Statistics for Business Unit 1 V3.docx

Statistics for Business_BBA_2

Statistics for Business_BBA_2

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:127433163

Submission Date

Feb 3, 2026, 1:16 PM GMT+5:30

Download Date

Feb 3, 2026, 1:18 PM GMT+5:30

File Name

Statistics for Business Unit 1 V3.docx

File Size

195.5 KB

29 Pages

5,841 Words

34,170 Characters

0% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 20 words)

Match Groups

0 Not Cited or Quoted 0%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

0% Internet sources
0% Publications
0% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **0** Not Cited or Quoted 0%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

0% Internet sources
0% Publications
0% Submitted works (Student Papers)

Unit 1: Introduction to Data

Learning Objectives

1. Understand the importance and methods of data organisation, including the need for classifying, tabulating, and presenting data in a structured format.
2. Differentiate between the major types of data, such as qualitative vs. quantitative data, and primary vs. secondary data.
3. Identify and apply the four levels of measurement scales—nominal, ordinal, interval, and ratio—used in statistical analysis.
4. Construct and interpret various statistical series, including individual, discrete, and continuous series, for effective data representation.
5. Summarise data meaningfully using appropriate tools and techniques that highlight the central tendencies, dispersion, and distribution patterns.
6. Familiarise with key statistical terminology to enhance comprehension and communication of statistical findings.
7. Apply knowledge through descriptive questions and a real-life case study, reinforcing practical understanding and analytical skills.

Content

1.0 Introductory Caselet

"Anita's Academic Analytics: Getting Smart About Your Data"

Anita, data coordinator at a senior secondary school in Pune, was getting lost with the piles of academic statistics that she had to deal with every term. From tests scores and attendance lists to student feedback and records of extracurriculars, each department would send her spreadsheets of messy data.

The principal had put her in charge of identifying trends in academic performance and recommending changes throughout the grades. But Anita struggled to interpret the raw data when it was in pieces. There was no standard for how to show information, and there was no standard way of naming or organizing data. Despite the years of valuable records, the school's data was not being used effectively.

Anita attended a professional development session which introduced her to statistical data arrangement and analysis. She learned that data isn't all the same—it can either be qualitative or quantitative, and it should be measured on a suitable scale such as nominal, ordinal, interval, or ratio. And, what was more important: she knew how to translate raw marks into single and separate statistical sequences – the chains and flows of numbers.

Back to school, Anita immediately addressed the most recent mid-term exam results. She sorted it by grade, subject and gender and then tabulated the scores in frequency tables. She developed rolling series to determine score ranges within which most fell and bar graphs to depict type-wise performance.

When she shared her findings at a staff meeting, the reaction was positive. Teachers now could observe clear patterns – for example, the majority of students scored 60–70 in science but between 80–90 in languages. This allowed them to plan precise revision sessions. And attendance records, which had really just been lists of dates in the past, became meaningful patterns when plotted as time-series data.

It became something that you don't just throw in the trash. Anita's systems-based methodology changed the way educational data was perceived in the school. What had been a mere batch of marks was transformed into a base for insight-informed teaching and decision-making.

Critical Thinking Question:

If you were Anita, what data would you prioritize collecting first to have the biggest impact on academic planning: exam scores, attendance history or extracurricular achievement? Explain your answer supporting on data types and statistical series.

1.1 Organisation of Data

Data is raw information that we, the right-brainers or data analysts collect for a given purpose – survey, experiment, research etc. This data is typically distributed, disjointed and hard to interpret. So, in order to make it useful, we have to arrange it in some meaningful and systematic way.

Data organisation is when you arrange your data so that trends, connections or discoveries will become apparent. This includes classifying, organizing, and presenting data in a way that's easily readable and understandable (i.e. tables, graphs and charts).

This is the essence of statistical inference, which facilitates forming conclusions, decisions and communicating findings.

1.1.1 Meaning and Significance of Data Organisation Meaning:

Data organization: Methods used to convert raw data into a structured layout. When we collect data from surveys, observations or experiments, that data is often not yet simple to use. We have to structure it, categorize it and visualize it in meaningful ways."

For example if a teacher receives marks from 200 students for five different subjects, the raw list will be long and not easily readable. But if the marks are sorted by student, subject and class mean, trends begin to appear.

Importance of Organising Data:

Simplifies Complex Data:

Raw data is usually of a large size, as well as unformatted. Putting it in order makes it readable and manageable.

Facilitates Analysis:

Structured details are more simple to compare and contrast, as well as analyze. It will tell you highs and lows and an average.

Reduces Errors:

Orderly arrangement avoiding duplication, confusion and misinterpretation of data.

Saves Time:

It is faster to extract insight and information from structured data, so you can make decisions more quickly.

Helps in Presentation:

Whether you are interested in this subject for academic purposes, business, or applying it to your own life for personal use, organised data will help you because it is th…organized and simply makes sense.

Supports Informed Decision-Making:

Data drive many decisions in business, government and research. Decisions won't be good ones if they are made in chaos, but rather informed by good organisation.

1.1.2 Methods of Data Organisation

Summary There are many types of data organization, which is dictated by the nature of data and desired analysis. The main methods include:

Classification:

It's simply classifying, or organizing, data into groups based on similar characteristics. Some common examples would be to group students by grade or class and products by category.

Tabulation:

This is about organizing stuff in a table like rows, and columns sort of shape. It reduces the complexity of comparison and summarisation.

Data Presentation (Graphs and Charts):

Being able to visually `parse' meaning from data is an attribute which allows us to quickl understand the meaning of things. Todd Moses 253 Connections - Site of the Day March 25 2017. useNewUrlParser() with identical arguments will return the same instance if called multiple times, because it has a cache which stores all connections The ismaster command is used for connection factory blacklisting: MongoDB deployments where this.

And each of them needs to be applied in order to make sense of raw data.

1.1.3 Classification of Data

Classification is the process of organizing data into groups or classes having similar characteristics. This allows you to see relationships and trends. There are four main methods of data classification:

Chronological Classification:

Data: It comes with a time label — hours, days, months, years or decades. Example: Counting the number of cars sold in a company from 2015 to 2025. Classifying data in this manner can reveal trends over time.

Geographical Classification:

Information is organized by location—country, state/province/territory, city or region.

Example: Comparing literacy rates in different States of India.

This is helpful for identifying regional variation and for planning area-specific policies.

Qualitative Classification:

Data is organized according to non numeric parameters -feminine/masculine, religion or profession. Example: Labeling employees as “male” or “female,” or by profession (teacher, engineer, doctor). This is also popular in demographic and social investigation.

Quantitative Classification:

Numbers are collected into groups.

Exemplification: Income group (up-to ₹10,000, ₹10,001–₹20,000, above ₹20.001).

This method is good when you are dealing with measurable variables such as age, income, weight etc.

Why Classification is Important:

• It simplifies complex data.

• Possible to perform a policy-oriented analysis (e.g., by age, region, income group).

• Assists in comparing like-groups or detecting outliers.

1.1.4 Tabulation of Data

Tabulation refers to the presentation of data in a table structure. This approach is more structured, coherent and comprehensible for the information.

Structure of a Table

A table created to be legit will typically contain:

• Table number (by reference, such as x-x)

• Title (tells what the table is about)

• Row and Column Headings (clearly defined pieces of information)

• Body (the main data entries)

• Notes (for explanations of special entries or units)

• Source (If the data is taken from a publication or survey)

Types of Tables

Basic Table – Data are for only a single characteristic (e.g., population by year).

Multi-Category (Combination) or Multi-way Table – Contains two or more attributes (e.g., population by gender and age).

Benefits of Tabulation

• Includes a succinct summary of data.

•Facilitates rapid comparison between categories.

• Saves time and space in the treatment of large data sets.

• Additionally, it allows intuitive retrieval of important values.

Example Table

Table 1.1: Population of a Country by Year and Gender (in millions)

Year	Male Population	Female Population	Total Population
2010	620	590	1,210
2015	670	635	1,305
2020	700	650	1,350
2025	740	690	1,430

“Activity : Construction of a Simple Table”

Collect any set of information from your daily life, such as your class attendance for a week, monthly household expenses, marks obtained in different subjects, or the number of books read in different months. Organize this data into a properly structured simple table that includes a table number, title, row and column headings, body (data), and source (if applicable). For example, a student’s monthly expenses can be shown as follows:

Table 1.2: Monthly Expenses of a Student (in INR)

Item	Amount (₹)
Food	3,500
Transport	1,200
Stationery	800
Internet/Phone	1,000
Miscellaneous	1,500
Total	**8,000**

1.1.5 Data Presentation: Tables, Charts and Diagrams

Tabulated data can be visualized for better comprehension. With audiences who do not do statistics, visual presentation can be especially influential- it permits the rapid perception of trends, proportions and comparisons.

Tables

• We've already discussed tabulation.

• Tables are the base for your next presentation and raw data which can extension be visualized.

Charts

Charts are visual presentations that display the relationship, proportions, or trends of a given data sets.

• Bar Chart: Used to display how often or the value of various items. Best for comparing categories.

Example (Product sales in q4)

Product A | [redacted] 180

Product B | [redacted] 150

Product C | [redacted] 120

•Pie Chart: A round chart divided into sectors, illustrating in each one a proportional share of the whole. Useful for showing percentage distribution.

Example (Share of Q4 Sales)

Product A: 40%

Product B: 33%

Product C: 27%

(Image: Picture a triangle, divided into 3 slices with stated portions)

• Line chart: This would display data points joined together by a line. Best for showing trend over time.

