

BAHIRJI SMARAK MAHAVIDYALAYA

BASMATHNAGAR

FACULTY OF COMMERCE AND MANAGEMENT

B.Com. First Year | Semester I | Level 4.5

Syllabus as per NEP-2020, effective from Academic Year 2026–2027

BUSINESS STATISTICS

Complete Study Notes

BY DR NARSINGH

Course Code: CCOMMT1102

Discipline Specific Minor 2 (DSM) | 4 Credits | 60 Teaching Hours

Unit	Title	Hours
1	Introduction to Business Statistics	10
2	Presentation of Data by Diagrams and Graphs	12
3	Measures of Central Tendency	14
4	Measures of Dispersion	12
5	Correlation Analysis	12

Prepared for the students of B.Com. First Year

Course Overview and Syllabus

Course Objectives

1. To acquaint students with the basic principles of statistics as applicable in business.
2. To enable students to understand the types of data and their uses in business operations.
3. To make students understand the measures of central tendency and their application in solving business problems.
4. To make students understand the measures of dispersion and their application in business operations.

Course Outcomes

1. Students will be acquainted with the basic principles and utility of statistics in business operations.
2. Students will be able to upgrade their skills of presentation and interpretation of data with the help of tables, charts and diagrams.
3. Students will be able to utilise the basic concepts of central tendency in different business operations.
4. Students will be able to apply different measures of dispersion in business operations.

Detailed Syllabus

Module	Topics	Hrs
1. Introduction to Business Statistics	1.1 Introduction, definitions, origin and history of statistics; 1.2 Nature, scope and limitations; 1.3 Importance of Business Statistics; 1.4 Concept of data, types of data (primary and secondary), tools of primary data collection: questionnaire, schedule and interview	10
2. Presentation of Data by Diagrams and Graphs	2.1 Introduction, data presentation techniques; 2.2 Objects of diagram, rules and importance of diagrammatic presentation, types and limitations of diagrams; 2.3 Graphs: introduction, types and usage	12
3. Measures of Central Tendency	3.1 Meaning and concept of central tendency, averages; 3.2 Mean: definition, merits and demerits, arithmetic mean, combined mean; 3.3 Median: definition, merits, demerits; 3.4 Mode: definition, merits, demerits; numerical problems	14
4. Measures of Dispersion	4.1 Meaning and concepts of dispersion; 4.2 Quartile deviation and its coefficient; 4.3 Mean deviation and its coefficient; 4.4 Standard deviation and its coefficient; skewness: meaning and definition; numerical problems	12
5. Correlation Analysis	5.1 Meaning, definition and types of correlation; variables, types of variables; 5.2 Karl Pearson's coefficient of correlation; 5.3 Probable error; 5.4 Numerical problems	12

How to Use These Notes

- Read the **definitions** (blue boxes) and **formulae** (green boxes) first; these are the most frequently asked items.
- Work through every **solved example** with a pen and calculator before looking at the answer.
- Use the **Quick Revision Questions** at the end of each unit and the **MCQs** and **formula sheet** at the end of the booklet for revision before tests and the semester-end examination.

UNIT 1: Introduction to Business Statistics

Teaching hours: 10 | Topics: 1.1 Introduction, definitions, origin and history | 1.2 Nature, scope and limitations | 1.3 Importance of Business Statistics | 1.4 Concept of data, primary and secondary data, tools of primary data collection

1.1 Introduction to Statistics

Every business decision, whether it is launching a new product, fixing a price, hiring staff or granting a loan, depends on facts and figures. Statistics is the science that collects these figures, organises them, presents them clearly and draws sensible conclusions from them. In daily life we meet statistics everywhere: the cricket batting average, the percentage of marks, the monthly rainfall, the sensex movement, the census count.

The word **statistics** is used in two senses. In the **plural sense**, it means numerical facts or data (for example, the sales figures of a company for five years). In the **singular sense**, it means the science or method of dealing with such data: collection, classification, presentation, analysis and interpretation.

Origin of the word

The word 'statistics' is believed to have come from the Latin word '**status**', the Italian word '**statista**', the German word '**statistik**' or the French word '**statistique**', all of which mean a political state. In early times, statistics was used by governments (kings) to gather information about population, land and wealth for taxation and war. That is why statistics was once called the 'Science of Kings' or 'Science of Statecraft'.

Definitions of Statistics

Definitions are broadly of two kinds: those that treat statistics as data (plural sense) and those that treat it as a method (singular sense).

Definitions as DATA (plural sense)

Prof. Horace Secrist: 'Statistics are aggregates of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other.'

Prof. A. L. Bowley: 'Statistics are numerical statements of facts in any department of enquiry placed in relation to each other.'

Definitions as a METHOD / SCIENCE (singular sense)

Croxtton and Cowden: 'Statistics is the science of collection, presentation, analysis and interpretation of numerical data.'

Prof. A. L. Bowley: 'Statistics may be called the science of counting' and 'the science of averages'.

Wallis and Roberts: 'Statistics is not a body of substantive knowledge but a body of methods for obtaining knowledge.'

Of all these, the definition of **Croxtton and Cowden** is the most widely accepted because it covers all the stages of a statistical investigation. Horace Secrist's definition lists the essential characteristics of statistical data.

Characteristics of Statistics (data)

- Statistics are **aggregates of facts**; a single figure like 'Ravi scored 70' is not statistics.
- They are **affected by multiple causes**: sales depend on price, advertising, season, income and competition.
- They are **numerically expressed**; qualitative facts such as honesty must be converted into numbers (ranks, scores).
- They are **collected systematically** for a **predetermined purpose**.
- They are **reasonably accurate** (estimated or enumerated) and **comparable** with each other.

Origin and History of Statistics

The growth of statistics can be traced through the following stages.

Period	Development
Ancient times	Egypt, Babylon, Greece and Rome carried out head counts and land surveys for taxation and military recruitment (e.g., Egyptian pyramid-building census, about 3050 B.C.).
Ancient India	Kautilya's Arthashastra (about 300 B.C.) records data on population, agriculture, trade and taxes. The Ain-i-Akbari (Abul Fazl, Akbar's period) gives detailed administrative and revenue statistics.
17th century	John Graunt (England) studied birth and death records, known as the father of vital statistics. Sir William Petty developed 'Political Arithmetic'. Probability theory began with Pascal and Fermat.
18th–19th century	Gottfried Achenwall used the term 'Statistik'. Laplace and Gauss developed probability and the normal curve. Adolphe Quetelet applied statistics to social science.
20th century	Karl Pearson (correlation, chi-square), R. A. Fisher (design of experiments, ANOVA), W. S. Gosset ('Student', t-test). In India, P. C. Mahalanobis founded the Indian Statistical Institute (1931) and the National Sample Survey (1950).
Today	Computers, big data, data analytics, machine learning and business intelligence tools have made statistics central to business.

1.2 Nature, Scope and Limitations of Statistics

Nature of Statistics

The nature of statistics means its basic character: is it a science or an art? Statistics is **both**. It is a **science** because it has a systematic body of knowledge, uses scientific methods and is based on cause–effect study. It is an **art** because it requires skill in applying methods to practical problems. Further, statistics is an **applied science** whose methods are used in almost every field. Its main features are:

- It is **numerical**: deals with quantitative facts.
- It deals with **groups or aggregates**, not individual items.
- It is **mathematical and logical** in method but **practical** in use.
- It reveals **hidden patterns, trends and relationships** in data.
- It is a **tool for decision-making under uncertainty**; its conclusions are valid on average and in the long run, not for every individual case.

Scope of Statistics

Scope means the area of use. Statistics has two broad branches:

Descriptive Statistics	Inferential Statistics
Collects, organises, summarises and presents data	Draws conclusions about a population from a sample
Uses tables, graphs, averages, dispersion	Uses sampling, estimation, hypothesis testing, forecasting
Example: average sales of last 12 months	Example: predicting next year's sales from this year's sample

The practical scope of statistics covers many fields:

- **Business and industry**: production planning, quality control, marketing research, pricing and demand forecasting, inventory control.
- **Economics**: national income, index numbers, price trends, consumption and demand analysis.
- **Banking and finance**: risk analysis, credit scoring, portfolio return and variability, insurance premium calculation.
- **Government and planning**: census, five-year plans, budgets, poverty and employment surveys.

- **Accounting and auditing:** sampling in audit, cost analysis, budgetary control.
- **Science, agriculture, health and education:** experiments, crop yield studies, clinical trials, examination results.

Limitations of Statistics

Statistics is a powerful tool, but it is not perfect. Prof. Newsholme said, 'Statistics are the straws out of which, like other economists, you have to make bricks.' The main limitations are:

1. **Does not study individuals.** It deals with groups, so a single case has no statistical value.
2. **Deals only with numerical (quantitative) data.** Qualitative aspects like honesty, intelligence or beauty can be studied only after they are converted into numbers.
3. **True only on average.** Conclusions are not exact for each item; e.g., average income of a town does not mean everyone earns that amount.
4. **Can be misused.** Statistics can be made to prove anything: 'Figures do not lie, but liars can figure.'
5. **Only one of the methods.** Statistical results must be supported by other evidence and expert judgement.
6. **Results depend on the quality of data and the skill of the user.** Wrong data or a wrong method yields wrong conclusions.
7. **Requires expert handling.** Only a trained person should interpret statistical results.

1.3 Importance of Business Statistics

Business Statistics is the application of statistical methods to business problems: collecting data about customers, sales, costs, production, finance and the market; analysing them; and using the results for decisions. In a competitive and uncertain market, a businessman who relies on guesswork is likely to fail, while one who relies on data can reduce risk and increase profit.

Area	Use of Statistics
Production	Quality control charts, deciding the production volume, machine efficiency, defect rates
Marketing	Market surveys, consumer preference, sales forecasting, effect of advertising on sales
Finance	Analysing profit and risk, ratio and trend analysis, budgeting, investment decisions
Human Resources	Wage and productivity analysis, labour turnover, absenteeism, manpower planning
Purchasing & Stores	Optimum order quantity, stock levels, supplier performance
Planning & Control	Setting targets, comparing actual with standard, spotting deviations early
Banking & Insurance	Loan default risk, premium calculation using mortality and accident rates

- It **simplifies complex data** into averages, percentages and graphs that managers can understand.
- It helps in **planning and forecasting** future demand, sales and prices.
- It **facilitates comparison** between firms, products, periods and regions.
- It helps in **policy formulation and decision-making** with reduced risk.
- It helps to **test the relationship between variables** (price and demand, advertising and sales).
- It enables **control and efficiency** by comparing actual performance with targets.

