
23	-2	4
40	15	225
27	2	4
25	0	0
23	-2	4
25	0	0
20	-5	25
223	**-2**	**362**

Mean $= \bar{x} = \frac{\sum x}{n} = \frac{223}{9} = 24.78$

The value 25 has the highest frequency (occurs thrice), hence mode is 25.

$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{362}{9} - \left(\frac{-2}{9}\right)^2} = \sqrt{40.17} = 6.337$

$S_k = \frac{Mean - Mode}{\text{S. D.}} = \frac{24.78 - 25}{6.337} = -0.0347$

Example 17 : Find the coefficient of skewness from the data given below:

Size	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Frequency	**7**	**10**	**14**	**35**	**102**	**136**	**43**	**8**

Solution:

x	**f**	**Fx**	**d=x-A =X-8**	**fd**	fd^2
3	7	21	-5	-35	175
4	10	40	-4	-40	160
5	14	70	-3	-42	126
6	35	210	-2	-70	140
7	102	714	-1	-102	102
8	136	1088	0	0	0
9	43	387	1	43	43
10	8	80	2	16	32
	355	**2610**	**-12**	**-230**	**778**

Mean $= \bar{x} = \frac{\sum fx}{\sum f} = \frac{2610}{355} = 7.352$

The value 8 has the highest frequency (136), hence mode is 8.

$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2} = \sqrt{\frac{778}{355} - \left(\frac{-230}{355}\right)^2}$ =$\sqrt{1.7717}$ = 1.337

$S_k = \frac{Mean - Mode}{S.D.} = \frac{7.352 - 8}{1.337} = -0.4846$

Example 18: Find the coefficient of skewness from the data given below:

Size	**12.5**	**17.5**	**22.5**	**27.5**	**32.5**	**37.5**	**42.5**	**47.5**
Frequency	**28**	**42**	**54**	**108**	**129**	**61**	**45**	**33**

Solution:

X	**f**	**Fx**	**d=x-A =X-27.5**	**fd**	fd^2

12.5	28	350	-15	-420	6300
17.5	42	735	-10	-420	4200
22.5	54	1215	-5	-270	1350
27.5	108	2970	0	0	0
32.5	129	4192.5	5	645	3225
37.5	61	2287.5	10	610	6100
42.5	45	1912.5	15	675	10125
47.5	33	1567.5	20	660	13200
	500	**15230**	**-12**	**1480**	**44500**

$$\text{Mean } = \bar{x} = \frac{\sum fx}{\sum f} = \frac{15230}{500} = 30.46$$

The value 32.5 has the highest frequency (129), hence mode is 32.5.

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2} = \sqrt{\frac{44500}{500} - \left(\frac{1480}{500}\right)^2} = \sqrt{80.234} = 8.957 = 8.96$$

$$S_k = \frac{Mean - Mode}{S.D.} = \frac{30.46 - 32.5}{8.96} = -0.2276$$

Example 19: Find the coefficient of skewness from the data given below:

x	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**	**70-80**
f	**5**	**6**	**11**	**21**	**35**	**30**	**22**	**11**

Solution:

X	**f**	**m**	**fm**	$\boldsymbol{d = \frac{m - 35}{10}}$	**fd**	$\boldsymbol{fd^2}$
0-10	5	5	25	-3	-15	45
10-20	6	15	90	-2	-12	24
20-30	11	25	275	-1	-11	11
30-40	21	35	735	0	0	0
40-50	35	45	1575	1	35	35
50-60	30	55	1650	2	60	120

60-70	22	65	1430	3	66	198
70-80	11	75	825	4	44	176
	141		**6605**		**167**	**609**

Mean $= \bar{x} = \frac{\sum fm}{\sum f} = \frac{6605}{141} = 46.84$

The class interval 40-50 has the highest frequency (35), hence modal class is 40-50.

$$Mode = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 40 + \frac{(35-21)}{(35-21)+(35-30)} \times 10 = 47.37$$

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2} \times i = \sqrt{\frac{609}{141} - \left(\frac{167}{141}\right)^2} \times 10 = \sqrt{2.916} \times 10 = 17.07$$

$$S_k = \frac{Mean - Mode}{S.D} = \frac{46.84 - 47.37}{17.07} = -0.031$$

Example 20: Find the coefficient of skewness from the data given below:

x	**0-5**	**5-10**	**10-15**	**15-20**	**20-25**	**25-30**	**30-35**	**35-40**
f	**2**	**5**	**7**	**13**	**21**	**16**	**8**	**3**

Solution:

X	**f**	**m**	**fm**	$\boldsymbol{d = \frac{m - 22.5}{5}}$	**fd**	$\boldsymbol{fd^2}$
0-5	2	2.5	5	-4	-8	32
5-10	5	7.5	37.5	-3	-15	45
10-15	7	12.5	87.5	-2	-14	28
15-20	13	17.5	227.5	-1	-13	13
20-25	21	22.5	472.5	0	0	0
25-30	16	27.5	440	1	16	16
30-35	8	32.5	260	2	16	32
35-40	3	37.5	112.5	3	9	27
	75		**1642.5**		**-9**	**193**

Mean $= \bar{x} = \frac{\sum fm}{\sum f} = \frac{1642.5}{75} = 21.9$

The class interval 20-25 has the highest frequency (21), hence modal class is 20-25.

$$Mode = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 20 + \frac{(21-13)}{(21-13)+(21-16)} \times 5 = 23.07$$

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2} \times i = \sqrt{\frac{193}{75} - \left(\frac{-9}{75}\right)^2} \times 5 = \sqrt{2.5589} \times 5 = 8$$

$$S_k = \frac{Mean - Mode}{S.D} = \frac{21.9 - 23.07}{8} = -0.146$$

Check Your Progress

1. Find the mean deviation (i) about the median (ii) coefficient of mean deviation for the data 3000,4000,4200,4400,4600,4800,5800.
2. Find the mean deviation (i) about the mean (ii) coefficient of mean deviation for the data 6, 7, 10, 12, 13, 4, 8, 12.
3. Find the mean deviation about mean for the following data:

x	3	5	6	8	10
f	7	3	4	5	1

4. Find the mean deviation about mean for the following data:

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	4	3	2	3	2

5. Find the mean deviation about median for the following data:

x	10	15	20	25	30	35
f	4	4	2	3	3	2

6. Find the mean deviation about median for the following data:

Class Interval	0-6	6-12	12-18	18-24	24-30
Frequency	8	10	12	9	5

7. Find the standard deviation for the following data: 25, 27,31,32,35

8. Find the standard deviation for the following data:

X	10	12	14	16	18	20	22
F	3	5	9	16	8	7	2

9. Find the standard deviation for the following data:

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	1	4	17	45	26	5	2

10. The temperature of two cities A and B in a winter season are given below. Find which city is more consistent in temperature changes?

A	18	20	22	24	26
B	11	14	15	17	18

11. From the observations 36, 32, 41, 38, 33, 37, 30, 35, 39, 35 calculate Karl Pearson's Coefficient of Skewness.

12. From the distribution given below, find out: Karl Pearson's Coeff. of Skewness

Height in cm	75	76	77	78	79	80	81	82	83
No. of Students	6	8	13	18	20	16	10	7	2

13. Calculate the Karl Person's Coefficient of Skewness from the following table:

Marks	70-80	80-90	90-100	100-110	110-120	120-130	130-140	140-150
Frequency	6	9	17	21	25	23	10	8

Answers

1. (i) 571.43 (ii)0.130

2. (i) 2.75 (ii) 0.305

3.1.9

4. 9.58

5.7.5

6.6.318

7.3.578

8.2.97

9.10.2

10. $\bar{x}$=22 ; CV_1= 12.85 ; $\bar{y}$=15; CV_2= 16.33 ; A is more consistent .

11.0.3

12. -0.14

13. -0.27

Let us Sum Up

- Mean deviation is a statistical measure used to give the average value of the absolute deviation with respect to the central point of the data.
- Mean deviation can be calculated about the mean, median, and mode.
- Mean deviation is less frequently used as compared to standard deviation.
- The square root of the average of the squared differences of data observations from the mean is called the standard deviation.
- Standard deviation is the positive square root of variance.
- Standard deviation is the indicator that shows the dispersion of the data points about the mean.
- The group for which the CV is less is more stable or more consistent.
- Skewness is the lack of symmetry and indicates lopsidedness of the curve.
- If in a distribution, the largest category of items does not occur at the centre of the distribution, but drifts to the left or to the right, then it is called a skewed distribution.

Unit -III Correlation Analysis

Chapter Objectives

At the end of the chapter, you should be able to

- Express quantitatively the degree and direction of the co-variation or association between two variables.
- Determine the validity and reliability of the co-variation or association between two variables.
- Provide a test of hypothesis to determine whether a linear relationship actually exists between the variables.

3.1 Introduction

The statistical methods discussed until now deal with analyzing data that involves only one variable. However, in many cases, it becomes necessary to examine data with two or more quantitative variables to find out if there is any relationship or connection between them. Studying such relationships helps describe how the variables are numerically related. Understanding these connections is useful for making conclusions or predictions in different situations. Here are a few examples where knowing the relationship between two variables can support better decision-making:

- Family income and expenditure on luxury items.
- Yield of a crop and quantity of fertilizer used.
- Sales revenue and expenses incurred on advertising.
- Frequency of smoking and lung damage.

In all these situations, to study how strong the relationship is between two variables, a statistical method called correlation analysis is used. Correlation analysis focuses on how two or more variables change together and how closely they are related. When this relationship can be described with numbers, correlation provides a suitable statistical tool to find, measure, and summarise this relationship in a compact mathematical form.

The *coefficient of correlation*, a descriptive statistic, expresses the magnitude and direction of statistical relationship between two variables.

The problem of examining the statistical relationship between two or more variables can be divided into the following sub problems and accordingly requires the development methods to answer these problems.

(i) Is there an association between two or more variables? If yes, what is the form and degree of that relationship?
(ii) Is the relationship strong or significant enough to be useful to arrive at a desirable conclusion?
(iii) Can the relationship be used to predict the most likely value of a dependent variable for the given value of independent variable or variables?

The first two questions will be answered in this chapter, while the third question will be answered in next chapter.

For correlation analysis, the data on values of two variables must come from sampling in pairs, one for each of the two variables. The pairing relationship should represent some time, place, or condition.

3.2 Definition

What is Correlation?

A statistical tool that helps in the study of the relationship between two variables is known as Correlation. It also helps in understanding the economic behaviour of the variable.

3.3 Types of Correlation

There are three broad types of correlations:
(i) Positive, negative and no correlation
(ii) Linear and non-linear
(iii) Simple, partial and multiple

In this chapter, we will discuss simple linear positive or negative correlation analysis.