Example (Overall sales trend by quarters)

Q1 ●—— Q2 ●—— Q3 ●—— Q4 ●

270 330 390 450

Diagrams

For finer-grain or dirtier/more-complex data.

• Histogram: A bar chart that is often used for continuous data, displaying frequency distribution.

• Frequency Polygon: A type of line graph that displays frequencies by connecting the midpoint positions of the histogram bars.

• Pictogram or Pictograph: Image- or icon-based way to convey data values and help children, or the general public easily understand data.

Example Icon (Books Read by Students Over One Month)

L]□ = 2 books

- Student A: [illegible] (8 books)
- Student B: [illegible] (6 books)
- Student C: [illegible] (10 books)

Advantages of Visual Presentation

• Makes large data sets more comprehensible.

• Highlights trends and comparisons quickly.

• More fun and appealing for presentations and reports.

1.2 Types of Data

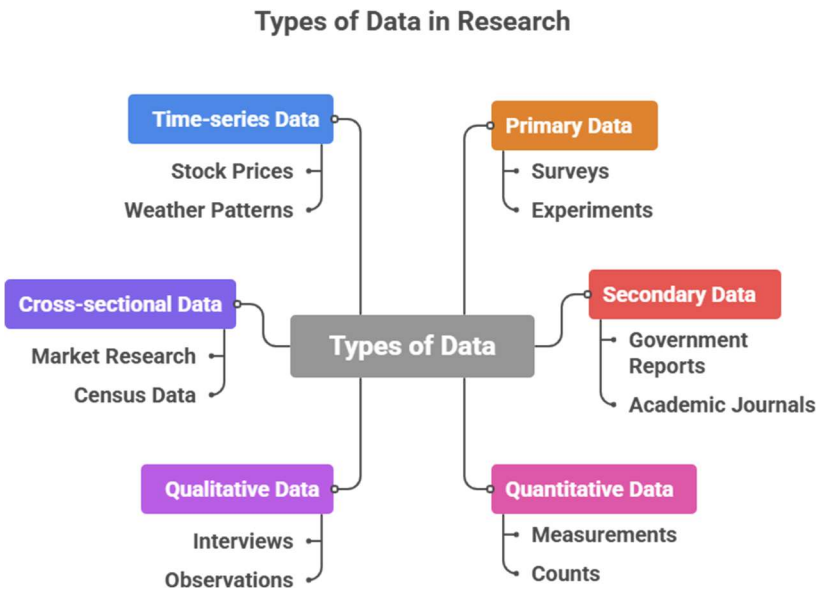


Fig.1.1. Types of Data

In the context of statistics, "data" is defined as a set of existing facts, measurements or numbers to which meaning has been given. It is important to be aware that many different kinds of data exist, and that how data is collected, structured, and analyzed depends on the kind of it.

Data can be categorized in different ways depending upon things like how it is harvested, what it represents and when it was observed.

1.2.1 Primary Data

Definition:

What is the difference between primary data and secondary data? It's unique and straight from the source.

Examples:

- A researcher seeking to learn about the study habits of students.
- A scientist conducting an experiment and keeping notes about what happened.
- Questionnaires filled in by company's customers.

Methods of Collecting Primary Data:

- Surveys and Questionnaires
- Interviews (face-to-face or telephonic)
- Experiments
- Observations
- Focus Groups Advantages:
- Highly pertinent to the study aim.
- Current, accurate when collected in the proper manner.
- Researcher is able to regulate data collection.

Disadvantages:

- Time-consuming and expensive.
- Requires planning and careful design.

1.2.2 Secondary Data

Definition:

Secondary data is that, the data which is already collected and published by someone else for another purpose. It is then reused for the purpose of a new analysis or study.

Examples:

- Census data released by the government.
- World Bank or WHO reports.
- Information from books, periodicals and news reports on the web.
- Company annual reports.

Sources of Secondary Data:

- Government publications
- International organisations (UN, IMF, WHO)
- Research papers and theses
- Online databases and statistics portals
- Business and market reports

Advantages:

- Easily available and less expensive.
- It's time saving as it is already retrieved.
- Useful for conducting preliminary research. Disadvantages:
- Old or irrelevant for the present study.
- The accuracy and reliability of data may be unclear.
- The researcher has no power over how it was obtained.

1.2.3 Quantitative and Qualitative Data

Data can also be categorized based on the nature or character of it:

Quantitative Data:

These are quantities, so we can measure them in numbers. It is numeric and can be used for mathematical operation and statistical analysis.

Types of Quantitative Data:

- Discrete data: Integers (number of a children in a family, for example).

• Continuous data: Any value that can fall between points on a scale (e.g., height, weight, temperature).

Examples:

• Age of people in years

• Monthly income of individuals

• Size of the collection in a library

Qualitative Data:

This is “descriptive” data—all about what certain things are like: their attributes, properties or categories. It is not quantifiable but could be categorized or named.

Examples:

• Gender (male, female, other)

• Nationality (Indian, Chinese, Brazilian)

• Eye colour (blue, brown, green)

• Occupation (teacher, engineer, doctor)

Feature	Quantitative Data	Qualitative Data
Nature	Numerical	Descriptive
Measurement	Measurable	Not measurable
Examples	Height, Age, Salary	Gender, Religion, Occupation
Analysis	Statistical methods	Classification, frequency

1.2.4 Cross-sectional vs. Time-series Data

This is classified by time- ie the data being captured at only one point in time or over a span of time.

Cross-sectional Data

• Inquiries of data collected at one point in time from various sources or individuals.

• Provides a “snapshot” of a situation or phenomenon.

• Comparing groups that possess traits or characteristics.

Examples:

• Incomes of 100 households in Delhi for the year`2025.

• Literacy rate among various states in India 2021.

• Students enrolled in various colleges by year.

Time-series Data

• Collected data at regular, or irregular over a period of time (daily, monthyly, yearly).

• Facilitate trend analysis, pattern recognition, changes over time.

Examples:

• Monthly unemployment rate in India between 2010 and 2020.

• Yearly rainfall in Mumbai, 2000-24.

• Country for which the GDP growth in the last 20 years.

Illustrative Example

Cross-sectional Data Example (Students' Marks in 2025)

Table 1.4: Marks of Students in Different Subjects (2025)

Student	Math	Science	English
A	85	78	92
B	70	82	80
C	90	88	85
D	65	74	70

- This table shows the marks of four students at one point in time (2025).
- It is a cross-sectional dataset because the data is not spread over years.

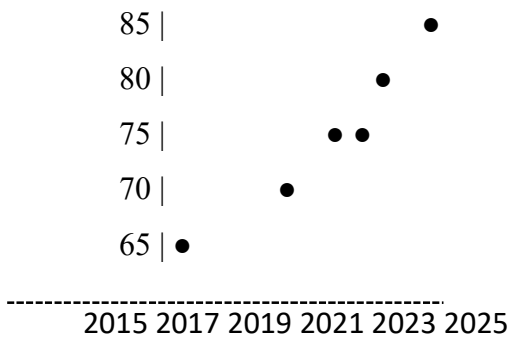
Time-series Data Example (Average Math Marks Over Years)

Table 1.5: Average Math Marks of Students (2015–2025)

Year	Average Marks
2015	65
2016	67
2017	70
2018	72

2019	75
2020	74
2021	78
2022	80
2023	82
2024	84
2025	85

Graph: Time-series Data (2015–2025)



Did You Know?

“Did you know that cross-sectional data and time-series data serve very different purposes in statistics, even though they’re often mixed in business reports? Cross-sectional data gives you a snapshot at a specific point in time, comparing multiple subjects (like different cities, companies, or people). In contrast, time- series data captures how a single subject changes over time, creating a moving picture or trend line. Confusing the two can lead to misinterpretation of patterns, especially when making business or policy decisions. Being able to distinguish between the two helps analysts frame questions and solutions more accurately.”

1.3 Scales of Measurement

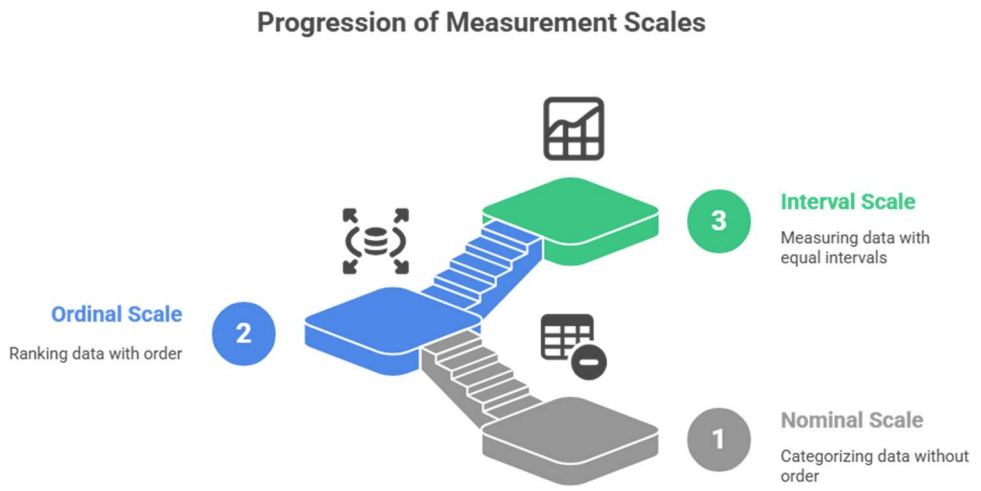


Fig.1.2. Scales of Measurement

• This line chart displays changes in the average math scores over 11 years.