1.4 Concept of Data

Data are the raw facts, figures and observations collected for a purpose. Data become **information** when they are processed and organised in a meaningful way, and information becomes **knowledge** when it is used for decisions. Data may be **quantitative** (numerical, such as sales in rupees, height) or **qualitative** (descriptive, such as gender, brand preference). Quantitative data can be **discrete** (counted: number of employees) or **continuous** (measured: weight, time).

Types of Data: Primary and Secondary

Primary Data

Data collected **first-hand, afresh and for the first time** by the investigator for a specific purpose. They are original in character. Example: a company interviewing 200 customers about a new soap.

Secondary Data

Data that were **already collected by someone else** for some other purpose and are now used by the investigator. They have already passed through a statistical process. Example: using NSSO reports, RBI Bulletin, Census of India, company annual reports or a published research paper.

Basis	Primary Data	Secondary Data
Meaning	Collected first-hand by the investigator	Taken from already available sources
Originality	Original	Not original (already processed)
Cost, time, effort	High	Low
Reliability	More reliable and suited to purpose	May not suit purpose; reliability must be verified
Methods	Questionnaire, schedule, interview, observation	Published and unpublished sources
Precaution	Training of investigators	Check suitability, adequacy and reliability

Sources of Secondary Data

- **Published sources:** government publications (Census, Economic Survey, RBI Bulletin, Statistical Abstract), reports of committees and commissions, publications of trade associations and chambers of commerce, newspapers and journals, research papers, company reports, international bodies (UN, World Bank, IMF).
- **Unpublished sources:** internal records of firms, diaries, letters, research theses and registers kept by organisations.

Precautions in using secondary data

Before using secondary data check: (1) **Reliability** of the source, (2) **Suitability** for the present purpose, (3) **Adequacy** of the data, and (4) the **definitions and units** used by the original collector.

Tools of Primary Data Collection

(A) Questionnaire

A questionnaire is a **set of written questions** relating to the problem, printed or mailed to respondents, who fill in the answers themselves and return it. It is used when respondents are literate and widely spread.

- **Types of questions:** open-ended (free answer), closed-ended (yes/no, multiple choice), and rating-scale questions.
- **Merits:** economical, covers a wide area, gives respondents time to think, no interviewer bias, privacy.
- **Demerits:** low response rate, not suitable for illiterate persons, no scope to clarify doubts, answers may be incomplete or wrong.

Qualities of a good questionnaire

Few, short, simple and clear questions; logical order (easy questions first); no ambiguous or leading questions; no personal or embarrassing questions; covering all aspects of the problem; with a covering letter and instructions; pre-tested by a pilot survey.

(B) Schedule

A schedule is also a **set of questions**, but it is filled in by a **trained enumerator** who visits the informants, asks questions and records the answers. It is useful when informants are illiterate.

- **Merits:** high response rate, suitable for illiterate people, accurate answers because the enumerator clarifies doubts, observation of respondent possible.
- **Demerits:** costly and time-consuming, needs trained enumerators, risk of enumerator bias, limited area.

Difference between Questionnaire and Schedule

Basis	Questionnaire	Schedule
Filled by	Respondent himself	Enumerator
Sent by	Post / email / online	Enumerator visits personally
Respondent	Must be literate	May be illiterate
Response rate	Low	High
Cost and time	Less	More
Area covered	Wide	Limited
Accuracy	Doubtful	More accurate

(C) Interview

An interview is a **face-to-face or telephonic conversation** between the investigator and the respondent in which questions are asked orally and answers are recorded. It may be **personal (direct)**, **telephonic**, **structured** (fixed questions) or **unstructured** (free discussion), and **individual** or **group**.

- **Merits:** direct contact, flexibility, deeper and more accurate information, the interviewer can observe facial expressions and clarify doubts, high response rate.
- **Demerits:** expensive and time-consuming, interviewer's bias, respondent may hide facts, requires skilled interviewers.

Other methods (for knowledge)

Direct personal observation (investigator watches and records, e.g., traffic count), **through local agents or correspondents** (used by newspapers), and **telephonic/online surveys** (Google Forms) are also used. These are not in the syllabus tools list but are useful in examples.

Unit 1: Quick Revision Questions**Q1. Define statistics and state its characteristics.**

Statistics, according to Croxton and Cowden, is the science of collection, presentation, analysis and interpretation of numerical data. Its characteristics are: aggregate of facts, affected by multiple causes, numerically expressed, systematically collected for a predetermined purpose, reasonably accurate and comparable.

Q2. Distinguish between primary and secondary data.

Primary data are first-hand, original and collected for a specific purpose, though costly and slow; secondary data are already available, collected earlier by others, cheap and quick, but must be checked for reliability and suitability.

Q3. State any four limitations of statistics.

Statistics does not study individuals, deals only with numerical data, is true only on average and can be misused by unskilled or dishonest persons.

UNIT 2: Presentation of Data by Diagrams and Graphs

Teaching hours: 12 | Topics: 2.1 Introduction, data presentation techniques | 2.2 Objects of diagrams, rules and importance of diagrammatic presentation, types of diagrams, limitations | 2.3 Graphs: introduction, types and usage

2.1 Introduction and Data Presentation Techniques

Raw data collected from a survey are usually bulky, scattered and difficult to understand. Before they can be analysed, they must be **edited, classified and presented** in a form that reveals their main features at a glance. Presentation of data is therefore the step between collection and analysis. A good presentation saves time, attracts attention, and makes comparison easy.

There are **three** main methods of presenting data:

Method	Description	Example
1. Textual (Descriptive)	Data are presented in the form of a paragraph or sentences	'Sales of the firm rose from Rs. 40 lakh in 2021 to Rs. 85 lakh in 2025.'
2. Tabular	Data are arranged in rows and columns in a systematic manner	Frequency table, classified table
3. Diagrammatic and Graphic	Data are shown through pictures, bars, circles, curves	Bar diagram, pie chart, histogram, ogive

Classification of data

Classification is the process of arranging data into groups or classes according to common characteristics. Bases of classification are: **geographical** (by place), **chronological** (by time), **qualitative** (by attribute such as gender) and **quantitative** (by magnitude such as marks or income).

Tabulation of data

Tabulation is the systematic arrangement of classified data in rows and columns. The main parts of a table are: table number, title, caption (column headings), stub (row headings), body, head-note, and foot-note/source note. A good table is simple, clear, has a suitable title and gives units.

Frequency distribution: key terms

Variable: a characteristic that varies (marks, income). **Frequency (f):** number of times a value occurs.

Class interval: e.g. 10–20. **Class limits:** lower (10) and upper (20) limits. **Class width (h):** upper limit minus lower limit.

Mid-value (m): $(\text{lower limit} + \text{upper limit}) / 2$. **Cumulative frequency:** running total of frequencies.

Exclusive method: the upper limit is excluded (a value 20 goes to class 20–30). **Inclusive method:** both limits are included (10–19, 20–29).

Example of a frequency distribution that will be used again in Units 3 and 4: marks of 50 students.

Marks (class)	0–10	10–20	20–30	30–40	40–50	Total
No. of students (f)	5	8	15	16	6	50
Mid-value (m)	5	15	25	35	45	–
Cumulative frequency (c.f.)	5	13	28	44	50	–

2.2 Diagrammatic Presentation of Data

A **diagram** is a visual form of presenting statistical data in which the numerical facts are shown through geometrical figures such as bars, squares, circles or pictures. Diagrams are drawn on plain paper and need no graph paper or scale of x–y axes (unlike graphs).

Objects (Purposes) of Diagrams

- To make complex data **simple and easily understandable** to common people.
- To present data in an **attractive and impressive** manner.
- To enable **quick comparison** between two or more sets of data.
- To help **memory**: a picture is remembered longer than a table.
- To convey facts to **illiterate persons** and the general public, for publicity and advertising.

Importance of Diagrammatic Presentation

- Gives a **bird's-eye view** of the entire data at a glance.
- Saves time and effort of studying figures.
- Is useful in **business reports, exhibitions, annual reports and advertisements**.
- Reveals **trends, ups and downs and relationships** easily.
- Has a **universal appeal**: understood by people of all levels of education.
- Helps in **checking the accuracy** of data by showing abnormal figures.

General Rules for Constructing Diagrams

1. Every diagram should have a **suitable title** that is clear, brief and placed at the top or bottom.
2. The diagram must be **neat, clean and attractive** and drawn to a **suitable scale** (scale should be indicated).
3. Select a **proper type** of diagram suited to the nature of data.
4. **Proportion**: the length and width should be in a good ratio (neither too tall nor too wide).
5. Different shades, colours or patterns should be used for different components, with an **index/legend** given.
6. The **source** of data and **footnotes** must be given.
7. The diagram should be **simple**, not overloaded, and the figures may be written on the diagram.
8. Bars should be of **equal width** with **equal gaps** between them.

Types of Diagrams

Diagrams are classified as: (A) **One-dimensional** (bar diagrams), (B) **Two-dimensional** (rectangles, squares, circles/pie), (C) **Three-dimensional** (cubes, cylinders), and (D) **Pictograms and Cartograms**.

(A) One-dimensional diagrams: Bar diagrams

In a bar diagram only the **length (height)** of the bar is considered; the width is kept the same. The bars may be vertical or horizontal.

1. Simple Bar Diagram: used to represent a single set of data (one variable), such as yearly sales or production. Bars are drawn on a common base, equal in width, and the length is proportional to the value. They may be arranged in chronological order or in ascending/descending order of size.

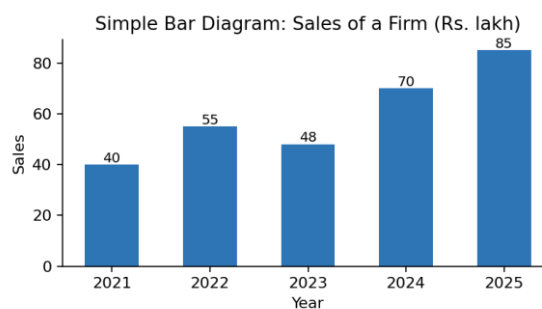


Figure 2.1 Simple bar diagram

2. Multiple (Compound) Bar Diagram: used to compare **two or more related** sets of data (e.g., sales and profit for each year). The bars of the related variables are drawn side by side in a group, each with a different shade, and a gap is left between groups.