3.3.1 Positive, Negative and No Correlation

Positive correlation occurs when two variables change in the same direction—both increasing or both decreasing together. In simple terms, as one variable rises, the other also rises, or as one falls, the other falls similarly. Negative (or inverse) correlation happens when variables move in opposite directions—one increases while the other decreases.

The following examples illustrate the concept of positive and negative correlation.

Positive Correlation

Two variables have a positive correlation when they move in the same direction: if one goes up, the other also goes up; if one goes down, the other also goes down. For instance, price and quantity supplied typically rise and fall together—higher prices encourage producers to supply more, while lower prices lead them to supply less. Other common examples include share dividends and share premiums moving together, and employees' years of experience tending to increase alongside their salaries.

Negative Correlation

Two variables are negatively correlated when they move in opposite directions: as one increases, the other decreases, and vice versa. In such a case, the variables do not grow or fall together but instead behave in reverse to each other. For example, the demand for a product and its price usually have a negative relationship: when the price goes up, people buy less, and when the price goes down, people tend to buy more. Similarly, the amount of tax a company pays and the dividend it can distribute often move in opposite directions, because higher taxes can reduce the profit available for dividends.

It may be noted here that the change (increasing or decreasing) in values of both the variables not be proportional or fixed.

No Correlation

Two variables are uncorrelated when changes in one have no link or pattern with changes in the other. In other words, knowing how one variable behaves does not help in predicting the behaviour of the other. For instance, a person's body weight has no meaningful connection with their hair colour, and the same is true for a person's height and hair colour—these pairs of variables do not move together in any systematic way.

3.3.2 Linear and Nonlinear Correlation

A linear correlation means that one variable changes at a steady rate when the other variable changes. In simple terms, if you plot the two variables on a graph, the points lie roughly along a straight line. The relationship is called linear when the changes in the

two variables keep a fixed proportion to each other, so that equal changes in one variable are associated with equal-size changes in the other. The following example illustrates a linear correlation between two variables x and y.

x	10	20	30	40	50
y	40	60	80	100	120

When these pairs of values of x and y are plotted on a graph paper, the line joining these points would be a straight line.

A non-linear (or curvilinear) correlation occurs when the change in one variable does not follow a steady, fixed rate compared with the change in the other variable. In such a relationship, the two variables are still related, but the pattern is curved or irregular rather than forming a straight line when plotted on a graph. The following example illustrates a non-linear correlation between two variables x and y.

x	8	9	9	10	10	28	29	30
y	80	130	170	150	230	560	460	600

When these pair of values of x and y are plotted on a graph paper, the line joining these points would not be a straight line, rather it would be curvi-linear.

3.3.3 Simple, Partial and Multiple Correlations

Correlation tells us how variables are related to each other. The difference between simple, partial and multiple correlation depends on how many variables we study at a time.

Simple correlation

Simple correlation is the relationship between only two variables.

Example: Crop yield and amount of fertilizer.

Example: Sales and money spent on advertisement.

Here, we look at how one variable changes when the other changes, ignoring all other factors.

Partial correlation

Partial correlation also focuses on two main variables, but it assumes that other related factors stay fixed (constant).

Example: Crop yield depends on fertilizer, rainfall, seed quality, soil type and pesticides. If we study only the relationship between yield and seed quality, while assuming that rainfall, soil, fertilizer and pesticides are at average levels, this is partial correlation.

Example: Sales depend on advertisement, price, product quality, competitors, distribution, etc.
If we measure only the correlation between sales and advertisement, keeping other factors fixed, we use partial correlation.

Multiple correlation

Multiple correlation studies the relationship between one main variable and several other variables at the same time.

Example: Employee satisfaction in a company may depend on salary, training, medical facilities, housing, children's education support and grievance handling.
When we study the combined effect of all these variables together on employee satisfaction, it is multiple correlation.

3.4 Methods of Studying Correlation

To understand the relationship between two variables, this section introduces covariance and the correlation coefficient as measures of how strongly two variables x and y are linearly related. These measures are calculated from sample data. The sample correlation coefficient, written as r, does not depend on the units of x and y (it is "scale free"), so its meaning stays the same whether x and y are measured in metres or centimetres, rupees or dollars.

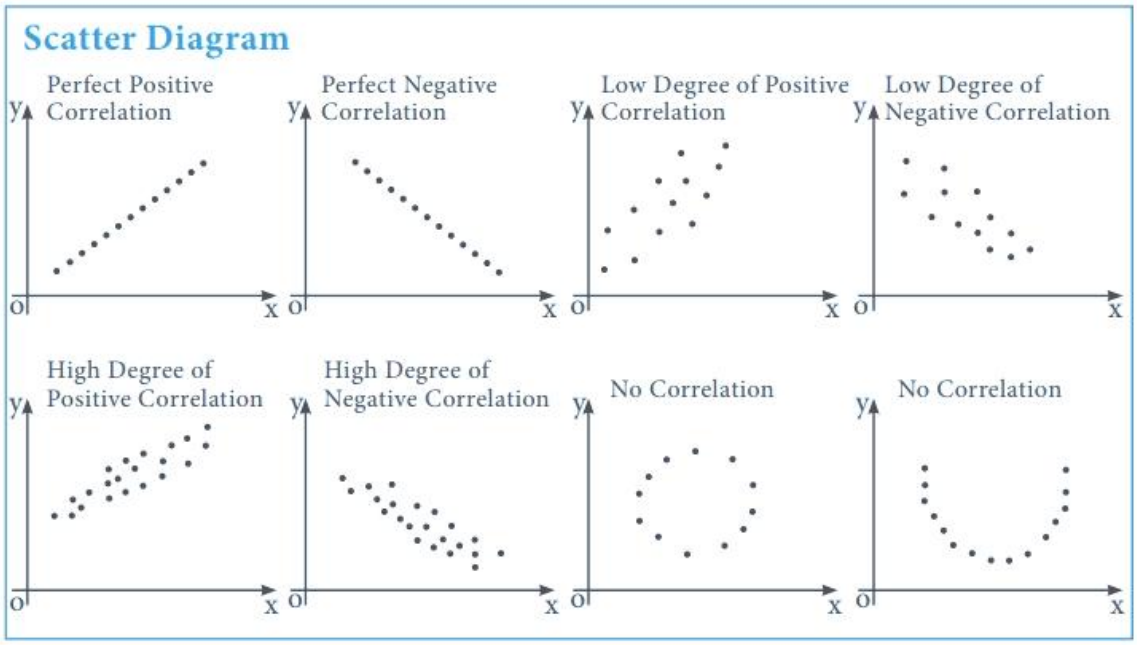
In this chapter, different methods will be explained for finding how large the correlation coefficient is between two variables x and y.

1. Scatter Diagram method
2. Karl Pearson's Coefficient of Correlation method
3. Spearman's Rank Correlation method
4. Method of Least-squares

3.4.1 Scatter Diagram Method

To understand the relationship between two variables, we can draw a scatter diagram. Think of x as height and y as weight. Put height on the horizontal (x) axis and weight on the vertical (y) axis, then plot each pair (x,y) as a dot on graph paper. The collection of dots is called a scatter diagram.

From this diagram, we can get a rough idea about the correlation between x and y. If the dots are grouped around a straight line, the relationship is called linear correlation. If the dots follow a curved pattern, the relationship is called non- linear (or curvilinear) correlation. By looking at whether the points slope upwards or downwards, we can also see if the correlation is positive or negative. They are illustrated in the following diagram:



3.4.2 Karl Pearson's Co-Efficient of Correlation Method

Karl Pearson's correlation coefficient measures quantitatively the extent to which two variables x and y are correlated. For a set of n pairs of values of x and y, Pearson's product moment correlation coefficient is given by

$$r = \frac{\text{Covariance}(x, y)}{\sqrt{\text{var}\, x}\sqrt{\text{var}\, y}} = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

$$where \quad Cov(x,y) = \frac{1}{n}\Sigma(x,\bar{x})(y,\bar{y})$$

$$\sigma_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} \quad \leftarrow \text{standard deviation of sample data on variable x}$$

$$\sigma_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}} \quad \leftarrow \text{standard deviation of sample data on variable y}$$

Substituting mathematical formula for Cov(x, y) and σ_x *and* σ_y, *we have*

$$r = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\sum\left(x-\bar{x}\right)^2}\sqrt{\sum\left(y-\bar{y}\right)^2}}$$

$$= \frac{n\Sigma xy - \Sigma x\Sigma y}{\sqrt{n\Sigma x^2-(\Sigma x)^2}\sqrt{n\Sigma y^2-(\Sigma y)^2}}$$

Step Deviation Method for Ungrouped Data

When actual mean values $\bar{x}$ and $\bar{y}$ are in fraction, the calculation of Pearson's correlation coefficient can be simplified by taking deviations of x and y values from their assumed means A and B, respectively. That is, $d_x =$ x - A and $d_y =$ y - B, where A and B are assumed means of x and y values. The formula becomes

$$r = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{n\Sigma {d_x}^2-(\Sigma d_x)^2}\sqrt{n\Sigma {d_y}^2-(\Sigma d_y)^2}}$$

Step Deviation Method for Grouped Data

When data on x and y values are classified or grouped into a frequency distribution, the formula is modified as:

$$r = \frac{n\Sigma fd_x d_y - \Sigma fd_x \Sigma fd_y}{\sqrt{n\Sigma fd_x^{\,2}-(\Sigma fd_x)^2}\sqrt{n\Sigma fd_y^{\,2}-(\Sigma fd_y)^2}}$$

Assumptions of Using Pearson's Correlation Coefficient:

(i) Pearson's correlation coefficient should be used when both variables x and y are measured on an interval or ratio scale (for example, height, weight, income, temperature).

ii) It works best when both x and y are approximately normally distributed and the relationship between them is roughly a straight-line (linear) relationship.

iii) The value of the correlation can change a lot if the range of values of x or y is cut off (truncated) or if their distributions are very far from normal. In such cases, r may not show the true strength of the relationship.

iv) A meaningful correlation usually appears when there is some real underlying link or influence between the two variables. If there is no such relationship, the correlation may turn out to be very small or close to zero.

Advantage and Disadvantages of Pearson's Correlation Coefficient

The correlation coefficient is a number between -1 and 1 that shows how strong the relationship is between two variables and whether it is positive or negative. A value close to 1 means a strong positive relationship, close to -1 means a strong negative relationship, and close to 0 means little or no linear relationship.

Main limitations of Pearson's correlation:

i) It always assumes that the relationship between the two variables is linear, even if in reality it is not.

ii) Its value can easily be misunderstood, so it must be interpreted carefully and in context.

iii) The coefficient is very sensitive to extreme (outlier) values in the data, which can distort the result.

iv) Compared with some simpler methods, the calculation of Pearson's r can take more time and effort, especially by hand.

3.4.3 The Coefficient of Determination

The coefficient of determination, denoted as represents the proportion of the total variability of the dependent variable, y that is explained by the independent variable, x. Since its value is presented as a proportion or percentage, it measures more precisely the extent or strength of association that exists between two variables x and y .