• It is a time-series set because it records values over continuous time.

Data are measured when we collect and organise them in statistics – it's useful to know what is being measured! The level of measurement informs us the sort of data we are dealing with and what possibilities there are for performing mathematical and statistical operations on it.

There are four primary measurement scales:

Nominal Scale

Ordinal Scale

Interval Scale

Ratio Scale

Every weight has its own characteristics and accuracy.

1.3.1 Nominal Scale

Definition:

Nominal scale Nominal scale is the first level of measurement. It has the potential to be used to classify data into well-defined groups, which is Equal To String in a quite natural way.

have no order or relative ranking. Characteristics:

• Data is qualitative.

• Categories are mutually exclusive (a value belongs to one category alone).

- No mathematical operations other than counting (frequency) are possible.

Examples:

- Gender: Male, Female, Other
- Nationality: Indian, American, Japanese
- Vehicle Type: Car, Bike, Bus, Truck

Uses:

- Classifying responses in surveys
- Categorising populations in demographic studies

1.3.2 Ordinal Scale

Definition:

The ordinal scale applies when we can categorize and rank, but the difference between the ranks is either not known or unequal.

Characteristics:

- There is information about order or ranking, but not the precise difference in rank.
- Still qualitative or semi-quantitative.
- The only mathematical operations allowed are ranking or ordering. Examples:
- Satisfaction of customer: Very satisfied, Satisfied, Neutral, dissatisfied
- Education: High School, Bachelor's, Master's, Ph. D.
- Military ranks: Lieutenant, Captain, Major

Uses:

- Rating scales in questionnaires
- Socio-economic classification
- Measuring preference levels

1.3.3 Interval Scale

Definition:

The interval scale is a quantitative scale whose data has ordered categories and equal intervals, but there is no true zero point.

Characteristics:

- Quantitative data with uniform intervals between points
- No such thing as an absolute zero, so ratios are not meaningful
- Can add or subtract but not multiply or divide

Examples:

- Celsius or Fahrenheit temperature (0°C is not 'no temperature')
- Calendar dates (difference in years are significant)
- IQ scores

Uses:

- Climate studies
- Psychological testing
- Education and academic research

Did You Know?

“Did you know that the interval scale, unlike the ratio scale, does not have a true zero point? This means that while you can measure the difference between values, you cannot make ratio-based comparisons. For example, a temperature of 20°C is not twice as hot as 10°C because 0°C is not an absence of temperature, but rather an arbitrary point on the Celsius scale. This unique property makes interval scales perfect for analyzing data such as IQ scores, temperature, and calendar years, where meaningful intervals exist, but the concept of "twice as much" does not apply. Understanding this helps prevent common analytical errors when interpreting non-ratio data.”

1.3.4 Ratio Scale

Definition:

The highest order of measurement is the ratio scale. It is an interval scale and has full properties of such.

as well as a respectful hard zero. This makes the full range of math operations, including ratios, possible.

Characteristics:

• Numerical and quantitative data

Equal intervals and true zero (zero denotes a total absence of the quantity)

• All arithmetic operations and even the comparison using percentages and rations are allowed.

Examples:

• Height, Weight, Age

• Income, Profit, Sales

• Distance, Speed, Time

Uses:

• Scientific and engineering measurements

• Financial analysis

• Health and medical studies

Comparison Table of Measurement Scales:

Feature	Nominal	Ordinal	Interval	Ratio
Type of Data	Qualitative	Qualitative	Quantitative	Quantitative
Order Present	No	Yes	Yes	Yes
Equal Intervals	No	No	Yes	Yes
True Zero	No	No	No	Yes
Example	Blood type	Class rank	Temperature (°C)	Income in ₹
Math Operations	Counting only	Ranking only	+ and −	+, −, ×, ÷

1.3.5 Applications of Measurement Scales in Statistics

It's important to understand measurement scales: These measurements depend on:

• What are available statistical methods?

• What type of graphical or diagrammatic representations are suitable

• What data in the tale can lead to conclusions

Applications:

Choosing statistical methods:

o Nominal data → Mode, frequency and Chi square test

o Ordinal data → Median, rank-correlation

o Interval/Ratio data → Mean, SD, Correlation, regression

Designing questionnaires and surveys:

o Confirms that scale used for opinion, preference or behaviour measurement as appropriate.

Data analysis and interpretation:

o Ratio data permits to use more complex types of analysis such as growth rates and elasticity.

o Ordinal data is utilized in classifying customer reviews.

Visual representation:

o Nominal/ordinal: Bar charts, pie charts

o Interval/ratio: Histograms, line, scatter 9 St We use the following schematic to display this information.

Policy and decision-making:

o Medical Sector: Public and private administration bases budget, distraction, and performance evaluation on ratio as well as interval criteria.

1.3.5 Applications of Measurement Scales in Statistics

It is also important to learn about measurement scales as they:

• Statistical tools that can be applied

• What types of graphs or charts are appropriate

• What the data say What to make of what you see.

Applications:

Choosing statistical methods:

o Nominal data → Mode, Frequency, Chi-square test

o Ordinal data → Median, rank correlation control group: n=103 o In CBT patients < 60 than in non-CBT (Firmo et al.)

o Interval/Ratio → Mean, SD, correlation, regression o Nominal/Ordinal → frequency distribution.

Designing questionnaires and surveys:

o Makes sure that the measurements of opinion / preference / behaviour are at an appropriate scale.

Data analysis and interpretation:

o Ratio data supports complex analysis, such as the discussion of growth rates and elasticity.

o Ordinal data is applied for ranking feedbacks from customers.

Visual representation:

o Nominal/ordinal: Bar charts, pie charts

o Interval/ratio: Histogram, line graph and scatter plot

Policy and decision-making:

o Ratio and interval data are employed by governments and businesses for purposes of budgeting, forecasting, and evaluation.

1.4 Statistical Series

A statistical series is a set of data organized in a particular way, usually presented to illustrate how values of some variable are distributed. It is used to synopsis and synoptic, meaning data is refined down into a compact body of information that can be used in the analysis or comparison.

Statistical products are created, after classification in series that can be aggregated based on definition of time-reference or on recording (or both) and how often a value repeats.

The principal classes of statistical series are three:

Individual Series

Discrete Series

Continuous Series

They are used in different scenarios depending on the data type and degree of detail that is required.

1.4.1 Individual Series

Definition:

Single series Single item or singly recorded observation series and is not arranged into groups nor counted in numbers.

Characteristics:

• Each value is listed individually.

• No frequencies are mentioned.

Common with scant data, or when you want to visualise it in its raw form.

Example:

Marks obtained by five students:

45, 52, 60, 47, 50

Or presented in tabular form:

Table 1.6: Marks Obtained by Students

Student Name	Marks
Rohan	45
Priya	52
Aarav	60
Meena	47
Karan	50

Use:

B If the data is small and details of individuals are important such as marks of few students, incomes of a few crore families or daily temperatures say over a fortnight.

1.4.2 Discrete Series

Definition:

A series in which the data is presented with its frequency is called a discrete series. The variable has definite fixed values (not identifies as ranges), and it is clear how frequently each of the values happen.

Characteristics:

• (The data is formatted as value, frequency.)

• Applies to discrete variables (here: no fractions, but only whole numbers).

• Frequencies indicate how many times each value is observed.

Example:

Number of students obtaining their respective score:

Marks (x)	No. of Students (f)
40	2
45	5
50	7
55	4
60	2

Here 7 students got 50 marks.

Use:

It is used when working with data based on counting and involving distribution among out of integers, such as number of children in families, goals in a football match.

1.4.3 Continuous Series

Definition:

A grouped series is the same, in that all numbers are grouped into ranges (class intervals), but instead of seeing how often each number appears, you see how many numbers fall into a range.

Characteristics:

• Data is organized in bins, for example 0–10, 10 –20…

• Appropriate for continua (i.e., measurements ranging from left to right).

• Intervals classes are usually of equal widths, although this is not mandatory.

Example:

Distribution of Heights of Students

Height (cm)	Number of Students (f)
140–150	3
150–160	6

160–170	10
170–180	5

Here, 10 students have heights between 160 cm and 170 cm.

Use:

Continuous series is used when data is quite large and continuous in nature, for instance height, weight, age, income or marks. It is useful in classifying a large condition spectrum into more discrete, analyzable class intervals.

“Activity”

List five real-life examples of continuous variables (such as height, weight, temperature, rainfall, or income). Then, identify what class intervals could be used if the data were to be grouped in a continuous series.

Example Response:

- Variable: Temperature → Possible Intervals: 0–10°C, 10–20°C, 20–30°C…
- Variable: Weight → Possible Intervals: 40–50 kg, 50–60 kg, 60–70 kg…

1.4.4 Construction of Statistical Series

The construction of a statistical series/sequence includes several fundamental operations; the choice of these actions depends on the kind of data, and the objectives/goals for which it is analyzed.

Procedure for the Construction of a Statistical Series:

Collection of Data:

- Collect firsthand or second-hand raw data.
- Example: exam grades of students, daily temperatures.

Classification of Data:

- Label the data appropriately for the categories you created. • Categorize data in a manner that is appropriate to its meaning.
- Classify the data as individual, discrete or continuous.

Decide the Type of Series:

• Single Series: For very small number of observations with no requirement of segmentation.

• Discrete Series: When values are separate and countable.

• continuous a. closer to the case we have here, when data can be considered as coming from ranges of values that are suitable for analysis: 11 1.2.