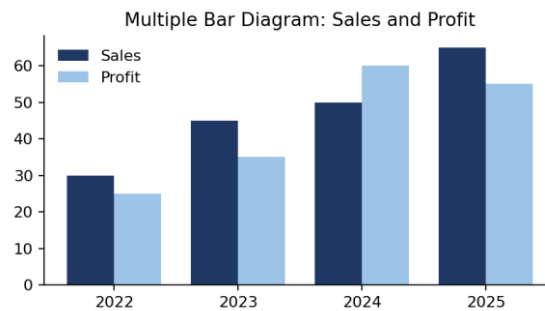


Figure 2.2 Multiple bar diagram

3. Sub-divided (Component) Bar Diagram: each bar is divided into **parts (components)** to show the total as well as the break-up (e.g., total sales split into products A, B and C). Different shades are used for each component.

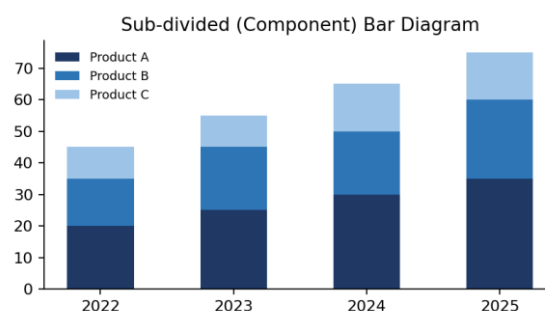


Figure 2.3 Sub-divided bar diagram

4. Percentage Sub-divided Bar Diagram: used when relative (percentage) composition is more important than absolute values. Every bar has the same height (100%), and the components are shown as percentages of the total. It is suitable to compare the structure of different groups.

5. Deviation (Bilateral) Bar Diagram: used for data having **both positive and negative values** such as profit and loss, or net exports; bars are drawn above and below the base line.

Simple vs multiple vs component

Simple bar: one variable. Multiple bar: two or more variables compared side by side. Component bar: total and its parts in one bar.

(B) Two-dimensional diagrams

In these diagrams, both **length and width (area)** are considered. The main forms are rectangles, squares and circles.

Rectangle diagram: area of the rectangle is proportional to the value. **Square diagram:** used when the values differ widely; side of the square is proportional to the square root of the value, i.e., side = $\sqrt{\text{value}}$. **Circle diagram:** radius is proportional to the square root of the value.

Pie diagram (Angular diagram): A circle is divided into sectors, each sector representing a component, with the angle of the sector proportional to its value. It is used to show the **break-up of a total** into parts and is very popular for showing a budget, cost structure or market share.

Angle of a sector in a pie diagram

Angle = $(\text{Value of the component} / \text{Total}) \times 360^\circ$

The total of all angles must equal 360° . Start the first sector at the 12 o'clock position and move clockwise, arranging the sectors in descending order of size if possible.

Example 1. A family spends its monthly income of Rs. 20,000 as follows: Food Rs. 8,000, Rent Rs. 5,000, Education Rs. 3,000, Clothing Rs. 2,000, Others Rs. 2,000. Draw a pie diagram.

Item	Expenditure (Rs.)	Calculation	Angle (degrees)
Food	8,000	$8000/20000 \times 360$	144°
Rent	5,000	$5000/20000 \times 360$	90°
Education	3,000	$3000/20000 \times 360$	54°
Clothing	2,000	$2000/20000 \times 360$	36°
Others	2,000	$2000/20000 \times 360$	36°
Total	20,000		360°

Pie Diagram: Family Budget (Rs. 20,000)

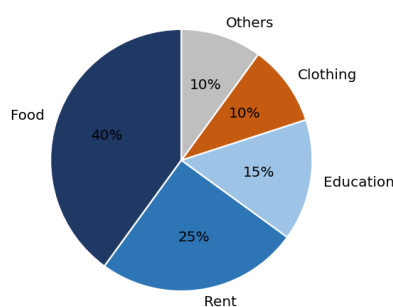


Figure 2.4 Pie diagram of the family budget

Merits of pie diagram: easy to understand, shows the proportion of each part to the whole, attractive. **Demerits:** difficult to compare many components or several pies; needs calculation of angles; not suitable when there are too many parts.

(C) Three-dimensional, pictograms and cartograms

- **Three-dimensional (cubes, cylinders, spheres):** volumes are used to show data with very large differences; side of cube = cube root of the value. These are rarely used because the volume is hard to judge visually.
- **Pictograms:** data are shown by **small pictures or symbols**, each symbol standing for a fixed quantity (e.g., one bus picture = 1,000 passengers). They are attractive and good for illiterate people but give only approximate figures.
- **Cartograms (statistical maps):** data are shown on a **map** by shades, dots or figures, e.g., population density of states, rainfall in districts.

Limitations of Diagrams

1. Diagrams give only **approximate values**; exact figures cannot be read.
2. They can show **only limited information**; detailed data cannot be presented.
3. They **can be misleading** if the scale is distorted or the base is not zero.
4. They **cannot be used for further statistical analysis** (mean, correlation etc.).
5. Different types need **skill and time** to draw; wrong choice of diagram can cause confusion.
6. They are **not suitable for experts** who need accurate data, and give no scope for complete analysis.

2.3 Graphs

A **graph** is a visual presentation of data drawn on **graph paper** using the x-axis (horizontal) and y-axis (vertical). Unlike diagrams, graphs are used for **precise** study of data, to **find values** (such as median, mode) and to study **relationships** (such as time series and frequency distributions).

Basis	Diagram	Graph
Drawn on	Plain paper	Graph paper

Basis	Diagram	Graph
Use	Popular appeal, comparison	Statistical analysis, finding values
Accuracy	Approximate	More accurate
Examples	Bar, pie, pictogram	Histogram, polygon, ogive, line graph
Audience	General public	Statisticians, researchers

Types of Graphs

(I) **Graphs of time series (historigrams):** time is plotted on the x-axis and the variable on the y-axis, and the points are joined by straight lines. This is called a **line graph**. It shows the trend over time (sales over years); multiple lines can compare several series.

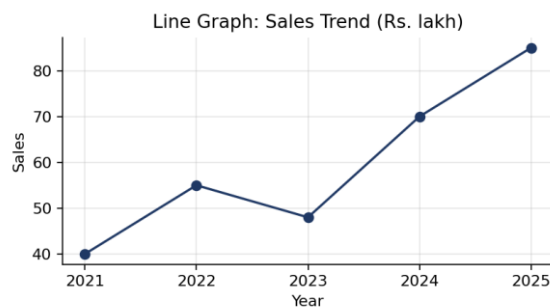


Figure 2.5 Line graph (time series)

(II) Graphs of frequency distributions:

1. **Histogram:** a set of **adjacent rectangles** erected on class intervals (x-axis) with heights proportional to frequencies (y-axis). There are no gaps between the bars because the data are continuous. For unequal class widths, frequency density (f / width) is plotted. The **mode** can be located from the histogram.
2. **Frequency polygon:** obtained by joining the **mid-points of the tops of the bars** of a histogram by straight lines, closing the figure at both ends at the x-axis. It is useful to compare two or more distributions on the same graph.
3. **Frequency curve:** a smoothed freehand polygon.
4. **Cumulative frequency curves (Ogives):** a graph of cumulative frequencies against class limits. A '**less than**' ogive is drawn by plotting upper limits against 'less than' cumulative frequencies (rising curve), and a '**more than**' ogive is drawn by plotting lower limits against 'more than' cumulative frequencies (falling curve). The point where the two intersect gives the **median**; also, quartiles, deciles and percentiles can be read from the ogive.

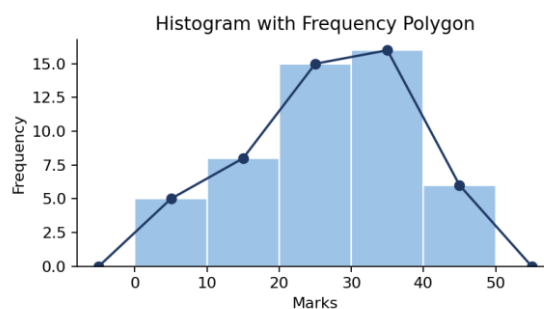


Figure 2.6 Histogram and frequency polygon (marks of 50 students)

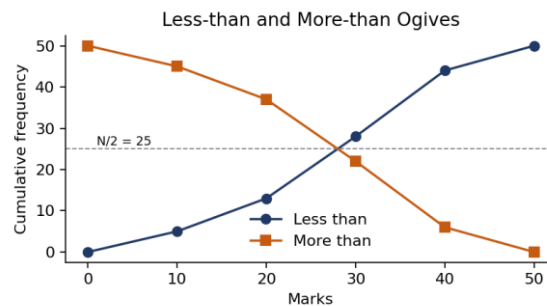


Figure 2.7 Less-than and more-than ogives

Usage of Graphs

- To study the **trend and fluctuations** of business variables like sales, profit or price over time.
- To **compare** two or more series on the same graph.
- To **locate positional values and mode**: the median from the ogive, mode from the histogram.
- To study **relationship between variables** (scatter diagram in Unit 5).
- To **forecast** by extending the trend line, and to detect abnormal values.
- Used by managers in **sales reports, production control charts and stock-market analysis**.

Unit 2: Quick Revision Questions

Q1. What is a diagram? State its objects and limitations.

A diagram is a visual device showing data through geometrical figures such as bars and circles. Its objects are to simplify data, attract attention, enable comparison and help memory. Its limitations are that it gives only approximate values, can mislead, shows limited detail and cannot be used for further analysis.

Q2. Distinguish between a histogram and a bar diagram.

A histogram has adjacent rectangles with no gaps and shows continuous frequency distributions (both width and height are meaningful). A bar diagram has separate bars with gaps and only the length matters.

Q3. How is a pie diagram constructed?

Find the total, compute each component's angle = $(\text{value} / \text{total}) \times 360$, draw a circle, and mark sectors with a protractor starting from the 12 o'clock position and moving clockwise; shade each sector and give an index.