The correlation coefficient r is often given more importance than it deserves and is used too frequently on its own. A better measure of how much two variables vary together in a linear way is its square, called the coefficient of determination, r^2. This value r^2tells us how much of the variation in one variable can be explained by the variation in the other.

Students and readers should get into the habit of looking at r^2(by squaring r) before deciding how strong is the linear relationship between two variables.

-Tuttle

Mathematically, the coefficient of determination is given by

$$r^2 = 1 - \frac{\text{Explained variability in y}}{\text{Total variability in y}}$$

$$= 1 - \frac{\Sigma(\text{y} - \hat{\text{y}})^2}{\Sigma(\text{y} - \bar{\text{y}})^2} = 1 - \frac{n\Sigma y^2 - a\Sigma y - b\Sigma xy}{n\Sigma y^2 - (\bar{y})^2}$$

Where $\hat{y} = a + bx$ and is the estimated value of y for given values of x. One minus the ratio between where these two variations is referred as the *coefficient of determination*.

The limits of two measures r and r^2 can be written as:

$$-1 \le r \le 1 \; and \; 0 \le r^2 \le 1$$

For example, let correlation variable x (height) and variable y (weight) be $r = 0.70$. Now the coefficient of determination $r^2 = 0.49$ or 49%, implies that 49% of the variable in variable y(weight) can be accounted for in terms of variable x(height). The remaining 51% of the variability may be due to other factors, say for instance, tendency to eat fatty foods.

It may be noted that even a relatively high correlation coefficient r = 0.70 accounts for less than 50 per cent of the variability. In this context, it is important to know that 'variability' refers to how values of variable y are scattered around its own mean value. That is, as in the above example, some people will be heavy, some average, some light. So we can account for 49% of the total variability of weight(y) in terms of height(x) if r = 0.70 .The greater the correlation coefficient, the greater the coefficient of determination, and the variability in dependent variable can be accounted for in terms of independent variable.

Example 1:

Find the coefficient of correlation between x and y.

x	1	2	3	4	5	6	7	8	9
y	12	11	13	15	14	17	16	19	18

Solution:

x	y	$dx = x - \bar{x}$	$dy = y - \bar{y}$	dx^2	dy^2	dxdy
1	12	-4	-3	16	9	12
2	11	-3	-4	9	16	12
3	13	-2	-2	4	4	4
4	15	-1	0	1	0	0
5	14	0	-1	0	1	0
6	17	1	2	1	4	2
7	16	2	1	4	1	2
8	19	3	4	9	16	12
9	18	4	3	16	9	12
45	**135**	**0**	**0**	**60**	**60**	**56**

$$\bar{x} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{135}{9} = 15$$

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

$$= \frac{56}{\sqrt{60\sqrt{60}}} = 0.93$$

Example 2:

Find the coefficient of correlation between x and y.

x	10	12	13	16	17	20	25
y	19	22	26	27	29	33	37

Solution:

x	y	x-A=dx	y-B=dy	dx^2	dy^2	dxdy
10	19	-6	-8	36	64	48
12	22	-4	-5	16	25	20
13	26	-3	-1	9	1	3
16	27	0	0	0	0	0
17	29	1	2	1	4	2
20	33	4	6	16	36	24
25	37	9	10	81	100	90
113	**193**	**1**	**4**	**159**	**230**	**187**

$$\bar{x} = \frac{113}{7} = 16\frac{1}{7}$$

$$\bar{y} = \frac{193}{7} = 27\frac{4}{7}$$

Take the assumed values A=16, B= 27

$$dx = x - A = x - 16$$

$$dy = x - B = y - 27$$

$$r = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{n\Sigma d_x{}^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y{}^2 - (\Sigma d_y)^2}}$$

$$= \frac{186.43}{\sqrt{158.86}\sqrt{227.71}} = \frac{186.43}{190.19} = 0.980$$

Example 3:

The following table gives indices of industrial production and number of registered unemployed people (in lakh). Calculate the value of the correlation coefficient.

Year	:	1991	1992	1993	1994	1995	1996	1997	1998
Index of Production	:	100	102	104	107	105	112	103	99
Number of Unemployed	:	15	12	13	11	12	12	19	26

Solution:

Calculations of Karl Pearson's correlation coefficient are shown in the table below:

Year	Production x	$d_x = (x - \bar{x})$	d_x^2	Unemployed y	$d_y = (y - \bar{y})$	d_y^2	$d_x d_y$
1991	100	-4	16	15	0	0	0
1992	102	-2	4	12	-3	9	+6
1993	104	0	0	13	-2	4	0
1994	107	+3	9	11	-4	16	-12
1995	105	+1	1	12	-3	9	-3
1996	112	+8	64	12	-3	9	-24
1997	103	-1	1	19	+4	16	-4
1998	99	-5	25	26	+11	121	-55
Total	**832**	**0**	**120**	**120**	**0**	**184**	**-92**

$$\bar{x} = \frac{\Sigma x}{n} = \frac{832}{8} = 104$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{120}{8} = 15$$

Applying the formula

$$r = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{n\Sigma d_x{}^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y{}^2 - (\Sigma d_y)^2}}$$

$$= \frac{-736}{\sqrt{960}\sqrt{1472}} = \frac{-92}{148.580} = -0.619$$

Since coefficient of correlation is negative 0.619, it indicates that there is a fairly large inverse correlation between the two variables. Hence we conclude that as the production index increases, the number of unemployed decreases and vice-versa.

Example 4:

The following table gives the distribution of items of production and also the relatively defective items among them, according to size groups. Find the correlation coefficient between size and defect in quality.

Size-group	**:**	**15-16**	**16-17**	**17-18**	**18-19**	**19-20**	**20-21**
No. of items	:	200	270	340	360	400	300
No. of Defective items	:	150	162	170	180	180	114

Solution:

Let group size be denoted by variable x and number of defective items by variable y. Calculations for Karl Pearson's correlation coefficient are shown below:

Size-Group	Mid-value m	$d_x = m - 17.5$	d_x^2	Percent of Defective Items	$d_y = y - 50$	d_y^2	$d_x d_y$
15-16	15.5	-2	4	75	+25	625	-50
16-17	16.5	-1	1	60	+10	100	-10
17-18	17.5	0	0	50	0	0	0
18-19	18.5	+1	1	50	0	0	0
19-20	19.5	+2	4	45	25	25	-10
20-21	20.5	+3	9	38	144	144	-36
		3	**19**		**18**	**894**	**-106**

Substituting values in the formula of Karl Pearson's correlation coefficient r,
we have

$$r = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{n\Sigma d_x{}^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y{}^2 - (\Sigma d_y)^2}}$$

$$= \frac{-636-54}{\sqrt{105}\sqrt{5040}} = \frac{-690}{727.46} = -0.949$$

Since value of r is negative, therefore size of groups and number of defective items are inversely correlated with high degree. Hence we conclude that when size of group increases, the number of defective items decreases and vice-versa.

Example 5:

The following data relate to age of employees and the number of days they reported sick in a month. Calculate Karl Pearson's coefficient of correlation and interpret it.

Employees	:	1	2	3	4	5	6	7	8	9	10
Age	:	30	32	35	40	48	50	52	55	57	61
Sick days	:	1	0	2	5	2	4	6	5	7	8

Solution:

Let age and sick days be represented by variables x and y, respectively. Calculations for value of correlation coefficient are shown below:

Age x	Sick days					
	$d_x = (x - \bar{x})$	d_x^2	y	$d_y = (y - \bar{y})$	d_y^2	$d_x d_y$
30	-16	256	1	-3	9	48
32	-14	196	0	-4	16	56
35	-11	121	2	-2	4	22
40	-6	36	5	1	1	-6
48	2	4	2	-2	4	-4
50	4	16	4	0	0	0
52	6	36	6	2	4	12
55	9	81	5	1	1	9
57	11	121	7	3	9	33
61	15	225	8	4	16	60
460	**0**	**1092**	**40**	**0**	**64**	**230**

$$\bar{x} = \frac{\Sigma x}{n} = \frac{460}{10} = 46$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{40}{10} = 4$$

Substituting values in the formula of Karl Pearson's correlation coefficient r, we have

$$r = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{n\Sigma d_x{}^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y{}^2 - (\Sigma d_y)^2}}$$

$$= \frac{2300}{\sqrt{10920}\sqrt{640}} = \frac{230}{264.363} = 0.870$$

Since value of r is positive, therefore age of employees and number of sick days are positively correlated to a high degree. Hence we conclude that as the age of an employee increases, he is likely to go on sick leave more often than others.

Example 6:

Find the coefficient of correlation between x and y from the following data:

$n = 10, \Sigma x = 60, \Sigma y = 60, \Sigma xy = 305, \Sigma x^2 = 400, \Sigma y^2 = 580$

Solution:

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{3050 - 3600}{\sqrt{4000 - 3600}\sqrt{5800 - 3600}}$$

$$= \frac{-550}{\sqrt{400}\sqrt{2200}}$$

$$= -0.5864$$

Example 7:

Calculate the coefficient of correlation from the following data:

$$n = 10, \Sigma x = 50, \Sigma y = -30, \Sigma xy = -115, \Sigma x^2 = 290, \Sigma y^2 = 300$$

Solution:

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{-1150 + 1500}{\sqrt{2900 - 2500}\sqrt{3000 - 900}}$$

$$= 0.3819$$

Example 8:

The following table gives the frequency, according to the marks, obtained by 67 students in an intelligence test. Measure the degree of relationship between age and marks:

Test Marks	Age in years				Total
	18	19	20	21	
200-250	4	4	2	1	11
250-300	3	5	4	2	14
300-350	2	6	8	5	21
350-400	1	4	6	10	21
Total	**10**	**19**	**20**	**18**	**67**

Solution:

Let age of students and marks obtained by them be represented by variables x and y. respectively. Calculations for correlation coefficient for this bivariate data is shown below:

x / d_x / y / dy		Age in years				Total f	f d_y	f d_y^2	f d_x d_y
		18	19	20	21				
		-1	0	1	2				
200-250	-1	4 4	0 4	-2 2	-2 1	11	-11	11	0
250-300	0	0 3	0 5	0 4	0 2	14	0	0	0
300-350	1	-2 2	0 6	8 8	10 5	21	21	21	16
350-400	2	-2 1	0 4	12 6	40 10	21	42	84	50
Total f		10	19	20	18	n=67	$\Sigma f d_y$ =52	$\Sigma f d_y^2$ =116	$\Sigma f d_x d_y$ =66
f d_x		-10	0	20	36	$\Sigma f d_x$ =46			
f d_x^2		10	0	20	72	$\Sigma f d_x^2$ =102			
f d_x d_y		0	0	18	48	$\Sigma f d_x d_y$ =66			

Substituting values in the formula of Karl Pearson's correlation coefficient, we have

$$r = \frac{n\Sigma fd_x d_y - \Sigma fd_x \Sigma fd_y}{\sqrt{n\Sigma fd_x{}^2 - (\Sigma fd_x)^2}\sqrt{n\Sigma fd_y{}^2 - (\Sigma fd_y)^2}}$$

$$= \frac{4422 - 2392}{\sqrt{6834 - 2116}\sqrt{7772 - 2704}} = \frac{2030}{\sqrt{4718}\sqrt{5068}} = \frac{2030}{4889.898} = 0.415$$

Because r is positive, the students' age and their intelligence test scores have a positive relationship. A correlation of 0.415 suggests a moderate positive correlation, so in general, as students get older, their scores on the intelligence test tend to increase.