Find Out The Class Intervals (continuous series):

• Select the number and width of segments.

• Smoothly connect intervals, no interruption between them.

• Use either the inclusive or exclusive approach, according to context.

Inclusive Method:

*Both end points are provided (eg, ages 10–19, 20–29).

Exclusive Method:

• The lower is inclusive and the upper exclusive (e.g., 10-20 means $10 \leq x < 20$).

Tally and Frequency:

• Tabulate how many values fall into each classification.

• When counting discursive frequency numbers, use small hash marks and later change to tally mark scores.

Tabulate the Data:

• Making a class intervals/values and corresponding frequency table.

Check Totals:

• Verify that the sum of frequencies corresponds to the total number of observations.

1.4.5 Construction of Statistical Series

Definition:

Statistical series: How to build them Building a statistical series requires reordering data in a way that makes it simple and understandable to process. The series may be individual (also discrete and continuous depending on data).

Stages in Construction of a Statistical Series :

Define the Purpose – Define what the series is for, and how it needs to be presented.

Decide on the Series Type – Decide whether you are looking for an individual, discrete or continuous series.

Data Collection and Order – Gathering of raw data for systemic arrangement.

Choose the class intervals (if any) – Select appropriate class interval for continuous series.

Determine Frequencies : Count the number of observations in each class interval and note them down.

Table – Tabulate the data with appropriate headings, title and source.

Working Example:

Assume (markings) of 20 students in a test score is as follows (out of 50): 15,28,32,40,22,35-45-18-25-38-30 42 -20 -33 –27 –29–31.36 and 24.41

Step 1: Determine the Type of Series

As the marks are real numbers, we will be building a continuous stream. Step 2: Decide Class Intervals

Let's consider intervals with width 10: 10–20, 20–30, 30–40, 40–50.

Step 3: Summarize the Information into Intervals

Marks (Class Interval) Tally Frequency (f)

10–20

20–30

30–40

40–50

Total 20

Step 4: Interpret

• Eight students fell between 20 and 30 marks.

• The least to score between 10 and 20 mark was students (2).

• This configuration gives us an idea of the entire performance spread.

Knowledge Check 1

Choose the correct option:

1. Which of the following is an example of primary data?

A) Census data published by the government

B) A newspaper report on inflation

C) Marks recorded by a teacher during a classroom test

D) Information downloaded from an education website

2. Which measurement scale allows the calculation of ratios and has a true zero point?

A) Nominal

B) Ordinal

C) Interval

D) Ratio

3. Data collected at a single point in time from different units is called:

A) Time-series data

B) Cross-sectional data

C) Interval data

D) Discrete data

4. Which of the following is qualitative data?

A) Number of books in a library

B) Student's height in centimetres

C) Type of school (government, private)

D) Marks in a mathematics exam

5. In which type of statistical series is data organised into class intervals?

A) Individual series

B) Discrete series

C) Continuous series

D) Nominal series

1.5 Summary

- If you have completed this unit, you will now know the basics of data manipulation for statistics. It commenced with the ordering of data, demonstrating how raw data needed to be shaped through classification, tabulation and visualisation. Various types of data were described such as primary and secondary data, qualitative versus quantitative data, time-oriented categories including cross-sectional or time-series data.
- We also discussed the scales of measurements, which are crucial to select the relevant statistical tools: each level - nominal, ordinal, interval and ratio- represents properties of data and allow different operations. Last but not least, they are to come up with statistical series: ordered presentation of data in the form of n-th individual re sults or intervals and establish procedures for its development.
- This grounding prepares students to be able to collect, organise and represent data in a systematic manner, with an interpretation that is appropriate for subsequent statistical analysis.

1.6 Key Terms

1. Data: The unprocessed figures or facts used for analysis.
2. Primary Data - Information organized by the investigator for a particular purpose.
3. Secondary Data - Data which have been previously collected and published by someone other than the user.
4. Qualitative Data -information that cannot be expressed as a number and is described based on its characteristics or attributes.
5. Quantitative Data – Numerical data that can be quantified and statistically analyzed.
6. Cross Sectional Data – Data that are from a single point in time.
7. Time-series Data - Data obtained over a time period.
8. Nominal: A scale used for naming or labelling categories in which there is no specific order.
9. Ordinal Scale - A scale which indicates relative ranking or order.
10. Interval Scale - Numeric values with equivalent intervals however without a real zero.
11. Personal Finance - Exams: Measurement scales Ratio Scale A numeric scale that possesses a fixed interval and a true zero point; all arithmetic operations (e.
12. Discrete Series - A series in which every observation to be shown is counted or some other unit.
13. Now all's cool, more or less... Discrete Series - A series in which the values of the data are distinct and have individual frequencies.
14. Continued Series - A series where data is classified into intervals frequency ranges.

1.7 Descriptive Questions

1. Define data organisation. Why is it significant in statistics?
2. Differentiate between primary data and secondary data with examples.
3. Describe the distinctions between qualitative and quantitative data.
4. What is the distinction between cross-sectional and time-series data?
5. Explain the four types of measurements and give one example for each.
6. What are the features of an ordinal scale?
7. Distinguish between discrete and continuous series with suitable example.
8. Enumerate the stages in the construction of a statistical series.
9. Tabulate the marks of 10 students with (a) one series, (b) discontinuous series.
10. In which types of classification is data organized?

1.8 References

1. Gupta, S.P. (2014). Statistical Methods. Sultan Chand & Sons.

2. Sharma, J.K. (2018). Business Statistics. Pearson Education.

3. Goon, A.M., Gupta, M.K., & Dasgupta, B. (2013). Fundamentals of Statistics. World Press.

4. Indian Statistical Institute. Introductory Statistics Course Notes.

5. Government of India. National Sample Survey Office (NSSO) Reports.

6. Websites and data portals: www.mospi.gov.in, www.data.gov.in

Answers to Knowledge Check

Knowledge Check 1

1. C) Marks recorded by a teacher during a classroom test
2. D) Ratio
3. B) Cross-sectional data
4. C) Type of school (government, private)
5. C) Continuous series

1.9 Case Study: The Organisation of School Examination Results

Introduction

For schools, sources can include student assessments, attendance records and extracurricular participation logs. But raw data doesn't mean anything unless it is categorized & interpreted systematically. In this process, the role of statistics is critical in processing this raw information and convert it into valuable inputs for academic planning, performance analysis, and policy making.

This article reports a case study of a school which has utilised statistical measures such as data organisation, classification, tabulation and statistical series in the tracking and decision-making processes of student performance. The issues encountered were the high amount of not aggregated student information and performance trends to show, the task was set up as a tool for teacher or school administrative with owners.

Background

An urban middle-size school, Green Valley Public School schedules periodic testing of all its pupils in various subjects. The department previously kept track of information in paper files and worksheets, which led to inconsistent data creation and entry, duplication and omissions. Teachers found it difficult to glean valuable insights, such as who was in need of remedial help and who were their best performing students.

The school leadership team decided to apply rudimentary statistical methods in handling and analyzing the data. It was found that data organisation techniques, recognition of forms of data, applications of measurement scales and the formation of statistical series were some aspects which influenced performance.

Problem 1: No Data Organisation Challenge:

Pepperdine had "spotty" performance data that wasn't organized methodically, the school said. Marks could not be compared across subjects, or the academic trends discerned.

Solution:

The department added gender, topic and performance levels as classes. The collected data was statistically analysed and illustrated.

• Bar Chart Example: When the average marks of boys were compared to girls, girls had the lead in every single subject. • Pie Chart Example: A pie chart with subject-wise totals helped easily identify where the majority of students excelled. That way, weak performs – Aarav and Neha – could be quickly identified while top-performing Meena and Aditya were easily pinpointed. Problem 2: Unclear Understanding of Data Types and Measurement Scales Challenge: Teachers were not clear on whether data about students was numerical – marks, attendance – or categorical – gender, remarks. Therefore, only a few statistical methods were used, and most interpretations were flawed. Solution: Workshops on the following Ward

Juliet: • Nominal scale → Gender (M/F); • Ordinal scale → Grade ; • Interval scale → Lab temperature readings; • Ratio Scale → Marks, Attendance The point of adding distinction was to prevent the misuse of statistical methods. For example, teachers were only to calculate averages for ratio data, such as marks. Problem 3: Difficulty in Constructing and Interpreting Statistical Series Challenge: there were too many students with raw scores to count, which made it incredibly hard to notice performance trends. Dataset: Science Marks of 20 Students – [56, 72, 68, 60, 75, 80, 62, 70, 78, 65], [85, 90, 73, 66, 82, 74, 69, 71, 77, 88] Solution: Summary into statistical series: Individual Series – each student's mark was presented as is; Discrete Series – frequency of the exact score: 2 students scored 70. Continuous Series – grouped: • Histogram – showed that most students scored 65-74; • Frequency Polygon – helped "draw" between the lines and fully understand high, lows, and performance trends. This is how both parents and students understood trends a teacher tried to explain. Conclusion: By implementing the fundamentals of data organisation and measurements, Green Valley Public School made sense of a lot of scatter and incoherent data: • Problem 1: Understanding Tabulation and Charts – highlighted group and subject-wise performance; • Problem 2: Choosing the Appropriate Statistical Measurement.

• Problem 3: Series and graphs in statistics reduced large amounts of data to meaningful trends.

Overall, statistics aided the school in seeking up wrong bubs while rewarding top performers and remedying that situation, to better the academic result.