UNIT 3: Measures of Central Tendency

Teaching hours: 14 | Topics: 3.1 Meaning of central tendency, averages | 3.2 Mean: definition, merits, demerits, arithmetic mean, combined mean | 3.3 Median | 3.4 Mode | Numerical problems

3.1 Meaning of Central Tendency and Averages

When a large body of data is collected, it is useful to find **one single value** that represents the whole group. Such a value tends to lie near the centre of the data, around which the other values cluster. It is called a **measure of central tendency** or an **average**. For example, instead of listing the marks of 500 students, we say 'the average mark is 54'.

Definition

Croxtan and Cowden: 'An average value is a single value within the range of the data that is used to represent all of the values in the series. Since an average is somewhere within the range of the data, it is sometimes called a measure of central value.'

Objectives (functions) of an average

- To get a **concise summary** of bulky data.
- To **facilitate comparison** between two or more groups.

- To help in **decision making** and policy formulation.
- To act as a **basis for further analysis**, such as dispersion and correlation.

Requisites of a good average

- It should be **rigidly defined** (no ambiguity) and **easy to understand and calculate**.
- It should be **based on all observations**.
- It should be **not unduly affected by extreme values**.
- It should have **sampling stability** and be **capable of further algebraic treatment**.

Types of averages

Mathematical averages	Positional averages
Arithmetic Mean (AM)	Median
Geometric Mean (GM)	Mode
Harmonic Mean (HM)	Quartiles, Deciles, Percentiles

In this syllabus we study the arithmetic mean, the median and the mode.

3.2 Arithmetic Mean (Mean)

Definition

The arithmetic mean is the value obtained by **dividing the sum of all observations by the number of observations**. It is the most commonly used average and is denoted by $\bar{X}$ (X-bar).

Calculation of AM

Formulae of Arithmetic Mean

Individual series (direct): $\bar{X} = \Sigma X / N$

Discrete series: $\bar{X} = \Sigma fX / \Sigma f$ ($= \Sigma fX / N$)

Continuous series (direct): $\bar{X} = \Sigma fm / N$, where m = mid-value of the class

Short-cut (assumed mean) method: $\bar{X} = A + \Sigma fd / N$, where $d = X - A$ (or $m - A$)

Step-deviation method: $\bar{X} = A + (\Sigma fd' / N) \times h$, where $d' = (m - A)/h$ and h = class width

Example 1 (Individual series). Find the mean of 10, 20, 30, 40, 50.

$$\bar{X} = \Sigma X / N = (10 + 20 + 30 + 40 + 50) / 5 = 150 / 5 = 30.$$

Example 2 (Discrete series). Find the mean of the following data.

X	10	20	30	40	50	Total
f	5	8	12	10	5	40
fX	50	160	360	400	250	1,220

$$\bar{X} = \Sigma fX / N = 1220 / 40 = 30.5.$$

Example 3 (Continuous series). Calculate the mean marks of the 50 students given in Unit 2 using the direct method.

Marks	Mid-value (m)	f	fm
0–10	5	5	25
10–20	15	8	120
20–30	25	15	375

Marks	Mid-value (m)	f	fm
30–40	35	16	560
40–50	45	6	270
Total		N = 50	Σfm = 1,350

$\bar{X} = \Sigma fm / N = 1350 / 50 = 27$ marks.

Same example by the step-deviation method. Take A = 25 and h = 10.

Marks	m	f	d' = (m – 25)/10	fd'
0–10	5	5	–2	–10
10–20	15	8	–1	–8
20–30	25	15	0	0
30–40	35	16	1	16
40–50	45	6	2	12
Total		50		Σfd' = 10

$\bar{X} = A + (\Sigma fd' / N) \times h = 25 + (10 / 50) \times 10 = 25 + 2 = 27$. Both methods give the same result, but the step-deviation method saves time when figures are large.

Combined mean

If the means and sizes of two or more groups are known, the mean of all groups taken together is the **combined mean**.

Combined Mean

$$\bar{X}_{12} = (N_1\bar{X}_1 + N_2\bar{X}_2) / (N_1 + N_2)$$

$$\text{For three groups: } \bar{X}_{123} = (N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3) / (N_1 + N_2 + N_3)$$

Example 4. The mean wage of 50 workers in Plant A is Rs. 60 and of 30 workers in Plant B is Rs. 70 (in hundreds). Find the combined mean wage.

$$\bar{X}_{12} = (50 \times 60 + 30 \times 70) / (50 + 30) = (3000 + 2100) / 80 = 5100 / 80 = \text{Rs. } 63.75 \text{ (hundreds), i.e., Rs. 6,375.}$$

Example 5 (Correction of mean). The mean of 100 items was found to be 50. Later it was discovered that an item 45 was wrongly read as 54. Find the correct mean.

$$\text{Incorrect } \Sigma X = 100 \times 50 = 5000. \text{ Correct } \Sigma X = 5000 - 54 + 45 = 4991. \text{ Correct mean} = 4991 / 100 = 49.91.$$

Example 6 (Missing frequency). The mean of the data below is 27; find the missing frequency f.

Class	0–10	10–20	20–30	30–40	40–50
Frequency	5	8	f	16	6
Mid-value	5	15	25	35	45

$N = 35 + f$, and $\Sigma fm = 25 + 120 + 25f + 560 + 270 = 975 + 25f$. Since $\text{Mean} = \Sigma fm / N$, we get $975 + 25f = 27(35 + f) = 945 + 27f$. Therefore $30 = 2f$, so $f = 15$. Method: always form the equation $\Sigma fm = \text{Mean} \times N$ and solve for the unknown.

Merits of Arithmetic Mean

- Simple to understand and easy to calculate; rigidly defined.
- Based on **all observations**.
- Capable of **further algebraic treatment** (combined mean, dispersion, correlation).
- Least affected by sampling fluctuations; most stable average.
- The sum of deviations of items from the mean is always **zero** ($\Sigma(X - \bar{X}) = 0$).

Demerits of Arithmetic Mean

- **Highly affected by extreme values** (a few very large or small items).
- Cannot be calculated for **open-end classes** without an assumption.
- Cannot be located graphically or by inspection.
- May give a **value that does not exist** in the data (e.g., 2.5 children per family).
- Cannot be used for qualitative data; may be misleading in highly skewed data.

3.3 Median (Positional Average)

Definition

The median is the **middle value** of a series when the items are arranged in **ascending or descending order**. It divides the series into two equal parts: 50% of the items are below it and 50% above it. **Prof. Connor:** 'The median is that value of the variable which divides the group into two equal parts, one part comprising all values greater, and the other all values less than the median.'

Formulae of Median

Individual series: Median = size of $((N + 1)/2)$ th item (after arranging in order).

Discrete series: Median = size of $((N + 1)/2)$ th item, found from the cumulative frequency (c.f.).

Continuous series: Median = $L + [(N/2 - c.f.) / f] \times h$

where L = lower limit of the median class, c.f. = cumulative frequency of the class preceding the median class, f = frequency of the median class, h = class width, and the median class is the class in which the $(N/2)$ th item lies.

Example 7 (Individual series, odd N). Find the median of 12, 7, 15, 9, 20, 11, 18.

Arranged: 7, 9, 11, 12, 15, 18, 20. $N = 7$. Median = $(7 + 1)/2 = 4$ th item = **12**.

Example 8 (Individual series, even N). Find the median of 5, 8, 12, 15, 20, 25. $N = 6$; median = $(6 + 1)/2 = 3.5$ th item = average of 3rd and 4th = $(12 + 15)/2 = \mathbf{13.5}$.

Example 9 (Discrete series).

X	10	20	30	40	50
f	4	7	15	8	6
c.f.	4	11	26	34	40

$N = 40$. $(N + 1)/2 = 20.5$ th item. The first c.f. that is greater than or equal to 20.5 is 26, which corresponds to $X = 30$. Median = **30**.

Example 10 (Continuous series). Find the median marks of the 50 students (data from Unit 2).

c.f. values: 5, 13, 28, 44, 50. $N/2 = 25$. The first c.f. greater than 25 is 28, so the **median class is 20–30**. $L = 20$, c.f. (preceding) = 13, $f = 15$, $h = 10$.

Median = $20 + [(25 - 13) / 15] \times 10 = 20 + (12/15) \times 10 = 20 + 8 = \mathbf{28 \text{ marks}}$.

Merits of Median

- Easy to understand and calculate; **not affected by extreme values**.
- Can be found **graphically** (ogive) and even when the data are incomplete, if the middle items are known.
- Suitable for **open-end classes** and for **qualitative data** that can be ranked (honesty, ability).
- The sum of absolute deviations from the median is the **minimum**.

Demerits of Median

- Not based on all observations; ignores the magnitude of extreme items.
- Needs arranging data in order: tedious for large data.
- Not capable of further algebraic treatment; affected by sampling fluctuation more than the mean.
- In an even series, it may not be an actual item.

3.4 Mode

Definition

The mode is the value that **occurs most frequently** in a series, i.e., the value having the **maximum frequency**. It is the most 'fashionable' or typical value. **Croxton and Cowden**: 'The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated.' Business uses: the most common shoe size or the best-selling product.

Mode: methods and formulae

Individual series: the value repeated the maximum number of times.

Discrete series: the value with the highest frequency (inspection); if not clear, use the grouping method.

Continuous series: $\text{Mode} = L + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2} \right] \times h$

where L = lower limit of the modal class (class with maximum frequency), $\Delta_1 = f_1 - f_0$, $\Delta_2 = f_1 - f_2$, f_1 = frequency of modal class, f_0 = frequency of the class before it, f_2 = frequency of the class after it, h = class width.

Empirical relationship (moderately skewed data): $\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$

Example 11 (Individual series). Find the mode of 3, 5, 2, 5, 6, 5, 7, 2. The value **5** occurs three times, which is the maximum. Mode = **5**.

Example 12 (Discrete series). X : 5, 6, 7, 8, 9 with f : 4, 9, 15, 10, 3. The maximum frequency is 15, corresponding to $X = 7$, so Mode = **7**.

Example 13 (Continuous series). Find the modal marks of the 50 students.

The maximum frequency is 16, so the **modal class is 30–40**. $L = 30$, $f_1 = 16$, $f_0 = 15$, $f_2 = 6$, $h = 10$. $\Delta_1 = 16 - 15 = 1$; $\Delta_2 = 16 - 6 = 10$.