Example 9:

A computer, while calculating the correlation coefficient between two variables x and y from 25 pairs of observations, obtained the following results:

$$n = 25, \quad \Sigma x = 125, \Sigma x^2 = 650 \quad \text{and} \quad \Sigma y = 100, \quad \Sigma y^2 = 460, \Sigma xy = 508$$

It was, however, discovered at the time of checking that he had copied down two pairs of observations as:

x	y
6	14
8	6

in-stead of

x	y
8	12
6	8

Obtain the correct value of correlation coefficient between x and y.

Solution:

The corrected values for termed needed in the formula of Pearson's correlation coefficient are determined as follows:

$$Correct \ \Sigma x = 125 - 6 - 8 + 8 + 6 = 125$$
$$Correct \ \Sigma y = 100 - 14 - 6 + 12 + 8 = 100$$
$$Correct \ \Sigma x^2 = 650 - (6)^2 - (8)^2 + (8)^2 + (6)^2 = 650 - 36 - 64 + 64 + 36 = 650$$
$$Correct \ \Sigma y^2 = 460 - (14)^2 - (6)^2 + (12)^2 + (8)^2 = 460 - 196 - 36 + 144 + 64 = 436$$
$$Correct \ \Sigma xy = 508 - 84 - 48 + 96 + 48 = 520$$

Applying the formula

$$r = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\sum\left(x-\bar{x}\right)^2}\sqrt{\sum\left(y-\bar{y}\right)^2}}$$

$$r = \frac{n\sum xy - \sum x\sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

$$= \frac{25X520 - 125X100}{\sqrt{25X650 - (125)^2}\sqrt{25X436 - (100)^2}}$$

$$r = 0.667$$

Thus, the correct value of correlation coefficient between x and y is 0.667.

Example 10:

Calculate the coefficient of correlation from the following bivariate frequency distribution:

Sales Revenue (Rs in lakh)	Advertising Expenditure(Rs in '000)				Total
	5-10	10-15	15-20	20-25	
75-125	4	1	-	-	5
125-175	7	6	2	1	16
175-225	1	3	4	2	10
225-275	1	1	3	4	9
Total	**13**	**11**	**9**	**7**	**40**

Solution:

Let advertising expenditure and sales revenue be represented by variables x and y, respectively. The calculations for correlation coefficient are shown below:

x Mid value(m) d_x=m-12.5			Advertising Expenditure				Total f	$f d_y$	$f d_y^2$	$f d_x d_y$
			5-10	10-15	15-20	20-25				
			7.5	12.5	17.5	22.5				
			-1	0	1	2				
Revenue y	Mid Value (m)	d_y= m-150								
75-125	100	-2	8 4	0 1	0 -	0 -	5	-10	20	8
125-175	150	-1	7 7	0 6	-2 2	-2 1	16	-16	16	3
175-225	200	0	0 1	0 3	0 4	0 2	10	0	0	0
225-275	250	1	-1 1	0 1	3 3	8 4	9	9	9	10
Total f			13	11	9	7	n=40	$\Sigma f d_y$ =-17	$\Sigma f d_y^2$ =45	$\Sigma f d_x d_y$ =21
$f d_x$			-13	0	9	14	$\Sigma f d_x$ =10			
$f d_x^2$			13	0	9	28	$\Sigma f d_x^2$ =50			
$f d_x d_y$			14	0	1	6	$\Sigma f d_x d_y$ =21			

Substituting values in the formula of Karl Pearson's correlation coefficient,

we have

$$r = \frac{n\Sigma fd_x d_y - \Sigma fd_x \Sigma fd_y}{\sqrt{n\Sigma fd_x{}^2 - (\Sigma fd_x)^2}\sqrt{n\Sigma fd_y{}^2 - (\Sigma fd_y)^2}}$$

$$= \frac{840 + 170}{\sqrt{1900}\sqrt{1511}} = \frac{1010}{1694.373} = 0.596$$

Since the value of r is positive, advertising expenditure and sales revenue are positively correlated to the extent of 0.596. Hence we conclude that as expenditure on advertising increases, the sales revenue also increases.

Exercise Problems 1:

1, Find the correlation coefficient by Karl Pearson's method between x and y interpret its value

x :	57	42	40	33	42	45	42	44	40	56	44	43
y :	10	60	30	41	29	27	27	19	18	19	31	29

(Answer: r = -0.554)

2. Calculate Karl Pearson's coefficient of correlation between age and playing habits from the data given below. Also calculate the probable error and comment on the value:

Age:	20	21	22	23	24	25
No. Of students:	500	400	300	240	200	160
Regular players:	400	300	180	96	60	24

(Answer: r = 0.005)

3. Find the coefficient of correlation between age and the sum assured from the following table:

Age Group (years)	Sum Assured (in Rs)				
	10,000	20,000	30,000	40,000	50,000
20-30	4	6	3	7	1
30-40	2	8	15	7	1
40-50	3	9	12	6	2
50-60	8	4	2	-	-

(Answer: r = -0.256)

4. The coefficient of correlation between two variables x and y is 0.3. The covariance is 9. The variance x is 16. Find the standard deviation of y series.

(Answer: $\sigma_y = 7.5$)

3.5 Spearman's Rank Correlation Method

This method for calculating a correlation coefficient between two variables was introduced by the British psychologist Charles Edward Spearman in 1904. Spearman's approach is now known as Spearman's rank correlation method. This method is applied to measure the association between two variables when only ordinal or rank data are available. In other words, this method is applied in a situation in which quantitative measure of certain qualitative factors such as judgement, leadership, colour, taste, cannot be fixed, but individual observations can be arranged in a definite order (also called rank). The ranking is decided by using a set of ordinal rank numbers, with 1 for the individual observation ranked first either in terms of quantity or quality; and n for the individual observation ranked last in a group of n pairs of observation. Mathematically, Spearman's rank correlation coefficient is defined as:

$$R = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

Where

R = rank correlation coefficient

R_1 = rank of observations with respect to first variable

R_2 = rank of observations with respect to second variable

d = R_1-R_2, difference in a pair of ranks

n = number of pairs of observations or individuals being ranked

The number '6' is placed in the formula as a scaling device, it ensures that the possible range of R is from -1 to 1. While using this method we may come across three types of cases.

Case I: When Ranks are given

When observations in a data set are already arranged in a particular order (rank), take the differences in pairs of observations to determine d. Square these differences and obtain the total Ed. Apply, formula to calculate correlation coefficient.

Example 11:

The rank correlation coefficient between debenture prices and share prices is 0.143. The sum of the squared differences between their ranks is 48. Using this information, find the number of paired observations n.

Solution:

The formula for Spearman's correlation coefficient is as follows:

$$R = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

Given R=0.143, $\Sigma d^2 = 48$ and n=7. Substituting values in the formula, we get

$$0.143 = 1 - \frac{288}{n(n^2 - 1)} = 1 - \frac{288}{n^3 - n}$$

$$0.143(n^3 - n) = (n^3 - n) - 288$$

$$n^3 - n - 336 = 0 \; or \quad (n-7)(n^2 + 7n + 48) = 0$$

This implies that either n-7=0, that is, n = 7 or $n^2 + 7n + 48 = 0$ But $n^2 + 7n + 48 = 0$ on simplification gives undesirable value of n because its discriminant b^2-4ac is negative. Hence n = 7.

Example 12:

Ten competitors in a beauty contest are ranked by three judges in the following order:

First judge	1	4	6	3	2	9	7	8	10	5
Second judge	2	6	5	4	7	10	9	3	8	1
Third judge	3	7	4	5	10	8	9	2	6	1

Use the method of rank correlation coefficient to determine which pair of judges have the nearest approach to common taste in beauty?

Solution:

Let x, y, z denote the ranks by 1st, 2nd,3rd judges respectively.

x	y	z	d_{xy}	d_{yz}	d_{zx}	d_{xy}^2	d_{yz}^2	d_{zx}^2
1	2	3	-1	-1	-2	1	1	4
4	6	7	-2	-1	-3	4	1	9
6	5	4	1	1	2	1	1	4
3	4	5	-1	-1	-2	1	1	4
2	7	10	-5	-3	-8	25	9	64
9	10	8	-1	2	1	1	4	1
7	9	9	-2	0	-2	4	0	4
8	3	2	5	1	6	25	1	36
10	8	6	2	2	4	4	4	16
5	1	1	4	0	4	16	0	16
						82	**22**	**158**

$$\rho_{xy} = 1 - \frac{6\Sigma d_{xy}^{\ 2}}{N(N^2-1)} = 1 - \frac{492}{990} = 0.503$$

$$\rho_{yz} = 1 - \frac{6\Sigma d_{yz}^{\ 2}}{N(N^2-1)} = 1 - \frac{132}{990} = 0.867$$

$$\rho_{xz} = 1 - \frac{6\Sigma d_{xz}^{\ 2}}{N(N^2-1)} = 1 - \frac{948}{990} = 0.04$$

Since the rank correlation coefficient between y and z is positive and highest among the three coefficients, judges y and z have the nearest approach for common taste in beauty.

Example 13: Find the rank correlation coefficient for the following data:

x	92	89	87	86	86	77	71	63	53	50
y	86	83	91	77	68	85	52	82	37	57

Solution: Let R_1 and R_2 denote the ranks in x and y respectively.

x	y	R_1	R_2	$d = R_1 - R_2$	d^2
92	86	1	2	-1	1
89	83	2	4	-2	4
87	91	3	1	2	4
86	77	4.5	6	-1.5	2.25
86	68	4.5	7	-2.5	6.25
77	85	6	3	3	9
71	52	7	9	-2	4
63	82	8	5	3	9
53	37	9	10	-1	1
50	57	10	8	2	4
					44.50

$$\rho = 1 - \frac{6\left[\Sigma d^2 + \frac{\Sigma m(m^2-1)}{12}\right]}{N(N^2-1)}$$

$$= 1 - \frac{6\left[44.5 + \frac{2(2^2-1)}{12}\right]}{990}$$

$$= 1 - \frac{6(44.5+0.5)}{990} = 1 - \frac{270}{990} = 0.727$$

Example 14:

The positions of 15 students in two subjects, A and B, are shown below. In each pair of brackets, the first number is the student's rank in subject A and the second number is the student's rank in subject B. (1.10), (2.7); (3,2), (4,6), (5,4), (6,8), (7,3), (8, 1), (9,11), (10, 15), (11,9),(12,5),(13,14),(14, 12),(15,13).