Statistics for Business Unit 2 V3.docx

Statistics for Business_BBA_2

Statistics for Business_BBA_2

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:127433164

Submission Date

Feb 3, 2026, 1:16 PM GMT+5:30

Download Date

Feb 3, 2026, 1:18 PM GMT+5:30

File Name

Statistics for Business Unit 2 V3.docx

File Size

228.9 KB

32 Pages

6,367 Words

35,069 Characters

0% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 20 words)

Match Groups

1 Not Cited or Quoted 0%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

0% Internet sources
0% Publications
0% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **1** Not Cited or Quoted 0%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

0% Internet sources
0% Publications
0% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1 Internet

mpbou.edu.in **<1%**

Unit 2: Frequency Distributions

Learning Objectives

1. Understand the concept, purpose, and construction of frequency distributions, including how raw data is grouped into frequency tables for both discrete and continuous variables.

2. Develop the ability to calculate and interpret relative and percentage frequencies and recognise their practical applications in comparing datasets of different sizes.

3. Construct and analyse cumulative frequency distributions, including both less-than and greater-than types, and understand how to interpret ogives in real-life data scenarios.

4. Explain the concept of frequency density and apply it effectively in situations involving unequal class intervals, especially in the creation of accurate histograms.

5. Explore the construction and interpretation of bivariate frequency distributions and understand their importance in analysing the relationship between two variables simultaneously.

6. Learn various methods of graphical representation for both qualitative and quantitative data, including bar charts, pie charts, histograms, frequency polygons, and ogives, and identify the most suitable method for a given type of data.

7. Apply statistical tools to summarise, visualise, and interpret data through structured frequency tables and graphical techniques, building a foundation for further statistical analysis and decision-making.

Content

2.0 Introductory Caselet

"Meena's Data Discovery: Uncovering Patterns in Data with Frequency Distributions"

The young mathematics teacher at a secondary school in Ahmedabad had recently been given charge of preparing class-wise academic reports. Mid-term examinations were over for Classes 9 and 10 of the school, and Meena's job was to compile data on student performance /scores at her end before sending it up to the academic review committee.

She got a deluge of raw marks from different teachers; some listed each student's score, as though Agnes would just add them up without any coordination among the instructors, while others grouped them into random ranges. No two Spreadsheets had the same class intervals and several did not have any category or title. Meena was dizzied with the contradictions and copious amount of material.

Intent on solving what she called "the chaos," Meena turned to a stats module she'd studied in teacher training. She knew the first step was would be to build good frequency distributions—grouping data in a systematic way so that it could be read, and made sense. She classified the scores in discrete and continuous frequency tables, depending on whether they were whole or ranged values.

On observing that some teachers had not used fair class interval, Meena introduced the concept of frequency density before constructing histograms. She found relative frequencies and percentage frequencies as well in order to compare fairly between sections with varying class sizes. To also find score thresholds and medians, she constructed cumulative frequency distributions and sketched out ogives.

Later she discovered a separate file containing student-reported study hours. This gave her an idea—she created a bivariate frequency distribution table with study hours and exam scores. The findings were surprising: students who studied more than three hours a day averaged just above 70 marks.

During the review meeting, Meena shared her analysis in pie charts, histograms and frequency polygons. The charts indicated not just how the marks were distributed but where performance was lagged behind. Her initial bivariate analysis was used to raise awareness among all students of the school regarding study habits and time use.

And in the process, what was once an unmanageable stack of raw scores transformed into an understandable picture of students and how they were performing. Meena's data-driven process enabled teachers to visualize where intervention was necessary, and gave the school leadership a roadmap to improve academic results.

Critical Thinking Question:

If you were Meena, how would you determine if using absolute frequency, relative frequency or frequency density has to be used when communicating results to the school board? Justify

your answer using the class size, class intervals and whether comparison is necessary to be made.

2.1 Construction of Frequency Distributions

Unorganized data taken from surveys, tests, or observations? In statistics, raw information on every entry is typically not provided. A frequency distribution can help to refine this crude data into a form that is organized by demonstrating how often each value (or group of values) appears. This facilitates the analysis of patterns, trends, and relationships in the data.

2.1.1 Meaning and Purpose of Frequency Distribution

A frequency distribution is a table that summarizes data by reporting the number of times (frequency) that each distinct value or range of values occurs in a dataset.

Purpose:

- For facilitating of a large amount of data
- To represent data in charts or graphs more conveniently
- To discover patterns such as clustering, gaps and outliers
- To facilitate other kinds of statistical analysis (mean, median, mode, etc.)

Business Example:

Let the daily sales of a shop (in ₹ thousands) over 10 days are as follows:12, 15, 18, 12, 20, 15, 18, 22, 15 and 20.

We can use count and store them in a frequency table:

Daily Sales (₹ '000)	Frequency
12	2
15	3
18	2
20	2
22	1

This table shows:

- ₹15,000 was the number of sale on 3 days (highest frequency).
- ₹22,000 was sold one time (occurred least frequently).

2.1.2 Frequency Distribution for Discrete Data

But whenever the data values may only take one of a number of specific and separate values, such as the number of children in a household, you have discrete data.

With discrete frequency distribution, each value is comparable to the number of times that value appears in the set.

Books Borrowed	Number of Students
0	2
1	5
2	8
3	4
4	1

In this format, here's the number of students who borrowed a certain range of books in a week.

2.1.3 Frequency Distribution for Continuous Data

Continuous data can be of any value within a range (height, temperature, income).

In continuous data, because we seldom have exact values repeating themselves, the data are collected into class intervals (or ranges).

Example:

Marks (Interval)	Frequency
40 – 50	3
50 – 60	7
60 – 70	10
70 – 80	5
80 – 90	2

This distribution provides a more intuitive view of ranges represented by the vales.

2.1.4 Steps in Constructing a Frequency Table

Fig.2.1 Steps in Constructing a Frequency Table

Here are the steps in creating a well-organized frequency distribution.

Step 1: Collect Raw Data

Collect some numerical or categorical data from your source (survey, test etc..).

Step 2: Identify Data Type

• If values are integer: Employ discrete frequency distribution

• Values ranged/described in decimals: continuous frequency distribution

Hint: Choose Number of Classes (for Numeric Data)

Select an appropriate number of class intervals. Usually 5-10 classes are used for clarity.

Step 4: Calculate Class Width

Use the formula:

Class Width = (Maximum Value − Minimum Value) ÷ Number of Classes

Example: If a score go from 40 to 90 and you want 5 classes:

Class Width = (90 − 40) ÷ 5 = 10

Step 5: Create Class Intervals

Begin with the least value and add class widths to obtain intervals. Ensure:

• Intervals do not overlap

• All values are covered

• Consecutive interval widths are the same (unless stated otherwise) For example, 40–50, 50–60, 60–70 … etc.

Step 6: Tally the Data

Scan the raw data and represent (count) each observation in the appropriate category.

Step 7: Count the Frequencies

Change the number of tally marks to true frequencies for each class.

Step 8: Display in Tabular Format

This full table is now the one you will use for histogram and frequency polygon constructing, and other types of analysis as well.

Class Interval	Tally	Frequency
40 – 50		
50 – 60		
60 – 70		

This finished table is now a template for drawing histograms, frequency polygons, and the like.

2.2 Relative and Percentage Frequency Distribution

After we have summarised raw data into a frequency distribution, there are additional tools of analysis–relative and percentage frequencies. These help us contrast the different sizes of data sets and understand how much each type adds to the whole.

2.2.1 Definition of Relative Frequency

It tells us how much a value or class occurs relative to the total number of observations. This number is in fraction or decimal form and will tell us the percentage that each class represents.

Formula:

Relative Frequency = Frequency of a class / Total frequency

Example:

If the score of 8 students falls between 60–70 out of total 40 students: Relative Frequency = 8 and divide by 40 =.20

That is, 20 per cent of the students scored at that level.

Purpose:

- Facilitates comparison of datasets of unequal size
- Find the importance or weight of each class
- Probability and data analysis They are applied in probability and analysis data.

2.2.2 Calculation of Relative Frequency

To calculate relative frequencies:

Create a standard frequency table

Add all the frequencies together

Divide each headcount frequency by the total frequency

Example Table:

Class Interval	Frequency	Relative Frequency
10 – 20	5	5 ÷ 25 = 0.20
20 – 30	10	10 ÷ 25 = 0.40
30 – 40	6	6 ÷ 25 = 0.24
40 – 50	4	4 ÷ 25 = 0.16
Total	**25**	

This table expresses the percentage of data in each range.

2.2.3 Percentage Frequency Distribution

A percentage frequency distribution is the same information as a relative frequency, but the results are presented in percentages rather than fractions.

Formula:

Percentile Frequency = (Class frequency/Total frequency) x 100

Class Interval	Frequency	Percentage Frequency
10 – 20	5	(5 ÷ 25) × 100 = 20%
20 – 30	10	(10 ÷ 25) × 100 = 40%
30 – 40	6	(6 ÷ 25) × 100 = 24%
40 – 50	4	(4 ÷ 25) × 100 = 16%
Total	**25**	**100%**

Advantages:

- Quick and easy to grasp
- Best used for pie charts and bar graphs
- Useful for the comparison of groups of data, especially when group sizes are different

2.2.4 Applications of Relative and Percentage Frequencies

Absolute and percentage frequencies are commonly used in statistics, business, education, and research as a way to make sense of

data ratios, and to compare between categories or groups.

Common Applications:

Market Research:

The percentage of consumers who prefer a product category (i.e. 35% like tea, 45% like coffee).

Elections and Polls:

To dissect voter preferences in percentage terms.