Mode = $30 + \left[\frac{1}{1 + 10} \right] \times 10 = 30 + 0.91 = \mathbf{30.91 \text{ marks}}$ (approximately).

Measure	Value for the 50-student data
Mean	27.00 marks
Median	28.00 marks
Mode	30.91 marks
Check: $3 \text{ Median} - 2 \text{ Mean}$	$3 \times 28 - 2 \times 27 = 30$ (close to the calculated mode)

Merits of Mode

- Simple and popular; can be located by **inspection** or graphically (histogram).
- **Not affected by extreme values**; can be used for open-end classes.
- Gives the most typical value, useful for qualitative data (most popular brand).

Demerits of Mode

- Not rigidly defined: a distribution may have **no mode** or **more than one mode** (bimodal, multimodal).
- **Not based on all items** and not capable of further algebraic treatment.

- Affected by the choice of class intervals; not suited to small data.

Comparison of Mean, Median and Mode

Basis	Mean	Median	Mode
Meaning	Sum ÷ number of items	Middle value	Most frequent value
Based on all items	Yes	No	No
Affected by extremes	Highly	No	No
Algebraic treatment	Possible	Not possible	Not possible
Graphical location	Not possible	Ogive	Histogram
Best use	Symmetrical data, further analysis	Skewed data, open-end classes	Most typical / popular item

Additional Numerical Problems (with Answers)

Problem 1. Find the mean of 12, 18, 24, 30, 36, 42. **Answer:** $\Sigma X = 162$, $N = 6$, mean = **27**.

Problem 2. The daily wages (Rs.) of 7 workers are 300, 350, 400, 450, 500, 550, 600. Find the mean and the median. **Answer:** mean = $3150/7 = 450$; median = 4th item = **450**.

Problem 3. Find the combined mean of two groups: 40 students with mean 55 and 60 students with mean 65. **Answer:** $(40 \times 55 + 60 \times 65)/100 = (2200 + 3900)/100 = 61$.

Problem 4. Find the mode of: 4, 6, 8, 6, 9, 6, 8, 4, 8, 8. **Answer:** 8 occurs four times, 6 occurs three times; mode = **8**.

Problem 5. Calculate the median of the following frequency distribution: Class 10–20, 20–30, 30–40, 40–50, 50–60 with $f = 4, 10, 16, 12, 8$.

Solution: c.f. = 4, 14, 30, 42, 50. $N = 50$; $N/2 = 25$ lies in the **30–40** class. $L = 30$, c.f. = 14, $f = 16$, $h = 10$. Median = $30 + [(25 - 14)/16] \times 10 = 30 + 6.875 = 36.875$.

Problem 6. For the same data find the mean by the direct method. Mid-values: 15, 25, 35, 45, 55; $fm = 60, 250, 560, 540, 440$; $\Sigma fm = 1850$. Mean = $1850/50 = 37$. Mode: modal class 30–40, $f_1 = 16$, $f_0 = 10$, $f_2 = 12$; $\Delta_1 = 6$, $\Delta_2 = 4$; Mode = $30 + (6/10) \times 10 = 36$.

Unit 3: Quick Revision Questions

Q1. What are the requisites of a good average?

It should be rigidly defined, simple to calculate, based on all observations, not unduly affected by extreme values, stable under sampling and capable of algebraic treatment.

Q2. Why is the median preferred over the mean in the case of income data?

Income distributions are skewed by a few very rich persons. The mean is pulled upward by these extreme values, whereas the median is not, so it better represents the typical income.

Q3. State the relationship between mean, median and mode.

For moderately asymmetrical data, Mode = 3 Median – 2 Mean. In a symmetrical distribution, mean = median = mode.

UNIT 4: Measures of Dispersion

Teaching hours: 12 | Topics: 4.1 Meaning and concepts of dispersion | 4.2 Quartile deviation and its coefficient | 4.3 Mean deviation and its coefficient | 4.4 Standard deviation and its coefficient | Skewness, numerical problems

4.1 Meaning and Concepts of Dispersion

An average gives a single representative value, but it does not tell us **how the items are spread** around it. Consider two firms whose workers have the same average wage of Rs. 500. In Firm A, every worker earns between Rs. 480 and Rs. 520; in Firm B, wages range from Rs. 100 to Rs. 900. The averages are equal, but the situations are very different. The extent to which items differ from the central value is called **dispersion** (variation, spread or scatter).

Definitions

Spiegel: 'Dispersion is the degree to which numerical data tend to spread about an average value.'

Simpson and Kafka: 'The measurement of the scatter of the mass of figures in a series about an average is called measure of variation or dispersion.'

Objectives of measuring dispersion

- To test the **reliability of an average**: small dispersion means the average is representative.
- To **compare the variability** (consistency) of two or more series.
- To control variability, e.g., in **quality control** of products.
- To serve as the **basis for other statistical techniques** like correlation, regression and testing.

Properties of a good measure of dispersion

It should be simple to understand and calculate, rigidly defined, based on all observations, not unduly affected by extreme values, capable of algebraic treatment and stable under sampling.

Types of dispersion measures

Absolute measures (with units)	Relative measures (coefficients, pure numbers)
Range	Coefficient of range
Quartile deviation (QD)	Coefficient of QD
Mean deviation (MD)	Coefficient of MD
Standard deviation (SD)	Coefficient of SD and coefficient of variation

Absolute measures are expressed in the units of the data (rupees, kg) and cannot be used to compare series with different units. **Relative measures** are ratios or percentages, free from units, and are used for comparison.

Range (for knowledge)

Range = Largest value (L) – Smallest value (S). Coefficient of range = $(L - S) / (L + S)$. It is the simplest measure but is based on only two items.

4.2 Quartile Deviation (Semi-Interquartile Range)

Meaning and Definition

Quartiles divide an arranged series into four equal parts: Q_1 (lower quartile, 25% items below), Q_2 (median) and Q_3 (upper quartile, 75% items below). **Quartile deviation** is half of the difference between the upper and lower quartiles, and hence is called the **semi-interquartile range**.

Formulae

Quartile Deviation (QD) = $(Q_3 - Q_1) / 2$

Coefficient of QD = $(Q_3 - Q_1) / (Q_3 + Q_1)$

Individual / discrete series: Q_1 = size of $((N + 1)/4)$ th item; Q_3 = size of $(3(N + 1)/4)$ th item

Continuous series: $Q_1 = L_1 + [(N/4 - c.f.) / f] \times h$; $Q_3 = L_3 + [(3N/4 - c.f.) / f] \times h$

Example 1 (Individual series). Find QD and its coefficient for 5, 7, 9, 11, 13, 15, 17, 19, 21.

$N = 9$ (already arranged). $Q_1 = (9 + 1)/4 = 2.5$ th item = $7 + 0.5 \times (9 - 7) = 8$. $Q_3 = 3(10)/4 = 7.5$ th item = $17 + 0.5 \times (19 - 17) = 18$.

$QD = (18 - 8)/2 = 5$. Coefficient of QD = $(18 - 8)/(18 + 8) = 10/26 = 0.385$.

Example 2 (Continuous series). Find QD for the marks of 50 students. c.f. = 5, 13, 28, 44, 50.

Q_1 : $N/4 = 12.5$ lies in the 10–20 class (c.f. 13). $L = 10$, c.f. before = 5, $f = 8$, $h = 10$. $Q_1 = 10 + [(12.5 - 5)/8] \times 10 = 10 + 9.375 = 19.375$.

Q_3 : $3N/4 = 37.5$ lies in the 30–40 class (c.f. 44). $L = 30$, c.f. before = 28, $f = 16$, $h = 10$. $Q_3 = 30 + [(37.5 - 28)/16] \times 10 = 30 + 5.9375 = 35.9375$.

$QD = (35.9375 - 19.375)/2 = 16.5625/2 = 8.28$ marks. Coefficient of QD = $16.5625 / 55.3125 = 0.30$.

Merits and demerits of QD

- **Merits:** simple to understand and calculate; not affected by extreme values because it uses the middle 50% of the data; can be used for open-end classes; not affected by extreme items.
- **Demerits:** ignores the first 25% and last 25% of the items; not based on all items; not capable of further algebraic treatment; affected by sampling fluctuations.

4.3 Mean Deviation (Average Deviation)

Meaning and Definition

Mean deviation is the arithmetic average of the **absolute deviations** (ignoring plus and minus signs) of the items from an average, usually the mean or the median (the mean deviation is minimum when taken about the median). It is denoted by MD or δ (delta). **Clark and Schkade:** 'Average deviation is the arithmetic mean of the absolute deviations of the items from some measure of central tendency.'

Formulae

Individual series: $MD = \sum |X - A| / N$ (A = mean, median or mode)

Discrete series: $MD = \sum f |X - A| / N$

Continuous series: $MD = \sum f |m - A| / N$ (m = mid-value)

Coefficient of MD = MD / A (A = the average about which MD is calculated; MD about mean divided by mean)

Example 3 (Individual series). Find the mean deviation about the mean and its coefficient for 4, 6, 10, 12, 18.

X	$ X - \bar{X} (\bar{X} = 10)$
4	6
6	4
10	0
12	2
18	8
$\Sigma X = 50$	$\Sigma d = 20$

Mean = $50/5 = 10$. MD = $20/5 = 4$. Coefficient of MD = $4/10 = 0.4$.

Example 4 (Continuous series). Mean deviation about the mean for the marks of 50 students (mean = 27).

Marks	m	f	$ m - 27 $	$f m - 27 $
0–10	5	5	22	110
10–20	15	8	12	96
20–30	25	15	2	30
30–40	35	16	8	128
40–50	45	6	18	108
Total		50		472

MD = $472/50 = 9.44$ marks. Coefficient of MD = $9.44/27 = 0.35$.

Merits and demerits of MD

- **Merits:** based on all observations; simple to understand; less affected by extreme values than SD; useful in economic and social studies.
- **Demerits:** ignores algebraic signs (mathematically unsound); not capable of further algebraic treatment; not suitable for open-end classes; less accurate and rarely used in advanced statistics.

4.4 Standard Deviation (SD)

Meaning and Definition

Standard deviation is the **positive square root of the arithmetic mean of the squared deviations** of the items from their arithmetic mean. It was introduced by **Karl Pearson** (1893) and is denoted by σ (sigma). It is also called the **root mean square deviation**. **Variance** = σ^2 (the square of SD).