Find Spearman's rank correlation coefficient.

Solution:

Since ranks of students with respect to their performance in two subjects are given, calculations for rank correlation coefficient are shown below:

Rank in A R_1	Rank in B R_2	$d = R_1 - R_2$	d^2

1	10	-9	81
2	7	-5	25
3	2	1	1
4	6	-2	4
5	4	1	1
6	8	-2	4
7	3	4	16
8	1	7	49
9	11	-2	4
10	15	-5	25
11	9	2	4
12	5	7	49
13	14	-1	1
14	12	2	4
15	13	2	4
			272

Applying the formula

$$R = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

$$= 1 - \frac{6(272)}{15^3 - 15} = 1 - \frac{1632}{3360} = 1 - 0.4857 = 0.5143$$

Since R = 0.5143 performance of students in two subjects is positively correlated to a moderate degree.

Example 15:

An office has 12 clerks. The long-serving clerks feel that they should have a seniority increment based on length of service built into their salary structure. An assessment of their efficiency by their departmental manager and the personnel department produces a ranking of efficiency. This is shown below together with a ranking of their length of service.

Ranking according	1	2	3	4	5	6	7	8	9	10	11	12

to length of service												
Ranking according to efficiency	2	3	5	1	9	10	11	12	8	7	6	4

Do the data support the clerks' claim for seniority increment?

Solution:

Since ranks are already given, calculations for rank correlation coefficient are shown below:

Rank According to Length of Service R_1	Rank According to Efficiency R_2	$d = R_1 - R_2$	d^2
1	2	-1	1
2	3	-1	1
3	5	-2	4
4	1	3	9
5	9	-4	16
6	10	-4	16
7	11	-4	16
8	12	-4	16
9	8	1	1
10	7	3	9
11	6	5	25
12	4	8	64
			178

Applying the formula

$$R = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

$$= 1 - \frac{6(178)}{12^3 - 12} = 1 - \frac{1068}{1716} = 0.378$$

Since $R = 0.378$ is a low degree positive correlation between length of service and efficiency, the claim of the for a seniority increment based on length of service is not justified.

Case 2: When Ranks are not given

When pairs of observations in the data set are not ranked as in Case 1, the ranks are assigned by taking either the highest value or the lowest value as 1 for both the variable's values.

Example 16 :

Quotations of index numbers of security prices of a certain joint stock company are given below:

Year	Debenture price	share price
1	97.8	73.2
2	99.2	85.8
3	98.8	78.9
4	98.3	75.8
5	98.4	77.2
6	96.7	87.2
7	97.1	83.8

Using the rank correlation method, determine the relationship between debenture prices and share prices.

Solution:

Let us start ranking from the lowest value for both the variables, as shown below:

Debenture Price (x)	R_1	Share Price (y)	R_2	$d = R_1 - R_2$	d^2
97.8	3	73.2	1	2	4
99.2	7	85.8	6	1	1
98.8	6	78.9	4	2	4
98.3	4	75.8	2	2	4
98.4	5	77.2	3	2	4
96.7	1	87.2	7	-6	36
97.1	2	83.8	5	-3	9
					62

Applying the formula

$$R = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

$$=1-\frac{6(62)}{7^3-7}=1-\frac{372}{336}=1-1.107=-0.107$$

There is a low degree of negative debenture prices and share prices of a certain joint stock company.

Case 3: When Ranks are Equal

When we give ranks to data (starting from rank 1 for either the largest or the smallest value), we may find that some observations have exactly the same value. In that case, each of those equal observations is given the average of the ranks they would have received if they were slightly different. For example, if two values are tied for 3rd place, we take the average of ranks 3 and 4, so each gets rank (3+4)/2=3.5(3+4)/2=3.5. If three values are tied at 3rd place, we average ranks 3, 4 and 5, so each gets rank (3+4+5)/3=4(3+4+5)/3=4.

When such tied ranks appear in the data, a correction (adjustment) has to be made in the usual Spearman rank correlation formula before calculating the coefficient.

$$R=1-\frac{6\{\Sigma d^2+\frac{1}{12}(m_1^3-m_1)+\frac{1}{12}(m_2^3-m_2)+\ldots\}}{n(n^2-1)}$$

where m_i(i = 1, 2, 3 ,...) stands for the number of times an observation is repeated in the data set for both variables.

Example 17 :

Obtain the rank correlation coefficient between the variables x and y from the following pairs of observed values.

x	50	55	65	50	55	60	50	65	70	75
y	110	110	115	125	140	115	130	120	115	160

Solution:

Let us start ranking from lowest value for both the variables as shown below. Moreover, certain observations in both sets of data are repeated, the ranking is done in accordance with suitable average value.

Variable x	Rank R_1	Variable y	Rank R_2	d = R_1 - R_2	d^2
50	2	110	1.5	0.5	0.25
55	4.5	110	1.5	3	9
65	7.5	115	4	3.5	12.25

50	2	125	7	-5	25
55	4.5	140	9	-4.5	20.25
60	6	115	4	2	4
50	2	130	8	-6	36
65	7.5	120	6	1.5	2.25
70	9	115	4	5	25
75	10	160	10	0	0
					134

It may be noted that in series x, 50 is repeated thrice ($m_1 = 3$), 55 is repeated twice ($m_2 = 2$), and 65 is repeated twice ($m_3 = 2$) . In series y, 110 is repeated ($m_4 = 2$) and 115 thrice ($m_5 = 3$).

$$R = 1 - \frac{6\{\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \ldots + \frac{1}{12}(m_5^3 - m_5)\}}{n(n^2-1)}$$

$$= 1 - \frac{6\{134 + \frac{1}{12}(3^3-3) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(3^3-3)\}}{10(10^2-1)}$$

$$= 1 - \frac{6[134 + 2 + 0.5 + 0.5 + 0.5 + 2]}{990}$$

$$= 1 - \frac{6(139.5)}{990} = 1 - \frac{837}{990} = 1 - 0.845 = 0.155$$

Advantages and Disadvantages of Spearman's Correlation Coefficient Method

Advantages

(i) This method is easy to understand and its application is simpler than Pearson's method.

(ii) This method is useful for correlation analysis when variables are expressed in qualitative terms like beauty, intelligence, honesty, efficiency, and so on.

(iii) This method is appropriate to use when both variables are measured on an interval or on a ratio scale.

(iv) The sample data of values of two variables is converted into ranks either in ascending order or descending order for calculating degree of correlation between two variables.

Disadvantages

(i) Values of both variables are assumed to be normally distributed and describing a linear (rather than colinear) relationship.

(ii) A large computational time is required when number of pairs of values of two variables exceed 30.

(iii) This method cannot be applied on a bivariate grouped data for correlation analysis.

Exercise Problems 2:

1. The ranking of 10 students in accordance with their performance in two subjects A and B are as follows:

A	6	5	3	10	2	4	9	7	8	1
B	3	8	4	9	1	6	10	7	5	2

Calculate the rank correlation coefficient and comment on its value.

(Answer: R = 0.782)

2. An examination of eight applicants for a clerical post was taken by a firm. From the marks obtained by the applicants in the accountancy and statistics papers, compute the rank correlation coefficient.

Applicant	A	B	C	D	E	F	G	H
Marks in accountancy	15	20	28	12	40	60	20	80
Marks in statistics	40	30	50	30	20	10	30	60

(Answer: R = 0)

3. An investigator collected the following data with respect to the socioeconomic status and severity of respiratory illness.

Patient	1	2	3	4	5	6	7	8
Socio-economic Status (rank)	6	7	2	3	5	4	1	8
Severity of illness (rank)	5	8	4	3	7	1	2	6

Calculate the rank correlation coefficient and comment on its value.

(Answer: R = 0.71)

4. The personnel department is interested in comparing the ratings of job applicants when measured by a variety of standard tests. The ratings of 9 applicants on interviews and standard psychological test are shown below:

Applicant	A	B	C	D	E	F	G	H	I
Interview	5	2	9	4	3	6	1	8	7
Standard test	8	1	7	5	3	4	2	9	6

Calculate Spearman's rank correlation coefficient and comment on its value.

(Answer: R = 0.871)

Unit IV : Regression Analysis

Structure

- **Regression lines**
- **Correlation Vs Regression**

4.1 REGRESSION

Regression is a mathematical measure of the average relationship between two or more variables (often of the original units of the data).

Fit Regression

The line of regression of **x on y** is given by:

$$X - \bar{X} = r \cdot \frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$$

$$X = a + b_{xy}Y$$

The line of regression of **y on x** is given by:

$$Y - \bar{Y} = r \cdot \frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

$$Y = a + b_{yx}X$$

Note:

- Both the lines of regression pass through mean value of X and Y

Angle between Two Lines of Regression

The angle **θ** between the two lines of regression is given by:

$$\tan\theta = \frac{r\sigma_Y}{\sigma_X} - \frac{r\sigma_X}{\sigma_Y}$$

Note:

- If **r = ±1**, one set of regression lines **coincides**.

Special Cases:

1. If **r = ±1**, then **θ = 0** Regression lines are parallel or coincide
2. If **r = 0**, the two regression lines (**y on x** and **x on y**) are **uncorrelated**.
3. If **r > 1**, the correlation between **x and y** is **positive**.

4.2 Regression Coefficients

Regression Coefficient of Y on X

$$b_{yx} = r\frac{\sigma_Y}{\sigma_X}$$

Regression Coefficient of X on Y

$$b_{xy} = r\frac{\sigma_X}{\sigma_Y}$$

- The relation between b_{xy} **and** b_{yx}:

$$b_{yx} \times b_{xy} = r^2$$

- The **correlation coefficient**:

$$r = \pm\sqrt{b_{yx} \times b_{xy}}$$

Formula for Regression Coefficients:

The regression coefficients $\mathbf{b_{xy}}$ and $\mathbf{b_{yx}}$ can be easily obtained using:

$$b_{xy} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(Y - \bar{Y})^2}$$

$$b_{yx} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

4.3 Comparison Between Correlation and Regression

Aspect	Correlation	Regression
Purpose	Measures the degree and direction of linear relationship between two variables.	Establishes a functional relationship to predict one variable based on another.
Type of Relationship	Symmetrical – no distinction between dependent and independent variables.	Asymmetrical – one variable is dependent, the other is independent.
Result	A single coefficient lies between -1 and $+1$.	Two regression equations: Y on X and X on Y.
Interpretation	Indicates strength and direction of relationship.	Helps in prediction and estimation.
Formula (Karl Pearson)	$r = \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}}$	$Y - \bar{Y} = b_{yx}(X - \bar{X})$
Change of Scale	Not affected by change of scale.	Affected by scale but not origin.
Use in Analysis	Primarily to understand association.	Used in predictive modeling.