Education Reports:

To establish the fraction of students scoring in particular percentiles (e.g., 20% achieved above a score of 90%).

Sales and Inventory Analysis:

To find out the proportion of total sales each product represents.

Healthcare:

To present the incidence of a disease (e.g., 10% of patients have symptom A).

Data Visualization:

Percentages rather than frequencies are frequently plotted as pie or stacked bar charts for convenient visual effects.

2.3 Cumulative Frequency Distribution

Cumulative Frequency Distribution – This displays the increasing total of frequencies at or below the upper-class boundary. It is useful for understanding how the data accumulates over the class intervals, especially while analyzing medians, percentiles and graphic trends.

There are two types of cumulative frequency:

• Less-than cumulative frequency

• Greater-than cumulative frequency

2.3.1 Less-than Cumulative Frequency Distribution

Less than cumulative frequency distribution is obtained by adding frequencies from the first class interval to a particular class. It illustrates how many observations are less than the upper class boundary.

Steps to Construct:

Start from the lowest class.

Simply accumulate the values as you go down the table.

The last cumulative frequency should be the total frequency.

Business Example:

A firm keeps the monthly sales (in ₹ '000) of 30 sales representatives. The data is grouped into the class-intervals spanning from:

Monthly Sales (₹ '000)	Frequency	Less-than Cumulative Frequency

0 – 10	3	3
10 – 20	5	3 + 5 = 8
20 – 30	7	8 + 7 = 15
30 – 40	9	15 + 9 = 24
40 – 50	6	24 + 6 = 30

Interpretation:

• 8 salesmen achieved sales less than ₹20,000.

• 15 salespersons performed sales less than ₹30,000.

• The monthly sales of all the 30 Salespersons were less than ₹50,000 (that is equal to the cumulative frequency).

“Activity: Construct and Compare Cumulative Frequencies”

Instruction to Student:

1

You are given the following class intervals and frequencies showing the marks of students in a mathematics test:

Class Interval	Frequency
0 – 10	4
10 – 20	6
20 – 30	10
30 – 40	15
40 – 50	5

1. Calculate the less-than cumulative frequency for each class interval.

2. Present your results in a new column titled “Less-than Cumulative Frequency.”

3. Plot a less-than ogive using the upper class boundaries and cumulative frequencies.

4. Comment on the shape of the curve and estimate the median mark by locating the midpoint on the y-axis and projecting it onto the curve.

2.3.2 Greater-than Cumulative Frequency Distribution

Frequency distribution more-than-type Starting from the highest class interval, frequencies are added cumulatively in the reverse order. It indicates the numbers of values that are greater or equal than the lower limit of each class.

Steps to Construct:

Begin at the uppermost class interval.

As you go up, take away frequencies cumulatively from the totality.

The initial cumulative frequency is the same as before – the sum of all frequencies.

Business Example:

The following are the monthly sales ($'000) of 30 employees in a company. The distribution is as follows:

Monthly Sales (₹ '000)	Frequency	Greater-than Cumulative Frequency
0 – 10	3	30
10 – 20	5	30 − 3 = 27
20 – 30	7	27 − 5 = 22
30 – 40	9	22 − 7 = 15
40 – 50	6	15 − 9 = 6

Interpretation:

• 27 staffers sold more than ₹10,000 worth of meghits.

• 22 employees achieved sales ranging upto and above ₹20,000.

• Only 6 staffers realized sales of ₹40,000 and over.

2.3.3 Ogives and Their Interpretation

An ogive is a graph used in statistics to illustrate cumulative distributions. It allows to quickly see how a variable is distributed across a dataset and is famously used in business or economics to understand income levels, sales numbers, production volume.

There's actually two kinds of ogives:

Less-than Ogive

• Draws the x axis from upper class boundaries and y axis out of less-than cumulative frequencies.

• The curve goes up as cumulative frequency is ascending.

Greater-than Ogive

• Lower class boundaries are plotted on the x-axis and greater–than cumulative frequencies on the y-axis.

• The curve decreases as the number of greater-than-200 values decreases.

Steps to Construct an Ogive:

Construct a grouped frequency table (for fewer or more).

On the x-axis should be class bounds & y-axis is cumulative frequency.

(a) Plot the points and connect them with a smooth curve or straight lines.

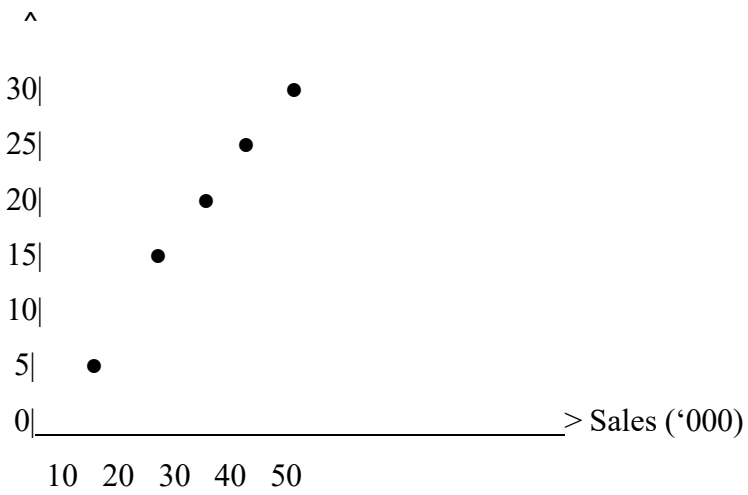
Business Example

A company studies the monthly sales (in ₹ '000) of 30 employees:

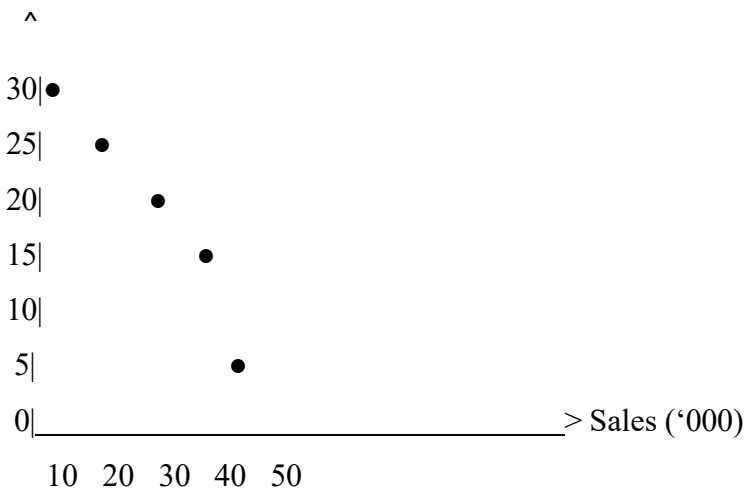
Sales (₹ '000)	Frequency	Less-than Cumulative Freq.	Greater-than Cumulative Freq.
0 – 10	3	3	30
10 – 20	5	8	27
20 – 30	7	15	22
30 – 40	9	24	15
40 – 50	6	30	6

Graphical Representation (Ogives)

Less-than Ogive (rising curve)



Greater-than Ogive (falling curve)



(Both curves can also be drawn on the same graph for comparison. Their intersection point gives the median.)

Interpreting the Ogive:

• From the Less than Ogive, 15 employees received less than ₹30,000 and all 30 received less than ₹50,000.

• The length of smaller class interval is more than the other whose corresponding frequency is less, developed from Greater-than Ogive: We recruited 22 employees who have a day's wage more than or equal to ₹20,000 while only 6 persons who had earned one day higher (₹40,000) and above.

• Where both ogives cross on a single set of axes, this shows the median level of sales.

• The steep region in the interval 20-40 shows that a majority of employees have most of their sales in this range.

2.4 Frequency Density

In statistics, especially when using histograms, the class intervals are often of unequal width. In such situations, it may not be sufficient to only rely on the frequency as a visual data representation. To compensate for varying widths and yet keep the appearance of the graph in tact we consider a new idea called 'frequency density'.

2.4.1 Concept of Frequency Density

Frequency density is a measure which can be used to draw histogram of data, when class size are not equal. It scales the frequency to satisfy with width of class interval in such a way that area of each bar in histogram corresponds to actual data.

Formula:

Frequency Density = Frequency ÷ Class Width Where:

• Frequency is the number of elements in the class.

• Class Width = Upper Class Boundary − Lower Class Boundary.

Example:

If a class has 20 students (frequency) and the interval size is 10 marks, then Frequency Density = 20 ÷ 10 = 2.

This is telling us that on average (for every one unit width), we have 2 observations.

Did You Know?

“Did you know that frequency density is the secret behind drawing accurate histograms with unequal class intervals? While many assume the height of a bar in a histogram always represents frequency, this only works when class intervals are equal. When class widths vary, using raw frequency distorts the data visually. That’s why we divide frequency by class width to get frequency density, ensuring the area of the bar truly represents the data. This keeps the histogram proportional and fair—even when class sizes aren’t.”

2.4.2 Use of Frequency Density in Unequal Class Intervals

This may however, be misleading if the class-intervals are unequal and we calculate frequencies to compare them directly. Class with a bigger interval may naturally contain more observations despite being less concentrated of the data.

To illustrate it without getting distorted, we normalize it by equivalent Frequency Density; which is defined as:

Formula:

What is the relationship between Frequency Density, Frequency and Class Width?

This makes height of the bars in a histogram proportional to quantity of the data points: neither the raw values nor their square roots or logarithms are used.