Formulae

Individual series (actual mean): $\sigma = \sqrt{\Sigma(X - \bar{X})^2 / N}$

Individual series (direct): $\sigma = \sqrt{\Sigma X^2 / N - (\Sigma X / N)^2}$

Continuous series (actual mean): $\sigma = \sqrt{\Sigma f(m - \bar{X})^2 / N}$

Continuous series (step deviation): $\sigma = h \times \sqrt{\Sigma fd'^2 / N - (\Sigma fd' / N)^2}$, where $d' = (m - A)/h$

Coefficient of SD = $\sigma / \bar{X}$

Coefficient of variation (CV) = $(\sigma / \bar{X}) \times 100$ (a percentage; used to compare consistency)

Example 5 (Individual series). Find SD and CV of 2, 4, 6, 8, 10.

X	d = X - 6	d ²
2	-4	16
4	-2	4
6	0	0
8	2	4
10	4	16
Total	0	40

$$\sigma = \sqrt{40/5} = \sqrt{8} = 2.83. \quad CV = (2.83/6) \times 100 = 47.1\%.$$

Example 6 (Continuous series). Standard deviation of the marks of 50 students. Mean = 27, take A = 25 and h = 10 for the step-deviation method.

Marks	m	f	d' = (m - 25)/10	fd'	fd' ²
0-10	5	5	-2	-10	20
10-20	15	8	-1	-8	8
20-30	25	15	0	0	0
30-40	35	16	1	16	16
40-50	45	6	2	12	24
Total		50		10	68

$$\sigma = 10 \times \sqrt{[68/50 - (10/50)^2]} = 10 \times \sqrt{1.36 - 0.04} = 10 \times \sqrt{1.32} = 10 \times 1.149 = 11.49 \text{ marks.}$$

$$\text{Variance} = \sigma^2 = 132. \quad \text{Coefficient of SD} = 11.49/27 = 0.4255. \quad CV = 42.55\%.$$

Check by actual mean method: $\Sigma f(m - 27)^2 = 5(484) + 8(144) + 15(4) + 16(64) + 6(324) = 2420 + 1152 + 60 + 1024 + 1944 = 6600$; $\sigma^2 = 6600/50 = 132$; $\sigma = 11.49$. The two methods agree.

Comparison of consistency using CV

Example 7. Firm A: mean wage Rs. 50, σ = Rs. 10. Firm B: mean wage Rs. 80, σ = Rs. 12. Which firm has more consistent (less variable) wages?

CV of A = $(10/50) \times 100 = 20\%$; CV of B = $(12/80) \times 100 = 15\%$. Since B has the smaller CV, **Firm B is more consistent**, even though its absolute SD is larger. **The series with the lower CV is more consistent (more stable); the series with the higher CV is more variable.**

Merits and demerits of SD

- **Merits:** based on all observations; rigidly defined; capable of **further algebraic treatment** (combined SD, correlation, regression, tests of significance); least affected by sampling fluctuations; the best and most widely used measure of dispersion.
- **Demerits:** difficult to understand and calculate; gives more weight to extreme items because deviations are squared; cannot be calculated for open-end classes without assumption.

Properties of SD

- SD is **independent of change of origin** (adding or subtracting a constant does not change it) but **depends on change of scale** (multiplying by k multiplies SD by |k|).
- SD is never negative. If all items are equal, SD = 0.
- In a normal distribution, mean $\pm 1\sigma$ covers about 68%, mean $\pm 2\sigma$ about 95% and mean $\pm 3\sigma$ about 99.7% of the items.

Measure	Based on all items	Affected by extremes	Further algebraic treatment	Best for
Range	No	Highly	No	Quick idea

Measure	Based on all items	Affected by extremes	Further algebraic treatment	Best for
Quartile deviation	No	No	No	Open-end data, skewed data
Mean deviation	Yes	Slightly	No	Simple analysis
Standard deviation	Yes	Yes	Yes	Scientific analysis

Skewness

Meaning and Definition

Skewness means **lack of symmetry**. Dispersion tells us how far items are spread, whereas skewness tells us the **direction and degree of asymmetry** of the distribution. **Morris Hamburg**: 'Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution.'

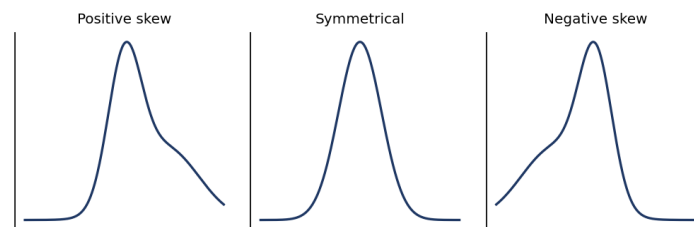


Figure 4.1 Types of skewness

Type	Shape	Relationship of averages
Symmetrical	Bell-shaped, both sides equal	Mean = Median = Mode; skewness = 0
Positively skewed	Long tail to the right	Mean > Median > Mode; skewness > 0
Negatively skewed	Long tail to the left	Mean < Median < Mode; skewness < 0

Measures of skewness

Karl Pearson's coefficient of skewness

$$Sk(P) = (\text{Mean} - \text{Mode}) / \sigma$$

If the mode is ill-defined: $Sk(P) = 3(\text{Mean} - \text{Median}) / \sigma$

Its value usually lies between -3 and +3.

Bowley's coefficient of skewness (based on quartiles)

$$Sk(B) = (Q_3 + Q_1 - 2 \text{ Median}) / (Q_3 - Q_1)$$

Its value lies between -1 and +1; useful for open-end classes.

Example 8. For the marks of 50 students: Mean = 27, Median = 28, Mode = 30.91, $\sigma = 11.49$, $Q_1 = 19.375$, $Q_3 = 35.9375$.

Karl Pearson: $Sk = (27 - 30.91)/11.49 = -0.34$. Using the median: $Sk = 3(27 - 28)/11.49 = -0.26$.

Bowley: $Sk = (35.9375 + 19.375 - 2 \times 28)/(35.9375 - 19.375) = (-0.6875)/16.5625 = -0.04$.

Interpretation: all three coefficients are negative, so the distribution is **slightly negatively skewed** (the tail extends to the lower marks); the Bowley measure shows the skewness is very small and the data are nearly symmetrical in the middle portion.

Example 9. If the mean = 50, mode = 45 and $\sigma = 10$ for a distribution, then $Sk = (50 - 45)/10 = +0.5$, i.e., positively skewed.

Additional Numerical Problems (with Answers)

Problem 1. Find the range, QD and coefficient of QD of the series 10, 20, 30, 40, 50, 60, 70. **Answer:** Range = 60; Q_1 = 2nd item = 20, Q_3 = 6th item = 60; QD = 20; coefficient = $40/80 = 0.5$.

Problem 2. Find SD of 3, 5, 7, 9, 11. **Answer:** mean = 7; $d^2 = 16, 4, 0, 4, 16 = 40$; $\sigma = \sqrt{8} = 2.83$.

Problem 3. Find the mean deviation about the mean of 2, 4, 6, 8, 10. **Answer:** mean = 6; $\sum |d| = 4 + 2 + 0 + 2 + 4 = 12$; MD = 2.4; coefficient = $2.4/6 = 0.4$.

Problem 4. The mean of a series is 40 and the CV is 25%. Find SD. **Answer:** $\sigma = CV \times \text{mean} / 100 = 25 \times 40/100 = 10$.

Problem 5. Series X: mean 100, σ 15; Series Y: mean 60, σ 12. Which is more variable? **Answer:** $CV(X) = 15\%$; $CV(Y) = 20\%$; **Y is more variable.**

Unit 4: Quick Revision Questions

Q1. Why is standard deviation considered the best measure of dispersion?

Because it is based on all observations, rigidly defined, amenable to algebraic treatment and least affected by sampling fluctuations; it is the foundation for correlation, regression and tests of significance.

Q2. Distinguish between absolute and relative measures of dispersion.

Absolute measures are expressed in the original units of data; relative measures are pure numbers (ratios or percentages) used to compare variability of series with different units or averages.

Q3. What is the use of the coefficient of variation?

It measures relative variability as a percentage and is used to compare the consistency of two or more series; the series with the smaller CV is more consistent.

UNIT 5: Correlation Analysis

Teaching hours: 12 | Topics: 5.1 Meaning, definition and types of correlation; variables and types of variables | 5.2 Karl Pearson's coefficient of correlation | 5.3 Probable error | 5.4 Numerical problems

5.1 Meaning, Definition and Types of Correlation

So far we studied **one variable** at a time (univariate analysis). In business, however, variables are often related: sales depend on advertising, demand depends on price, and profit depends on output. The study of such **two-variable (bivariate)** relationships is called **correlation analysis**.

Definitions

Croxtan and Cowden: 'When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.'

A. M. Tuttle: 'Correlation is an analysis of the covariation between two or more variables.'

Simply: correlation is a statistical technique which measures the **direction and degree of the linear relationship** between two variables.

Variables and types of variables

A **variable** is a characteristic or quantity that can take different values (price, income, age, marks). Types of variables are:

Basis	Types	Example
Nature of values	Discrete (counted, whole numbers) / Continuous (measured, any value in a range)	Number of employees / weight, time

Basis	Types	Example
Measurement	Quantitative (numerical) / Qualitative (attributes)	Sales in Rs. / gender, brand
Role in a relationship	Independent (cause, explanatory) / Dependent (effect, response)	Advertising / sales
Time dimension	Constant (fixed value) / Random (value depends on chance)	Number of days in a week / daily rainfall

Types of correlation

1. Positive and negative correlation (based on direction)

- **Positive correlation:** the two variables move in the **same direction** (increase together or decrease together). Example: advertising expenditure and sales; height and weight.
- **Negative (inverse) correlation:** the variables move in **opposite directions**. Example: price and demand; temperature and sale of woollens.
- **Zero correlation (no correlation):** the variables are unrelated, $r = 0$. Example: shoe size and intelligence.