Example 1:

Given the following data for marks in Economics (X) and Statistics (Y):

Marks in Economics (X)	25	28	35	32	31	36	29	38	34	32
Marks in Statistics (Y)	43	46	49	41	36	32	31	30	33	39

1. Find the regression equations of Y on X, and regression equation of X on Y.
2. Find the coefficient of correlation between marks in Economics and Statistics.
3. Find the most likely marks in Statistics when the marks in Economics are 30.

Step 1: Compute Necessary Components

X(Economics)	Y (Statistics)	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-24
29	31	-3	-7	9	49	21
38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
Total	**Sum = 320**			**140**	**394**	**-93**

Summarized Calculations

- Mean of X:

$$\bar{X} = \frac{\sum X}{n} = \frac{320}{10} = 32$$

- Mean of Y:

$$\bar{Y} = \frac{\sum Y}{n} = \frac{380}{10} = 38$$

- Variance of X (used for standard deviation):

$$\sigma_X^2 = \frac{\sum (X - \bar{X})^2}{n} = \frac{140}{10} = 14$$
$$\sigma_X = \sqrt{14} = 3.74$$

- Variance of Y:

$$\sigma_Y^2 = \frac{\sum (Y - \bar{Y})^2}{n} = \frac{394}{10} = 39.4$$
$$\sigma_Y = \sqrt{39.4} = 6.31$$

- Covariance:

$$\mathrm{Cov}(X, Y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n} = \frac{-93}{10} = -9.3$$

Regression Coefficients

- **Regression coefficient of Y on X**:

$$b_{yx} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X^2} = \frac{-9.3}{14} = -0.664$$

- **Regression coefficient of X on Y**:

$$b_{xy} = \frac{\text{Cov}(X,Y)}{\sigma_Y^2} = \frac{-9.3}{39.4} = -0.234$$

Regression Equation of Y on X:

$$Y - 38 = -0.664(X - 32)$$

$$Y = -0.664X + 59.26$$

Regression Equation of X on Y:

$$X - 32 = -0.234(Y - 38)$$

$$X = -0.234Y + 40.47$$

Correlation Coefficient

$$r = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{-9.3}{(3.74 \times 6.31)} = -0.394$$

This indicates a weak negative correlation.

Prediction: Most Likely Marks in Statistics When X = 30

Using the regression equation **Y on X**:

$$Y = -0.664(30) + 59.26$$

$$Y = 39.33$$

So, when the Economics marks are **30**, the most likely Statistics marks are **39**.

Example 2:

Obtain the equations of the regression lines from the following data using the method of least squares. Hence, find the coefficient of correlation between X and Y.

Also, estimate the value of:

(i) Y when X=38

(ii) X when Y=18

Given Data:

X	22	26	29	30	31	31	34	35
Y	20	20	21	29	27	24	27	31

Solution

Step 1: Calculate Mean of X and Y

The mean X is:

$$\bar{X} = \frac{\sum X}{n} = \frac{22 + 26 + 29 + 30 + 31 + 31 + 34 + 35}{8}$$

$$\bar{X} = \frac{238}{8} = 29.75$$

The mean Y is:

$$\bar{Y} = \frac{\sum Y}{n} = \frac{20 + 20 + 21 + 29 + 27 + 24 + 27 + 31}{8}$$

$$\bar{Y} = \frac{199}{8} = 24.875$$

Step 2: Create the Calculation Table

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
22	20	-7.75	-4.875	60.0625	23.7656	37.7813
26	20	-3.75	-4.875	14.0625	23.7656	18.2813
29	21	-0.75	-3.875	0.5625	15.0156	2.9063
30	29	0.25	4.125	0.0625	17.0156	1.0313
31	27	1.25	2.125	1.5625	4.5156	2.6563
31	24	1.25	-0.875	1.5625	0.7656	-1.0938
34	27	4.25	2.125	18.0625	4.5156	9.0313
35	31	5.25	6.125	27.5625	37.5156	32.1563

$$\sum(X-\bar{X})^2 = 123.4375, \quad \sum(Y-\bar{Y})^2 = 126.875, \quad \sum(X-\bar{X})(Y-\bar{Y}) = 100.25$$

Step 3: Compute Regression Coefficients

- **Regression coefficient of Y on X**:

$$b_{yx} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$$

$$b_{yx} = \frac{100.25}{123.4375} = 0.812$$

- **Regression coefficient of X on Y**:

$$b_{xy} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(Y - \bar{Y})^2}$$

$$b_{xy} = \frac{100.25}{126.875} = 0.79$$

Step 4: Regression Equations

1. **Regression Equation of Y on X**:

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$Y - 24.875 = 0.812(X - 29.75)$$

$$Y = 0.812X + 0.75$$

2. **Regression Equation of X on Y**:

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$X - 29.75 = 0.79(Y - 24.875)$$

$$X = 0.79Y + 10.1$$

Step 5: Compute Correlation Coefficient

$$r = \sqrt{b_{yx} \times b_{xy}}$$

$$r = \sqrt{0.812 \times 0.79}$$

$$r = \sqrt{0.6415} = 0.801$$

Step 6: Predictions

(i) **Estimate Y when X=38:**

Using Y=0.812X+0.75Y

$$Y = 0.812(38) + 0.75$$

$$Y = 30.546$$

(ii) **Estimate X when Y=18:**

Using X=0.79Y+10.1X

$$X = 0.79(18) + 10.1$$

$$X = 24.32$$

Unit V: Index Numbers

Structure

Overview

Learning Objectives

- **Introduction to Index Numbers**
- **Construction of Price Index Numbers**
- **Unweighted Index Numbers**
- **Weighted Index Numbers**
- **Tests of Adequacy of Index Number Formulae**

Overview

Index Numbers are statistical tools that help us understand relative changes in variables such as prices, quantities, and values over time or space. This unit focuses on introducing index numbers, their characteristics, types, and construction. A significant emphasis is laid on methods for calculating both weighted and unweighted price index numbers. Practical problem-solving is central to this unit, which holds 80% weightage.

Learning Objectives

By the end of this unit, the learners will be able to:

- Define and understand the concept of index numbers.
- Identify and explain the characteristics and uses of index numbers.
- Distinguish between different types of index numbers.
- Apply appropriate methods to construct price index numbers.
- Evaluate the adequacy of different index number formulae using statistical tests.

5.1 Introduction

An index number is a statistical measure that shows changes in a variable or a group of related variables over time. It is typically expressed as a percentage, with a base year set at 100. Index numbers are particularly useful in economic and business studies to analyze trends, compare price levels and evaluate inflation.

5.1.1 Definition

An index number is a statistical measure designed to show changes in a variable (or group of variables) with reference to a base year.

5.2 Characteristics of Index Numbers

- Relative measure
- Percentage-based
- Base year = 100
- Purpose-specific
- Average of price relatives
- Time/location comparison

---

1. **Relative Measurement:** Index numbers show relative changes over time.
2. **Expressed in Percentages:** The base year is assigned 100; other values are relative to it.
3. **Averages of Ratios:** They are computed using various averages of price or quantity ratios.
4. **Purpose-Oriented:** Constructed for specific uses like tracking price or quantity changes.
5. **Comparability:** Allow for comparison across time or regions.
6. **Base Year Reference:** Comparisons are always made with respect to a base year.

5.3 Uses of Index Numbers

- Cost of living analysis (CPI)
- Business forecasting
- Policy formulation
- Wage revision
- Economic comparisons

1. **Measuring Inflation and Deflation:** Helps in analyzing price level fluctuations.
2. **Cost of Living Adjustments:** Used to revise wages and pensions.
3. **Policy Making:** Assists governments in forming fiscal and monetary policies.
4. **Business Forecasting:** Guides pricing and investment decisions.
5. **International Comparisons:** Compares economic conditions between countries.

5.4 Types of Index Numbers

1. **Price Index Numbers** – Measure changes in prices over time.

   e.g., CPI, WPI

2. **Quantity Index Numbers** – Measure changes in quantities consumed, produced, or sold.

   e.g., Industrial Output

3. **Value Index Numbers** – Reflect changes in total value (price × quantity).

   e.g., Total Sales Value

4. **Consumer Price Index (CPI)** – Measures retail prices of consumer goods and services.

   e.g., Cost of Living Index – retail price variations

5. **Wholesale Price Index (WPI)** – Measures average changes in wholesale prices.

   Specialized Index – e.g., Agricultural Price Index

5.5 Construction of Price Index Numbers

Steps in Construction:

1. **Selection of Base Year:** Should be a normal year.
2. **Selection of Items:** Representative commodities must be chosen.
3. **Collection of Prices:** Data for base and current year.
4. **Selection of Method:** Choose appropriate formula.

1. Price Index Numbers

Measure changes in prices over time.

Example: CPI, WPI

Problem:

Calculate the price index using the simple average of price relatives:

Base Year Prices: [100, 200, 300], Current Year Prices: [110, 220, 330]

Step-by-Step Solution:

Relative prices: $110/100\times100 = 110$, $220/200\times100 = 110$, $330/300\times100 = 110$

Average of relatives $= (110 + 110 + 110) / 3 = 110$

2. Quantity Index Numbers

Measure changes in quantities consumed, produced, or sold.

Example: Industrial Output

Problem:

Calculate the quantity index using the simple aggregate method:

$Q_0 = [50, 70, 100]$, $Q_1 = [60, 90, 120]$

Step-by-Step Solution:

$\Sigma Q_0 = 220$

$\Sigma Q_1 = 270$

Index $= (270 / 220) \times 100 = 122.73$

3. Value Index Numbers

Reflect changes in total value (price × quantity).

Example: Total Sales Value

Problem:

Given:

$P_0 = [10, 20]$, $Q_0 = [5, 10]$

$P_1 = [15, 25]$, $Q_1 = [6, 12]$

Step-by-Step Solution:

Base value $= 10\times5 + 20\times10 = 50 + 200 = 250$

Current value $= 15\times6 + 25\times12 = 90 + 300 = 390$

Value Index $= (390 / 250) \times 100 = 156$

4. Consumer Price Index (CPI)

Measures retail prices of consumer goods and services.

Example: Cost of Living Index

Problem:

Item Prices (base): [10, 20], Quantities: [5, 10], Prices (current): [12, 25]

Step-by-Step Solution:

$P_1Q_0 = 12\times5 + 25\times10 = 60 + 250 = 310$

$P_0Q_0 = 10\times5 + 20\times10 = 50 + 200 = 250$

$CPI = (310 / 250) \times 100 = 124$

5. Wholesale Price Index (WPI)

Measures average changes in wholesale prices.