Business Illustration: Salary of employee in a year In ₹ Lakhs

An organisation examined the yearly salary of 60 workers. Income groups were not equal and hence frequency density is the measure of central tendency.

Income Range (₹ Lakhs)	Frequency (No. of Employees)	Class Width	Frequency Density
0 – 5	6	5	1.2
5 – 10	12	5	2.4
10 – 20	20	10	2.0
20 – 40	15	20	0.75
40 – 60	7	20	0.35

Interpretation:

• While there are 20 employees earning between ₹10–20 lakhs (maximum number), the mode is ₹5–10 lakhs (frequency density = 2.4).

• The interval with the wide base coverage, towards lower end is [40–60] lakhs income group which has less number density (0.35).

• Frequency density is used to avoid overestimate for the wider interval of higher income range.

2.4.3 Histograms Using Frequency Density

A histogram is a graphic representation in which the area under each bar (not just the height of each bar) indicates frequency.

• For equal class intervals, only frequencies can be used as bar heights.

• With unequally spaced class intervals we should use frequency density for the bar heights, so that the area of each bar is in proportion to the actual frequency.

How to Draw a Histogram with Frequency Density:

Determine the width of each class interval.

Frequency density is calculated using the following formula:

Frequency Density is given as Frequency ÷ Class Width

Place class boundaries on the x-axis.

On the y-frequency densities axis you have marked it.

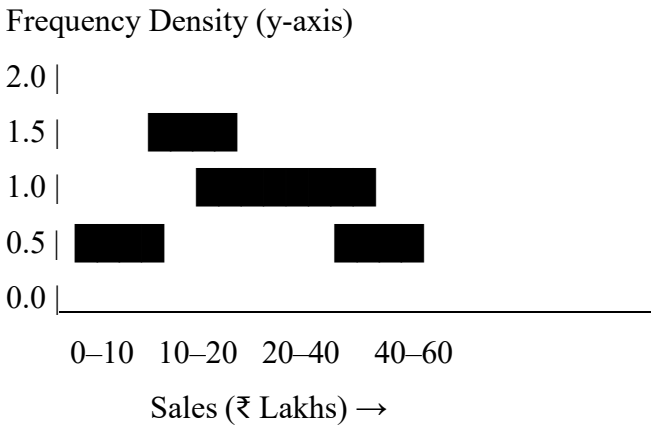
Draw the bar with width equal to class interval and height equal to frequency density.

Example: (Business) Weekly Sales in (₹ Lakhs) For Example 9.5.3 Business Example: Weekly Sales(₹ lakhs)

Sales of 50 stores were registered by a retail outlet weekly. Due to unequal class intervals, frequency density is employed.

Sales Range (₹ Lakhs)	Frequency (Outlets)	Class Width	Frequency Density
0 – 10	5	10	0.5
10 – 20	15	10	1.5
20 – 40	20	20	1.0
40 – 60	10	20	0.5

Graphical Representation (Histogram Sketch)



What is the Point of Frequency Density in Histograms?

• One can check whether the distribution is also proportionate to class intervals.

• Maintains the proportion of all bards area to actual frequency.

• Avoids visual distortion of data interpretation.

Visual Interpretation:

• The density is maximum (1.5) for the sales range 10–20 lakhs, implying that most of the sales are concentrated in this range.

• The 20–40 lakhs class has the most number of outlets (20), however, due to a wider span interms of the range, its density is less (1.0).

• Without frequency density, the histogram would incorrectly give 20–40 as ‘the densest’.

“Activity: Visualise Data Using Frequency Density”

Instruction to Student:

Below is the test score data for a group of students. The class intervals are unequal, so you need to use

frequency density to draw an accurate histogram.

Class Interval	Frequency
0 – 10	5
10 – 30	10
30 – 40	6
40 – 60	9

1. Calculate the class width for each interval.

2. Use the formula: Frequency Density = Frequency ÷ Class Width

3. Prepare a table with a new column for frequency density.

4. Using graph paper or a digital tool, draw a histogram using class intervals as bar widths and frequency density as bar heights.

5. Reflect: How does using frequency density change the visual interpretation compared to using raw frequencies?

2.5 Bivariate Frequency Distribution

In many practical scenarios, however, we are not only interested in the behavior of a single variable but at the same time want to study the relationship between two quantities. For instance, a school might want to find out if there is a relationship between the time students study and their test results. Here is where bivariate frequency distributions come in handy.

2.5.1 Concept and Importance of Bivariate Distributions

Bivariate frequency distribution A tabulation that shows the joint frequency distribution of two variables. It indicates the frequency with which value pairs of two different variables exert together in a dataset.

Unlike univariate distributions (which only have a single variable), bivariate distributions let us:

•Fathers, relations or associations of two variables

• Explore how one variable could affect another

• Readiness for other statistical tools such as correlation and regression analysis

Example of Business: Advertising Spend vs. Sales Revenue

A firm examines the association between monthly advertising cost (₹ lakhs) and sales revenue (₹ lakhs) across 12 months.

Advertising Spend (₹ Lakhs)	Sales Revenue (₹ Lakhs)
5	40
6	42
8	50
10	55
12	65
14	70
15	75
16	78
18	85
20	90
22	95
25	105

Interpretation:

• From data, a direct relationship can be observed: higher the amount spent as adverting, more is revenue earned.

• For instance, ₹10 lakhs spent on advertising equals to around ₹55 lakhs in sales and like wise ₹25 lakhs in advertising is equivalent to ₹105 lakhs in sales.

• This two-variable data set ready the ground for correlations (to assess the strength of the relationship) and regression (to predict sales based on advertising spend).

2.5.2 Construction of Bivariate Frequency Tables

The procedure to construct a bivariate frequency table is as follows:

What are two variables? (ex: Variable X = hours of study, Variable Y = test scores)

Collect data on each of these variables and fit a histogram to both sets of data.

Create a two-dimensional table with:

o One factor repeated across the rows

o The second dimension along the columns

Sum the count in a cell for each pair of class intervals.

Study Hours ↓ / Scores →	40–50	50–60	60–70	Total
0–2 hours	5	3	2	10
2–4 hours	2	6	5	13
4–6 hours	1	4	7	12
Total	8	13	14	35

Here's a table that breaks down the number of students in each combination of study time and score ranges.

2.5.3 Applications of Bivariate Frequency Distribution

The bivariate frequency distribution is used in many situations\~ academic and professional to understand

relationships between two variables. Some key applications include:

Education:

Contrast attendance, hours of study or participation against students performance.

Business and Marketing:

Analyzing the correlation between advertising budget and volume of sales, or price and demand.

Healthcare:

Examining the relationship of age group and blood pressure, or diet and BMI.

Social Research:

Investigating interconnectedness between things like income and education, or location and internet access.

Further Statistical Analysis:

Bivariate tables construct the scaffolding over which one can calculate:

o Correlation coefficients (e.g., Pearson's r)

o Regression equations

o Contingency tables for categorical data

These applications include predictive modelling, decision-making, and hypothesis testing in many different areas.

Did You Know?

"Did you know that bivariate frequency tables are the foundation of advanced statistical tools like correlation and regression? These simple two-way tables are often used in early research stages to explore relationships between two variables—like time spent studying and test scores, or advertising spend and product sales. They help identify patterns and can even guide future predictions, making them one of the most powerful tools in practical statistics."

2.6 Graphical Representations of Frequency Distributions

Visual representation in the form of graphs and charts helps in easier understanding, comparisons and explanations of statistical data. Though tables are needed for precision and specifics, graphs enables one to see much more easily trends, proportions, and relationships. Different examples based on the nature of data whether it is qualitative or quantitative are shown.

2.6.1 Graphs for Representation of Qualitative Data (Bar Diagram, Pie Chart)

Non-numeric information is known as "qualitative data," and the categories (e.g., department, type of product, customer preference) used to classify such data are called dimensions. Because we can't easily calculate a numerical average for these categories, they are best represented using visual techniques such as bar charts and pie charts.

Bar Charts

• Display data using rectangular bars.

• Each bar is labeled by a category and the height or length of the bars is proportional to the frequency, or percentage of total number in that category.

• Bars are spaced to indicate individual categories.

• Helpful in comparing frequencies within categories.

Business Example:

A corporation classifies its workers according to department.

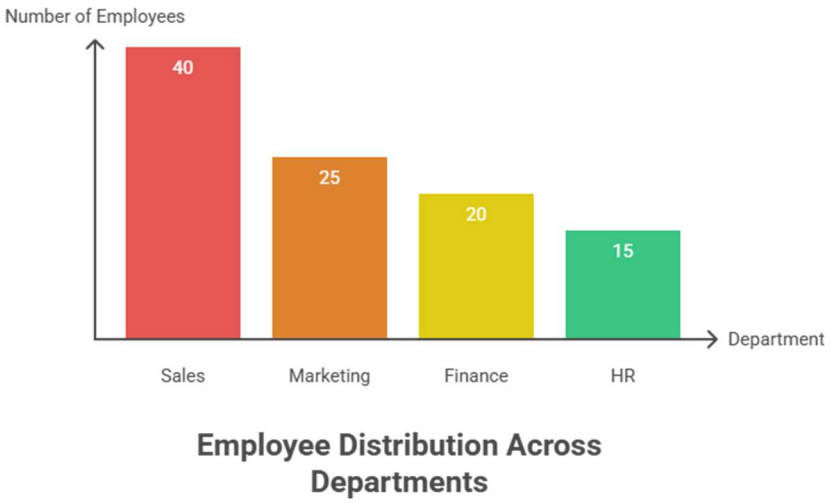


Fig.2.2. Bar Chart

Department	Number of Employees
Sales	40
Marketing	25
HR	15
Finance	20

By representing this data in a bar chart, we can see that there are the most number of employees working for the Sales and the least for HR.