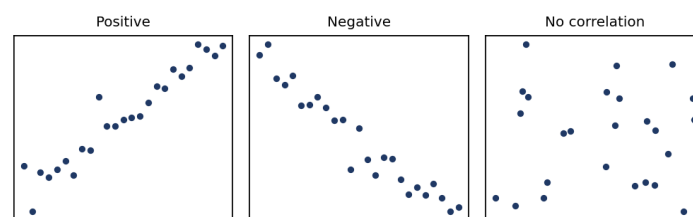


Figure 5.1 Scatter diagrams showing types of correlation

2. Linear and non-linear correlation (based on the ratio of change)

- **Linear correlation:** the change in one variable bears a **constant ratio** to the change in the other; the scatter points lie close to a straight line.
- **Non-linear (curvilinear) correlation:** the ratio of change is not constant; the points form a curve (e.g., rainfall and crop yield, where too much rain reduces yield).

3. Simple, multiple and partial correlation (based on number of variables)

- **Simple correlation:** study of two variables only.
- **Multiple correlation:** study of one variable against two or more others together (e.g., yield versus rainfall and fertiliser).
- **Partial correlation:** relationship between two variables while keeping the others constant.

Methods of studying correlation

- **Scatter diagram** (graphic): plot the pairs of values as dots; the pattern shows the direction and strength.
- **Karl Pearson's coefficient of correlation** (algebraic; most widely used).
- **Spearman's rank correlation** and **concurrent deviation method** (for knowledge).

Degree of correlation

Value of r	Interpretation
+1	Perfect positive correlation
+0.75 to +0.99	High (strong) positive correlation
+0.25 to +0.75	Moderate positive correlation
0 to +0.25	Low (weak) positive / negligible correlation
0	No correlation
-0.25 to 0	Low negative correlation
-0.75 to -0.25	Moderate negative correlation
-0.99 to -0.75	High negative correlation

Value of r	Interpretation
-1	Perfect negative correlation

Important cautions

Correlation does not mean causation. A high r between two variables does not prove that one causes the other: both may be influenced by a third factor, or the relationship may be a chance coincidence (spurious correlation).

Pearson's r measures only **linear** relationship; r = 0 does not mean no relationship of any kind, e.g., a curved relation.

5.2 Karl Pearson's Coefficient of Correlation

This is the most popular and mathematically exact method, devised by **Karl Pearson** (1867–1936). It is also called the **product-moment correlation coefficient** and is denoted by r. It measures the degree of linear relationship between two variables X and Y.

Formulae for r

Basic definition: $r = \text{Cov}(X, Y) / (\sigma_x \cdot \sigma_y)$, $\text{Cov}(X, Y) = \Sigma(X - \bar{X})(Y - \bar{Y}) / N$

Deviations from actual means: $r = \Sigma xy / \sqrt{(\Sigma x^2 \cdot \Sigma y^2)}$, where $x = X - \bar{X}$ and $y = Y - \bar{Y}$

Direct method (raw values): $r = [N\Sigma XY - (\Sigma X)(\Sigma Y)] / \sqrt{[N\Sigma X^2 - (\Sigma X)^2] \cdot [N\Sigma Y^2 - (\Sigma Y)^2]}$

Assumed mean method: $r = [N\Sigma dxdy - (\Sigma dx)(\Sigma dy)] / \sqrt{[N\Sigma dx^2 - (\Sigma dx)^2] \cdot [N\Sigma dy^2 - (\Sigma dy)^2]}$

Properties of r

- r lies between **-1 and +1**: $-1 \leq r \leq +1$.
- r is **independent of change of origin and scale** (adding a constant or multiplying by a positive constant does not change r).
- r is a **pure number**, free of units, so series with different units can be correlated.
- r is the **geometric mean of the two regression coefficients**: $r = \sqrt{b_{xy} \cdot b_{yx}}$.
- If X and Y are independent, $r = 0$ (but $r = 0$ does not always mean independence).
- r^2 is called the **coefficient of determination**; it shows the percentage of the variation in Y explained by X.

Steps for calculating r (deviation method)

1. Calculate the means $\bar{X}$ and $\bar{Y}$.
2. Find deviations $x = X - \bar{X}$ and $y = Y - \bar{Y}$.
3. Compute x^2 , y^2 and the product xy ; find Σx^2 , Σy^2 , Σxy .
4. Substitute in $r = \Sigma xy / \sqrt{(\Sigma x^2 \cdot \Sigma y^2)}$ and interpret.

Example 1 (Positive correlation). Find the correlation coefficient between advertising expenditure X (Rs. thousand) and sales Y (Rs. lakh).

X	Y	$x = X - 14$	$y = Y - 21$	x^2	y^2	xy
10	15	-4	-6	16	36	24
12	18	-2	-3	4	9	6
14	20	0	-1	0	1	0
16	24	2	3	4	9	6
18	28	4	7	16	49	28
Total 70	105	0	0	40	104	64

$\bar{X} = 70/5 = 14$; $\bar{Y} = 105/5 = 21$. $r = \Sigma xy / \sqrt{(\Sigma x^2 \cdot \Sigma y^2)} = 64 / \sqrt{(40 \times 104)} = 64 / \sqrt{4160} = 64 / 64.50 = +0.992$.

Interpretation: there is a very high positive correlation; higher advertising is associated with higher sales.

Example 2 (Negative correlation). Price X (Rs.) and quantity demanded Y (units) of a commodity.

X	Y	$x = X - 30$	$y = Y - 58$	x^2	y^2	xy
10	90	-20	32	400	1024	-640
20	70	-10	12	100	144	-120
30	60	0	2	0	4	0
40	40	10	-18	100	324	-180
50	30	20	-28	400	784	-560
Total 150	290	0	0	1000	2280	-1500

$r = -1500 / \sqrt{(1000 \times 2280)} = -1500 / 1509.97 = -0.993$. Very high negative correlation: as the price rises, demand falls (the law of demand).

Example 3 (Direct method). Find r for X = 2, 4, 5, 6, 8, 11 and Y = 18, 12, 10, 8, 7, 5.

X	Y	X^2	Y^2	XY
2	18	4	324	36
4	12	16	144	48
5	10	25	100	50
6	8	36	64	48
8	7	64	49	56
11	5	121	25	55
$\Sigma X = 36$	$\Sigma Y = 60$	$\Sigma X^2 = 266$	$\Sigma Y^2 = 706$	$\Sigma XY = 293$

$N = 6$. Numerator = $N\Sigma XY - \Sigma X\Sigma Y = 6(293) - (36)(60) = 1758 - 2160 = -402$. Denominator = $\sqrt{[(6 \times 266 - 36^2)(6 \times 706 - 60^2)]} = \sqrt{[(1596 - 1296)(4236 - 3600)]} = \sqrt{(300 \times 636)} = \sqrt{190800} = 436.8$.

$r = -402 / 436.8 = -0.920$: a high negative correlation.

Merits and demerits of Karl Pearson's method

- **Merits:** gives both the **degree and the direction** of correlation; based on all items; algebraically exact and widely accepted; forms the basis for regression analysis.
- **Demerits:** assumes a **linear** relationship; affected by extreme values; calculation is lengthy; affected by sampling errors; the interpretation of r can be misleading without knowing the probable error.

5.3 Probable Error (P.E.)

Meaning

The **probable error** of the correlation coefficient is a statistical measure that helps to **test the reliability and significance** of the calculated value of r from a sample. It indicates the limits within which the true population correlation coefficient is expected to lie with a 50% chance (probability). Its use presumes that the sample is **random** and the population is approximately **normal**.

Formula and limits

P.E. = $0.6745 \times (1 - r^2) / \sqrt{N}$ (N = number of pairs)

Limits of the population correlation: $r \pm \text{P.E.}$

Standard error (S.E.) = $(1 - r^2) / \sqrt{N}$, and **P.E. = $0.6745 \times \text{S.E.}$**

Interpretation rules

1. If $r < \text{P.E.}$, there is **no evidence of correlation** (r is not significant).
2. If $r > 6 \times \text{P.E.}$, the correlation is **significant** (definitely exists).
3. If r lies between P.E. and $6 \times \text{P.E.}$, nothing definite can be said.
4. The condition N is **fairly large** (at least about 10 pairs) should be satisfied for the test to be reliable.

Example 4. Find the probable error for Example 1 ($r = 0.992$, $N = 5$).

$r^2 = 0.984$; $1 - r^2 = 0.0154$ (approximately). $\sqrt{N} = \sqrt{5} = 2.236$. **P.E. = $0.6745 \times 0.0154 / 2.236 = 0.0046$** . $6 \times \text{P.E.} = 0.028$. Since $r = 0.992$ is much greater than 0.028, the correlation is **significant**. The limits of the population correlation are 0.992 ± 0.0046 , i.e., from 0.987 to 0.997.

Example 5. If $r = 0.8$ and $N = 25$, find P.E. and the limits. **P.E. = $0.6745 \times (1 - 0.64) / \sqrt{25} = 0.6745 \times 0.36 / 5 = 0.0486$** . Limits: $0.8 \pm 0.0486 = 0.751$ to 0.849 . Also, $6 \times \text{P.E.} = 0.29 < 0.8$, so r is significant.

Example 6. If $r = 0.3$ and $N = 16$, **P.E. = $0.6745 \times 0.91 / 4 = 0.1534$** ; $6 \times \text{P.E.} = 0.92$. Since r (0.3) is greater than P.E. but less than $6 \times \text{P.E.}$, the value is **not definitely significant**.

5.4 Additional Numerical Problems (with Answers)

Problem 1. Calculate r from: $N = 10$, $\Sigma X = 100$, $\Sigma Y = 150$, $\Sigma X^2 = 1100$, $\Sigma Y^2 = 2500$, $\Sigma XY = 1550$.

Solution: Numerator = $N\Sigma XY - \Sigma X\Sigma Y = 10(1550) - 100 \times 150 = 15500 - 15000 = 500$. Denominator = $\sqrt{[(10 \times 1100 - 100^2)(10 \times 2500 - 150^2)]} = \sqrt{[(11000 - 10000)(25000 - 22500)]} = \sqrt{(1000 \times 2500)} = 1581.1$. Therefore $r = 500 / 1581.1 = +0.316$ (moderate positive correlation).

Problem 2. If $\Sigma xy = 120$, $\Sigma x^2 = 100$ and $\Sigma y^2 = 225$, find r. **Answer:** $r = 120 / \sqrt{(100 \times 225)} = 120 / 150 = +0.8$.

Problem 3. If $\text{Cov}(X, Y) = 12$, $\sigma_x = 4$ and $\sigma_y = 5$, find r. **Answer:** $r = 12 / (4 \times 5) = 0.6$.