Example: WPI for bulk commodities

Problem:

Base Prices: [50, 70, 80], Current Prices: [60, 90, 100]

Step-by-Step Solution:

$\Sigma P_0 = 200$

$\Sigma P_1 = 250$

$WPI = (250 / 200) \times 100 = 125$

6. Specialized Index - Agricultural Price Index

Indexes focusing on specific sectors like agriculture.

Example: Agricultural Price Index

Problem:

Base Prices: [100, 120], Current Prices: [130, 150]

Step-by-Step Solution:

Relative prices: $130/100\times100 = 130, 150/120\times100 = 125$

$Average = (130 + 125) / 2 = 127.5$

5.5.1 UNWEIGHTED INDEX NUMBERS

Definition:

Unweighted index numbers are statistical tools used to measure changes in variables (usually price or quantity) over time. In this method, all items are given equal weight or importance, regardless of their actual quantity or value.

Methods of Calculation:

1. Simple Aggregative Method

Formula: Index Number $= (\Sigma P_1 / \Sigma P_0) \times 100$

2. Simple Average of Price Relatives Method

Formula: Index Number $= (1/n) \times \Sigma(P_1 / P_0 \times 100)$

Where:

$P_1 =$ Price in current year

$P_0 =$ Price in base year

$n =$ Number of commodities

Solved Examples

Example 1: Simple Aggregative Method

Commodity	Price in 2020 (₹)	Price in 2025 (₹)
A	20	30
B	50	60
C	10	15

Index $= (\Sigma P_1 / \Sigma P_0) \times 100$

$= (105 / 80) \times 100 = 131.25$

Example 2: Simple Average of Price Relatives

Commodity	2020 (₹)	2025 (₹)
X	10	15
Y	20	30
Z	40	60

Index Number $= (1/n) \times \Sigma(P_1 / P_0 \times 100)$

$= (1/3) \times[(10/15)x100+(30/20)x100+(60/40)x100]$

$=1/3(150+150+150) = 150$

Example 3: Simple Aggregative Method

Commodity	Base Year (₹)	Current Year (₹)
Soap	25	35
Oil	100	120
Rice	50	65

Index $= (220 / 175) \times 100 = 125.71$

Example 4: Simple Average of Relatives

Item	Base Price (₹)	Current Price (₹)
A	40	44

B	80	88
C	100	110

Relatives = 110, 110, 110 → Index = 110

Example 5: Mixed Example

Commodity	2019 (₹)	2024 (₹)
Pen	10	12
Pencil	5	6
Paper	50	60
Ink	20	25

Aggregative Index = (103 / 85) × 100 = 121.18

Average of Relatives = 120, 120, 120, 125 → Index = 121.25

Example 6: Calculate the simple aggregate price index for the following:

Commodity	**Price in 2020 (P_0)**	**Price in 2024 (P_1)**
A	20	25
B	10	15
C	30	40

Solution:

Step 1: List the prices in the base year (P_0) and current year (P_1):

Commodity	Price in 2020 (P_0)	Price in 2024 (P_1)
A	20	25

B	10	15
C	30	40

Step 2: Calculate the sum of prices in each year:

$\Sigma P_0 = 20 + 10 + 30 = 60$

$\Sigma P_1 = 25 + 15 + 40 = 80$

Step 3: Apply the formula:

$P = (80 / 60) \times 100 = 133.33$

Therefore Index Number = 133.33

Interpretation:

The overall price level increased by 33.33% from 2020 to 2024.

2.Simple Average of Price Relatives Method

Formula:

$$P = \frac{\sum \left(\frac{P_1}{P_0} \times 100 \right)}{n}$$

Example 7:

Using the data above, calculate the index number using the average of price relatives.

Solution:

Commodity	P_1	P_0	$\frac{P_1}{P_0} \times 100$
A	25	20	125
B	15	10	150
C	40	30	133.33

Step 1: Calculate price relatives:

A: $(25 / 20) \times 100 = 125$

B: $(15 / 10) \times 100 = 150$

C: $(40 / 30) \times 100 = 133.33$

Step 2: Compute the average:

$P = (125 + 150 + 133.33) / 3 = 136.11$

Answer: Index Number = 136.11

5.5.2 Weighted index numbers

Definition:

Weighted index numbers are statistical tools that measure changes in prices or quantities over time, where each item is assigned a weight according to its relative importance or quantity consumed. Unlike unweighted index numbers, weighted indices provide a more accurate picture by considering how much of each item is used.

Methods of Calculation:

1. **Laspeyres Method**

Formula: Index $= (\Sigma P_1 \times Q_0 / \Sigma P_0 \times Q_0) \times 100$

2. **Paasche Method**

Formula: Index $= (\Sigma P_1 \times Q_1 / \Sigma P_0 \times Q_1) \times 100$

Where:

P_0 = Price in base year

P_1 = Price in current year

Q_0 = Quantity in base year

Q_1 = Quantity in current year

3. Fisher's Ideal Index:

$$P = \sqrt{\left(\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_1Q_1}{\sum P_0Q_1}\right)} \times 100$$

Example 1: Calculate **Laspeyres' Price Index**

Commodity	P_0	P_1	Q_0
A	10	12	5
B	8	10	10
C	12	15	8

Solution:

Step 1: Calculate P_1Q_0 and P_0Q_0 for each:

P_1Q_0: $12\times5 = 60$, $10\times10 = 100$, $15\times8 = 120 \rightarrow \sum P_1Q_0 = 280$

$$\sum P_1Q_0 = (12 \times 5) + (10 \times 10) + (15 \times 8) = 60 + 100 + 120 = 280$$

P_0Q_0: $10\times5 = 50$, $8\times10 = 80$, $12\times8 = 96 \rightarrow \sum P_0Q_0 = 226$

$$\sum P_0Q_0 = (10 \times 5) + (8 \times 10) + (12 \times 8) = 50 + 80 + 96 = 226$$

Step 2: Apply the formula:

$$P_L = \frac{\sum P_1Q_0}{\sum P_0Q_0} \times 100$$

$P_L = (280 / 226) \times 100 = 123.89$

Answer: Laspeyres' Index = 123.89

$$P_L = \frac{280}{226} \times 100 = 123.89$$

Example 2: Calculate **Paasche's Price Index**

Commodity	P_0	P_1	Q_1
A	10	12	6
B	8	10	12
C	12	15	7

Solution:

Step 1: Calculate P_1Q_1 and P_0Q_1 for each:

P_1Q_1: $12\times6 = 72$, $10\times12 = 120$, $15\times7 = 105 \rightarrow \sum P_1Q_1 = 297$

$$\sum P_1Q_1 = (12 \times 6) + (10 \times 12) + (15 \times 7) = 72 + 120 + 105 = 297$$

P_0Q_1: $10\times6 = 60$, $8\times12 = 96$, $12\times7 = 84 \rightarrow \sum P_0Q_1 = 240$

$$\sum P_0Q_1 = (10 \times 6) + (8 \times 12) + (12 \times 7) = 60 + 96 + 84 = 240$$

Step 2: Apply the formula:

$$P_P = \frac{\sum P_1Q_1}{\sum P_0Q_1} \times 100$$

$P_P = (297 / 240) \times 100 = 123.75$

Answer: Paasche's Price Index = 123.75

$$P_P = \frac{297}{240} \times 100 = 123.75$$

Example 3:Calculate **Fisher's Ideal Index**

Commodity	P_0	P_1	Q_0
A	10	12	5
B	8	10	10
C	12	15	8

(a) Write the formula for Fisher's Ideal Index.

(b) Calculate the Fisher's Ideal Index using the given values.

(c) Interpret the result.

Solution:

a)
$$P = \sqrt{\left(\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_1Q_1}{\sum P_0Q_1}\right)} \times 100$$

Step-by-Step Answer:

Step 1: **Laspeyres' Price Index**

$$\sum P_1Q_0 = (12 \times 5) + (10 \times 10) + (15 \times 8) = 60 + 100 + 120 = 280$$

$$\sum P_0Q_0 = (10 \times 5) + (8 \times 10) + (12 \times 8) = 50 + 80 + 96 = 226$$

$$P_L = \frac{\sum P_1Q_0}{\sum P_0Q_0} \times 100$$

$P_L = (280 / 226) \times 100 = 123.89$

Step 2: Calculate **Paasche's Price Index**

$$\sum P_1Q_1 = (12 \times 6) + (10 \times 12) + (15 \times 7) = 72 + 120 + 105 = 297$$

$$\sum P_0Q_1 = (10 \times 6) + (8 \times 12) + (12 \times 7) = 60 + 96 + 84 = 240$$

$$P_P = \frac{\sum P_1Q_1}{\sum P_0Q_1} \times 100$$

Step 3: Fisher's Ideal Index is the geometric mean of Laspeyres' and Paasche's Index Numbers.

$$P = \sqrt{\left(\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_1Q_1}{\sum P_0Q_1}\right)} \times 100$$

i.e., $PF = \sqrt{(PL \times PP)}$

Substitutes the value of $P_L = 123.89$, and $P_P = 123.75$

$P_F = \sqrt{(123.89 \times 123.75)} = \sqrt{(15334.74)}$

$= 123.82$ (approx)

Step 4: Interpretation

The Fisher's Ideal Index is 123.82, which means that on average, the price level has increased by 23.82% over the base year.

Since it is the geometric mean of two well-known indices, and it satisfies both major tests of adequacy, it is considered an ideal index.

Fisher's Ideal Index = 123.82

5.8 Test of Adequacy of Index Number Formulae:

Definition: Tests of adequacy (or consistency) of index number formulae are mathematical checks used to judge the reliability and validity of index number formulas. A good index number should satisfy the Time Reversal Test and Factor Reversal Test.

1. Time Reversal Test

A formula satisfies the Time Reversal Test if the product of the index number from time 0 to 1 and from 1 to 0 equals 100^2.

Formula: $P_{01} \times P_{10} = 10000$

✓ Satisfied by: Fisher's Ideal Index, Geometric Mean Method

✗ Not satisfied by: Laspeyres, Paasche, Simple Aggregative

2. Factor Reversal Test

A formula satisfies the Factor Reversal Test if the product of the price index and the quantity index equals the value index.

Formula: $P_{01} \times Q_{01} = V_{01} = (\Sigma P_1Q_1 / \Sigma P_0Q_0)$

Satisfied by: Fisher's Ideal Index

Not satisfied by: Laspeyres, Paasche, Simple Aggregative

3.**Circular Test:**

The product of the indices from period A to B, B to C, and C to A should be 1 (or 100 if using percentage form).

Only some indices satisfy this test, such as the geometric mean method.