Problem 4. Find r between X: 1, 2, 3, 4, 5 and Y: 2, 4, 6, 8, 10. **Answer:** $Y = 2X$ exactly, so $r = +1$ (perfect positive correlation). Check: $x = -2, -1, 0, 1, 2$; $y = -4, -2, 0, 2, 4$; $\Sigma xy = 20$, $\Sigma x^2 = 10$, $\Sigma y^2 = 40$; $r = 20 / \sqrt{400} = 1$.

Problem 5. Find the probable error if $r = 0.6$ and $N = 36$. **Answer:** **P.E. = $0.6745 \times (1 - 0.36) / 6 = 0.6745 \times 0.64 / 6 = 0.072$** . Since $6 \times \text{P.E.} = 0.432 < 0.6$, r is significant.

Unit 5: Quick Revision Questions

Q1. What is correlation? State its types.

Correlation is the statistical measure of the direction and degree of linear relationship between two variables. Types: positive and negative, linear and non-linear, simple, multiple and partial.

Q2. State the properties of Karl Pearson's coefficient of correlation.

It lies between -1 and $+1$, is independent of change of origin and scale, is a pure number, is the geometric mean of the regression coefficients, and its square is the coefficient of determination.

Q3. What is probable error and what is its use?

Probable error is $0.6745 \times (1 - r^2) / \sqrt{N}$. It is used to test the reliability of r: if $r > 6 \times \text{P.E.}$, correlation is significant; if $r < \text{P.E.}$, it is insignificant.

Formula Sheet and Exam Preparation

A. One-page Formula Sheet

Topic	Formula
Class mid-value	$m = (\text{Lower limit} + \text{Upper limit}) / 2$
Pie diagram angle	$(\text{Value} / \text{Total}) \times 360^\circ$
AM (individual / discrete / continuous)	$\Sigma X/N; \Sigma fx/N; \Sigma fm/N$
AM (step deviation)	$A + (\Sigma fd'/N) \times h$
Combined mean	$(N_1\bar{X}_1 + N_2\bar{X}_2)/(N_1 + N_2)$
Median (individual)	size of $((N + 1)/2)$ th item
Median (continuous)	$L + [(N/2 - c.f.)/f] \times h$
Mode (continuous)	$L + [\Delta_1/(\Delta_1 + \Delta_2)] \times h, \Delta_1 = f_1 - f_0, \Delta_2 = f_1 - f_2$
Empirical relation	$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$
Quartile deviation, coefficient	$(Q_3 - Q_1)/2; (Q_3 - Q_1)/(Q_3 + Q_1)$
Mean deviation, coefficient	$\Sigma f X - A /N; MD / A$
Standard deviation	$\sqrt{\Sigma f(m - \bar{X})^2/N}$ or $h\sqrt{\Sigma fd'^2/N - (\Sigma fd'/N)^2}$
Coefficient of SD, CV	$\sigma/\bar{X}; (\sigma/\bar{X}) \times 100$
Karl Pearson's skewness	$(\text{Mean} - \text{Mode})/\sigma$ or $3(\text{Mean} - \text{Median})/\sigma$
Bowley's skewness	$(Q_3 + Q_1 - 2 \text{ Median})/(Q_3 - Q_1)$
Karl Pearson's r	$\Sigma xy/\sqrt{\Sigma x^2 \Sigma y^2}$ or $[N\Sigma XY - \Sigma X\Sigma Y]/\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}$
Probable error	$0.6745 (1 - r^2)/\sqrt{N}$; significant if $r > 6 \text{ P.E.}$

B. Semester-End Examination Pattern (60 Marks)

- Total **4 questions** of **15 marks** each = 60 marks.
- **Q.1 is compulsory** and has sub-questions covering **all modules** of the syllabus.
- Attempt any **3 questions from Q.2 to Q.6** covering all modules.
- **Q.6** has four sub-questions, of which **any two** are to be solved.
- Continuous assessment (CCA) is for **40 marks**: two tests of 10 marks each, assignment/oral/presentation of 12 marks, and attendance up to 8 marks (4 marks for 60–75%, 6 marks for 75–85%, 8 marks for 85% and above). Total = 100 marks.

Preparation tip

Numerical problems on **mean, median, mode, quartile deviation, mean deviation, standard deviation and Karl Pearson's correlation** are the most scoring. Practise at least two problems of each type, show all steps, write the formula first and give the final answer with units and a one-line interpretation.

C. Important Theory Questions

Long answer (10–15 marks)

1. Define statistics. Explain its nature, scope and limitations.
2. Explain the importance of Business Statistics with suitable examples.
3. Define primary and secondary data. Explain the sources and methods of collecting primary data.
4. What is a diagram? Explain the objects, importance and rules of diagrammatic presentation.

5. Explain the types of diagrams. Discuss their limitations.
6. What is a graph? Explain the types and uses of graphs. Distinguish between diagrams and graphs.
7. Define an average. Discuss the requisites of a good average. State the merits and demerits of mean, median and mode.
8. What is dispersion? Explain its objects and the absolute and relative measures of dispersion.
9. Explain quartile deviation, mean deviation and standard deviation, with their merits and demerits.
10. What is skewness? Explain the types and measures of skewness.
11. Define correlation. Explain its types. Explain Karl Pearson's coefficient of correlation and its properties.
12. What is probable error? Explain its calculation and use in interpreting the correlation coefficient.

Short notes (5 marks)

Origin and history of statistics; Descriptive and inferential statistics; Questionnaire and schedule; Interview method; Pie diagram; Histogram; Ogive; Combined mean; Positional averages; Coefficient of variation; Karl Pearson's coefficient of skewness; Scatter diagram; Coefficient of determination; Correlation and causation.

Important numerical problems

1. Mean, median and mode of a continuous series (use the 50-student data in this booklet to practise).
2. Combined mean and corrected mean; finding a missing frequency.
3. Quartile deviation and its coefficient for individual and continuous series.
4. Mean deviation and its coefficient about the mean and median.
5. Standard deviation, variance, coefficient of variation; comparing the consistency of two series.
6. Karl Pearson's and Bowley's coefficient of skewness.
7. Karl Pearson's coefficient of correlation and probable error.

D. Multiple Choice Questions (MCQs)

Choose the correct option. The answers are given at the end of the list.

1. The word 'statistics' is derived from the Latin word: (a) Statista (b) Status (c) Statistik (d) Statistique
2. Who defined statistics as 'the science of collection, presentation, analysis and interpretation of numerical data'? (a) Bowley (b) Secrist (c) Croxton and Cowden (d) Wallis and Roberts
3. Data collected for the first time by the investigator are called: (a) secondary (b) primary (c) tertiary (d) grouped
4. A questionnaire is filled in by: (a) the enumerator (b) the respondent (c) the supervisor (d) the statistician
5. The best diagram to show the break-up of a total into parts using angles is: (a) simple bar (b) histogram (c) pie diagram (d) pictogram
6. Frequency polygon is obtained by joining: (a) upper limits (b) mid-points of the tops of histogram bars (c) lower limits (d) class widths
7. The median can be located graphically from the: (a) histogram (b) ogive (c) pie chart (d) pictogram
8. The sum of deviations of items from the arithmetic mean is always: (a) maximum (b) one (c) zero (d) negative
9. The average that is most affected by extreme values is: (a) median (b) mode (c) mean (d) quartile
10. In a symmetrical distribution: (a) mean > median (b) mean = median = mode (c) mean < mode (d) mean > mode
11. The mode of 2, 3, 3, 4, 5, 5, 5, 6 is: (a) 3 (b) 4 (c) 5 (d) 6
12. Quartile deviation is equal to: (a) $Q_3 - Q_1$ (b) $(Q_3 + Q_1)/2$ (c) $(Q_3 - Q_1)/2$ (d) Q_3/Q_1
13. Standard deviation was introduced by: (a) Fisher (b) Karl Pearson (c) Bowley (d) Gauss
14. The square of standard deviation is called: (a) range (b) variance (c) coefficient of variation (d) mean deviation
15. The series with a smaller coefficient of variation is: (a) more variable (b) more consistent (c) skewed (d) symmetric
16. If mean = 50, mode = 40 and $\sigma = 20$, Karl Pearson's coefficient of skewness is: (a) 0.5 (b) 2 (c) -0.5 (d) 0
17. The value of Karl Pearson's coefficient of correlation lies between: (a) 0 and 1 (b) -1 and +1 (c) $-\infty$ and $+\infty$ (d) 0 and ∞

18. Price and demand of a normal commodity generally show: (a) positive correlation (b) negative correlation (c) zero correlation (d) perfect correlation
19. The probable error of $r = 0.6$ with $N = 36$ is approximately: (a) 0.072 (b) 0.72 (c) 0.0072 (d) 0.6
20. The correlation is significant if: (a) $r < P.E.$ (b) $r = 0$ (c) $r > 6 P.E.$ (d) $r = P.E.$

Answers: 1-b, 2-c, 3-b, 4-b, 5-c, 6-b, 7-b, 8-c, 9-c, 10-b, 11-c, 12-c, 13-b, 14-b, 15-b, 16-a, 17-b, 18-b, 19-a, 20-c.

(Check for Q16: $S_k = (50 - 40)/20 = 0.5$.)

E. Reference Books (as per syllabus)

1. S. C. Gupta, **Fundamentals of Statistics**, 10th Edition, Himalaya Publishing House, New Delhi, 2022.
2. Dr. B. N. Gupta, **Business Statistics**, 1st Edition, Sahitya Bhavan Publishing House, Agra, 2022.
3. S. C. Gupta and Indra Gupta, **Business Statistics**, Himalaya Publishing House, 2nd Edition, 2013.
4. S. M. Shukla and S. P. Sahai, **Business Statistics**, 1st Edition, Sahitya Bhavan Publishing House, Agra, 2023.
5. Dr. Maroti Kachave and Dr. Ashok Kalam, **Business Statistics**, 1st Edition, Suman Publication, Latur, 2021.
6. Dr. Maroti Kachave, **Business Statistics**, Kailas Publication, Chh. Sambhajinagar, Edition I, 2024.
7. Dr. C. K. Harnawale, Dr. R. B. Rampure and Dr. Rajesh Goje, **Business Statistics and Mathematics**, NEWMAN Publication.
8. Dr. B. G. Khatal, **Sankhyiki Paddhati** (Marathi), Prashant Publication, Jalgaon.

— End of Notes —