Test	Requirement	Satisfied by
Time Reversal	$P_{01} \times P_{10} = 100$	Fisher
Factor Reversal	$P \times Q = V$	Fisher
Circular	$P_{01} \times P_{12} \times P_{20} = 100$	Not all

Example 4: Show that Fisher's index satisfies the time reversal test using the values:

(or)

The Time Reversal Test is a condition used to determine the consistency of an index number formula.
You are given the following value:
Fisher's Price Index from time 0 to 1 (P_{01}) = 123.82

(a) Define the Time Reversal Test.
(b) Calculate the reverse index (P_{10}).
(c) Show that the Fisher's Index satisfies the Time Reversal Test.
Answer:

Step 1: Definition
The Time Reversal Test states that if we compute the price index from time 0 to time 1 (P_{01}) and then reverse the computation from time 1 to time 0 (P_{10}), the product of the two indices should be equal to 100.
Mathematically:
$P_{01} \times P_{10} = 100$

Step 2: Given

$P_{01} = 123.82$

To find P_{10}, we use the reciprocal:

$P_{10} = 100 / P_{01} = 100 / 123.82 = 0.8074 \times 100 = 80.74$

Step 3: Verify the Test

$P_{01} \times P_{10} = 123.82 \times 80.74 = 10000.15$

Now, divide by 100 (as both were percentage indices):

10000.15 / 100 = 100 (approximately)

Conclusion:

Since $P_{01} \times P_{10} \approx 100$, the Fisher's Ideal Index satisfies the Time Reversal Test.

$$P_{01} = 123.82, \quad P_{10} = \frac{1}{1.2382} \times 100 = 80.74$$

$$P_{01} \times P_{10} = 123.82 \times 80.74 \approx 10000 \Rightarrow \frac{10000}{100} = 100\checkmark$$

Hence, **time reversal test satisfied**.

Example 5:

Show that Fisher's Ideal Index Number is known to satisfy both the Time Reversal Test and the Factor Reversal Test.

The following data given:

• Fisher's Price Index (P) = 123.82

• Fisher's Quantity Index (Q) = 118.5

(or)

(a) State the Factor Reversal Test condition for an ideal index number.

(b) Verify whether the above values satisfy the Factor Reversal Test.

(c) Interpret the meaning of your result in terms of value change.

Step-by-Step Answer:

Step 1: Recall the Factor Reversal Test condition:

The product of the Fisher's Price Index and Quantity Index should equal the Value Index.

Mathematically: P × Q = V (Value Index)

Step 2: Multiply the given values:

Fisher's Price Index (P) = 123.82

Fisher's Quantity Index (Q) = 118.5

P × Q = (123.82 × 118.5) / 100 = 146.73

Step 3: Interpretation:

The Value Index (V) represents the combined effect of changes in both prices and quantities.

The result 146.73 means that the total value has increased by approximately 46.73% over the base year.

Conclusion:

Since the product of the Price Index and Quantity Index equals the Value Index, the Factor Reversal Test is satisfied. Hence, Fisher's Index is validated as an ideal index number.

$$P \times Q = \frac{123.82 \times 118.5}{100} = 146.74 \Rightarrow \text{Value Index} = \frac{\sum P_1 Q_1}{\sum P_0 Q_0}$$

Questions and Answers

Q1. What is an index number?

A: It is a statistical measure showing the relative change in a variable over time, usually expressed as a percentage.

Q2. Write any two characteristics of index numbers.

(i) Expressed in percentages, (ii) Averages of ratios.

Q3. Give the formula for Fisher's Ideal Index.

$$P = \sqrt{\left(\frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times \frac{\sum P_1 Q_1}{\sum P_0 Q_1}\right)} \times 100$$

Q4. Differentiate between unweighted and weighted index numbers.

Unweighted indices give equal importance to all items; weighted indices assign importance based on quantities or other criteria.

Q5. What are the tests of adequacy for index number formulae?

Time reversal test, factor reversal test, and circular test.

Practice Exercises

1. Simple Aggregative Method

Item	Base Year Price (₹)	Current Year Price (₹)
Milk	30	36
Sugar	40	60
Rice	50	55
Oil	100	120

Task: Calculate the index number using the appropriate method.

2. Simple Average of Relatives

Item	Base Year Price (₹)	Current Year Price (₹)
Soap	25	35
Paste	40	50
Shampoo	60	72
Powder	30	36

Task: Calculate the index number using the appropriate method.

3. Both Methods

Item	Base Year Price (₹)	Current Year Price (₹)
A	100	120
B	150	180

C	200	250

Task: Calculate the index number using the appropriate method.

4. Simple Aggregative Method

Item	Base Year Price (₹)	Current Year Price (₹)
Shoes	800	1000
Belt	200	250
Shirt	500	650

Task: Calculate the index number using the appropriate method.

5. Simple Average of Relatives

Item	Base Year Price (₹)	Current Year Price (₹)
Tea	200	220
Coffee	400	500
Sugar	50	70
Milk	30	40

Check Your Progress-I

1. Define index numbers and list their uses.
2. What are the types of index numbers? Give examples.
3. Explain the steps in constructing price index numbers.
4. Differentiate between Laspeyres' and Paasche's methods.
5. Why is Fisher's index called the Ideal Index?

Check Your Progress-II

Problem 1: Calculate Laspeyres' Price Index

Given: $P_0 = [10, 8, 12]$, $P_1 = [12, 10, 15]$, $Q_0 = [5, 10, 8]$

Step-by-Step Answer:

$\Sigma P_1Q_0 = 280$

$\Sigma P_0Q_0 = 226$

$P_L = (280 / 226) \times 100 = 123.89$

Laspeyres' price Index=123.89

Problem 2: Calculate Laspeyres' Price Index

Given: $P_0 = [5, 7, 6]$, $P_1 = [6, 9, 8]$, $Q_0 = [10, 5, 15]$

Step-by-Step Answer:

$\Sigma P_1Q_0 = 225$

$\Sigma P_0Q_0 = 175$

$P_L = (225 / 175) \times 100 = 128.57$

Laspeyres' Price Index= 128.57

Problem 3: Calculate Paasche's Price Index

Given: $P_0 = [10, 8, 12]$, $P_1 = [12, 10, 15]$, $Q_1 = [6, 12, 7]$

Step-by-Step Answer:

$\Sigma P_1Q_1 = 297$

$\Sigma P_0Q_1 = 240$

$P_P = (297 / 240) \times 100 = 123.75$

Paasche's Price Index =123.75

Problem 4: Calculate Paasche's Price Index

Given: $P_0 = [5, 7, 6]$, $P_1 = [6, 9, 8]$, $Q_1 = [12, 8, 10]$

Step-by-Step Answer:

$\Sigma P_1Q_1 = 224$

$\Sigma P_0Q_1 = 176$

$P_P = (224 / 176) \times 100 = 127.27$

$=127.27$

Problem 5: Given $P_L = 123.89$ and $P_P = 123.75$. Compute Fisher's Ideal Index .

Step-by-Step Answer:

$P_F = \sqrt{(123.89 \times 123.75)} = \sqrt{15334.74} = 123.82$

Fisher's Ideal Index=123.82

Problem 6: Given PL = 128.57 and PP = 127.27. Compute Fisher's Ideal Index.

Step-by-Step Answer:

$P_F = \sqrt{(128.57 \times 127.27)} = \sqrt{16364.71} = 127.91$

Fisher's Ideal Index =127.91

Problem 7: Given $P_{01} = 123.82$, compute P_{10} and verify Time Reversal Test.

Step-by-Step Answer:

$P_{10} = 100 / 123.82 = 80.74$

$P_{01} \times P_{10} = 123.82 \times 80.74 = 10000.15 / 100 = 100$

Test satisfied

Problem 8: Given $P_{01} = 127.91$, compute P_{10} and verify Time Reversal Test.

Step-by-Step Answer:

$P_{10} = 100 / 127.91 = 78.17$

$P_{01} \times P_{10} = 127.91 \times 78.17 \approx 9999.9 / 100 \approx 100$

Test satisfied

Problem 9: Given P = 123.82 and Q = 118.5. Verify Factor Reversal Test.

Step-by-Step Answer:

$P \times Q = (123.82 \times 118.5) / 100 = 146.73$

Test satisfied, V = 146.73

Problem 10:Given P = 127.91 and Q = 115.4. Verify Factor Reversal Test.

Step-by-Step Answer:

$P \times Q = (127.91 \times 115.4) / 100 = 147.61$

Test satisfied, V = 147.61

Practice Exercises:

1. Calculate the weighted index number using the Laspeyres Method

Commodity	P_0 (₹)	P_1 (₹)	Q_0/Q_1 (Units)
X	5	6	10
Y	8	9	15
Z	4	5	12

2. Calculate the weighted index number using the Paasche Price Method

Commodity	P_0 (₹)	P_1 (₹)	Q_0/Q_1 (Units)
M	30	36	20
N	50	55	25
O	40	45	15

3. Calculate the weighted index number using the Laspeyres Method

Commodity	P_0 (₹)	P_1 (₹)	Q_0/Q_1 (Units)
Rice	60	75	40
Wheat	45	50	30
Oil	120	135	10

4. Calculate the weighted index number using the Paasche Method

Commodity	P_0 (₹)	P_1 (₹)	Q_0/Q_1 (Units)
Milk	25	30	50
Bread	15	18	40
Butter	40	45	20

5. Calculate the weighted index number using the .Laspeyres Price Method

Commodity	P_0 (₹)	P_1 (₹)	Q_0/Q_1 (Units)
Tea	100	120	10
Coffee	200	250	8
Sugar	50	60	20

6. Calculate the weighted index number using the Fisher's Ideal Index Method

Item	p0	q0	p1	q1
Item 1	5	4	6	5
Item 2	10	6	12	5

7 . Calculate the weighted index number using the Fisher's Ideal Index Method

Item	p0	q0	p1	q1
Item 1	3	6	4	7
Item 2	7	8	9	9

8. . Calculate the weighted index number using the Fisher's Ideal Index Method

Item	p0	q0	p1	q1
Item 1	2	10	3	12
Item 2	5	15	6	14

9. . Calculate the weighted index number using the Fisher's Ideal Index Method

Item	p0	q0	p1	q1
Item 1	8	5	10	6
Item 2	12	10	14	11

10. . Calculate the weighted index number using the Fisher's Ideal Index Method

Item	p0	q0	p1	q1
Item 1	4	7	5	8
Item 2	6	9	7	10

Let Us Sum Up

Index numbers are vital in economic and business studies for analyzing and comparing changes in variables like prices and quantities. They are constructed through well-defined steps using different methods. Among various formulas, Fisher's Ideal Index is widely appreciated for satisfying both time and factor reversal tests. Understanding and applying these concepts help in effective decision-making.

Glossary

Term	Meaning
Base Year	A reference year assigned index value 100.
Index Number	Statistical measure of relative change.
Laspeyres' Index	Uses base year quantities as weights.
Paasche's Index	Uses current year quantities as weights.
Fisher's Ideal Index	Geometric mean of Laspeyres and Paasche indices.
Time Reversal Test	Index should be consistent in reverse time calculation.
Factor Reversal Test	Price × Quantity index = Value index.