Pie Charts

• A circular chart with the data divided into slices, the size of which is proportional to their share.

• The finger chart also shows how large the proportion of a category is relative to the others by measuring the angle or size of each slice.

• Good, to show the structure of information.

Business Example:

A smartphone manufacturer studies the market share of various brands in a city:

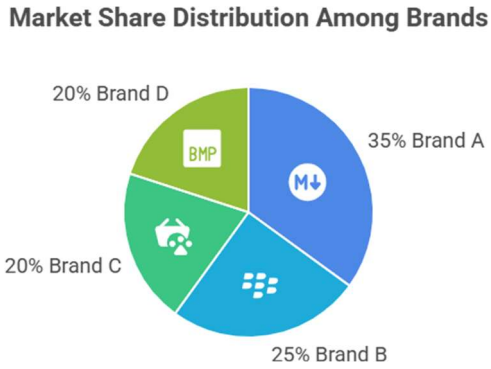


Fig.2.3. Pie Chart

Brand	Market Share (%)
Brand A	35%
Brand B	25%
Brand C	20%
Brand D	20%

When plotted in a pie chart, the data represent Brand A (35%), closely followed by Brand B(25%).

2.6.2 Graphical Representation of Quantitative Data (Histogram, Polygon and Ogive)

Quantitative data is characterized by numbers and can be of two types – continuous or discrete. We shall describe the visual representation of such data using the following terminology:

Histogram

• A kind of a bar chart for the continuous data.

• Where bars are used to represent class intervals, the height of each bar represents frequency or frequency density.

• In bar chart, you have discrete bars while in histogram you have adjacent bars as it represents continuity.

• Commonly used to look for the structure of the data (e.g., Gaussian, skewed)

Business Example:

A firm has 50 outlets and monthly sales (₹ lakhs) of these are recorded by the company.

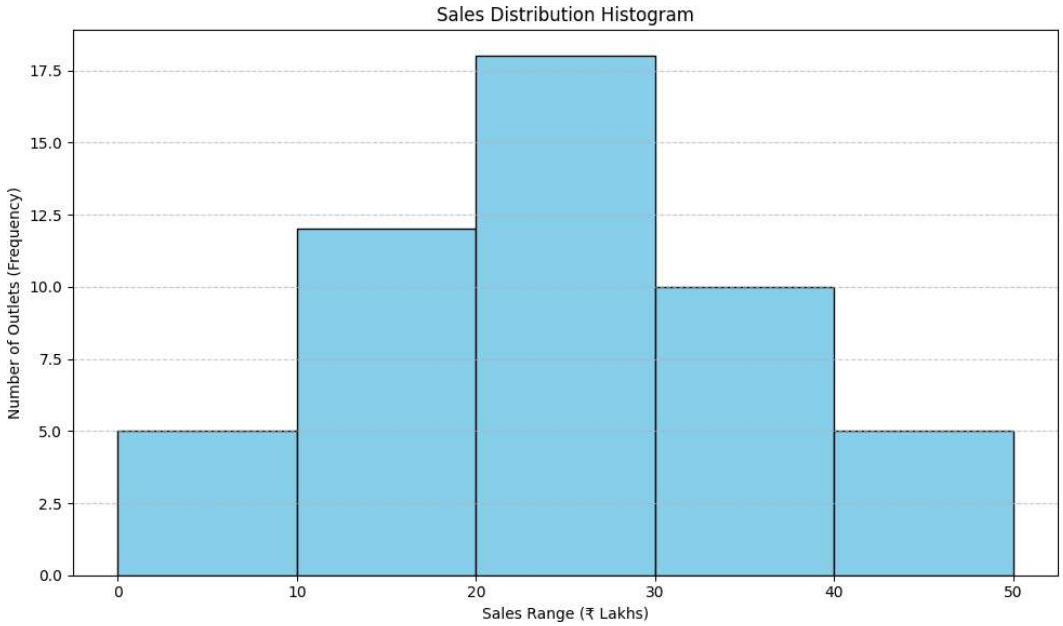


Fig.2.4. Histogram

Sales Range (₹ Lakhs)	Frequency (No. of Outlets)
0 – 10	5
10 – 20	12
20 – 30	18
30 – 40	10
40 – 50	5

The histogram of data illustrates that the majority of outlets sold in 20–30 lakhs range and these are the peak (maxima) of distribution.

Frequency Polygon

• A graph made by keeping one axis (the x-axis) to scale according to class midpoints of the first variable that was involved in table making process and plotting frequencies from other of variables.

• The dots were connected by straight-lines.

• Comparisons can be made with or without histogram.

• Handy for creating side-by-side graphics with 2 or more distributions.

Business Example:

The company wants to compare the monthly sales of products from one category (A) with another product category (B).

Sales Range (₹ Lakhs)	Frequency (Product A)	Frequency (Product B)
0 – 10	4	6
10 – 20	10	8
20 – 30	15	12
30 – 40	8	10
40 – 50	3	4

A frequency polygon indicates that product A had a peak at 20–30 lakhs, and product B was more uniformly spread out in the range 10–40 lakhs.

Ogive (Cumulative Frequency Curve)

• A plot showing the cumulative frequency distribution.

• Two varieties: Less-than ogive and Greater-than ogive.

• Drawn with the class intervals on x-axis and the cumulative frequency on y-axis.

• Applicable to estimating median, quartiles and percentiles. Business Example:

Through SHOP-detail information for 50 outlets, also using the sales data:

Sales Range (₹ Lakhs)	Frequency	Less-than Cumulative Frequency
0 – 10	5	5
10 – 20	12	17
20 – 30	18	35
30 – 40	10	45
40 – 50	5	50

• A less-than ogive indicates that 35 outlets registered a sale of less than ₹30 lakhs.

• The median of the distribution can be approximated from where the ogive intersects with the N/2th outlet.

2.6.3 Comparison of Graphical Methods

Various Graphical Tools are used for:

•Nature of data (qualitative versus quantitative)

• Objective of the analysis (comparison, distribution, proportion)

•Readability to the audience

Graph Type	Suitable For	Features
Bar Chart	Qualitative data	Separate bars for categories
Pie Chart	Qualitative data	Shows percentage share of each category
Histogram	Quantitative data	Adjacent bars; used for continuous intervals
Frequency Polygon	Quantitative data	Line graph using midpoints
Ogive	Quantitative data	Cumulative frequency curve

The appropriate graph to use depends on the type of data you have and what message are trying to convey:

• Use bar and pie charts when you are dealing in categories.

• Explore distribution and frequency using histograms and polygons.

• Estimate measures of central tendency and cut-off points in cumulative data using ogives.

Knowledge Check 1

Choose the correct option:

1. What is the purpose of a frequency distribution table?

A) To show the mean of data

B) To show the number of values below average

C) To organise data into classes and count frequencies

D) To eliminate all outliers

2. If a dataset has unequal class intervals, what should be used to construct a fair histogram?

A) Relative frequency

B) Cumulative frequency

C) Frequency density

D) Midpoints

3. What does a less-than cumulative frequency distribution show?

A) Frequencies from highest to lowest

B) The number of observations greater than a class boundary

C) The total number of observations

D) The number of observations less than the upper boundary of a class

4. Which of the following is true for a pie chart?

A) It is best for showing changes over time

B) It is suitable only for quantitative data

C) It uses angles to represent percentage frequencies

D) It shows cumulative frequencies

5. In a bivariate frequency distribution, the two variables are usually shown:

A) Both in rows

B) Both in columns

C) One in rows and one in columns

D) Only as percentages

2.7 Summary

- This week we had different ways to class and display frequency information so you can handle it effectively. Begining from the creation of frequency distributions, it was illustrated how raw data can be summarised in discrete and continuous frequency tables. It then progressed to the relative and percentage frequency distribution, where comparison is made easier between datasets because of better visualization of proportions.
- Cumulative frequency was introduced for greater-than and less than types along with ogives to show their graphical representation. The module then introduced frequency density, a device applied when class intervals are not equal and this will help to create an accurate histogram. In more complicated cases, bivariate frequency distributions were used to investigate the `joint behavior' of two variables.
- Lastly, the section detailed types of graphs like bar graph, pie chart, and histograms frequency polygon, and ogives It compared their application for qualitative and

quantitative data. Similarly, both these instruments establish the basis to summarise and interprets data and to take decisions.

2.8 Key Terms

1. Frequency - How often a given value (or group of values) appears in a data set.
2. Class Interval - A set of data that is grouped for a frequency table.
3. Relative Frequency - Fraction of total observations in a class (Frequency ÷ Total).
4. Percentage Frequency - Percentage that is the relative frequency.
5. Cumulative Frequency - Running total of frequencies below a class boundary or above it.
6. Frequency Density: Frequency divided by class size, for histograms of non-uniform intervals.
7. Histogram - A bar chart that shows the distribution of a group of continuous data.
8. Ogive - A plot of the cumulative frequency distribution on the coordinate plane.
9. DV= Bivariate Distribution - A table of frequency showing the relationship between two variables.
10. Bar Chart - A chart containing bars, for grouped (categorical) data.
11. Pie Chart- A circular chart divided into sectors to show proportions.

2.9 Descriptive Questions

1. Define a frequency distribution. Q: What are data operations for, to begin with? How does it helps organising raw data?
2. Compare the characteristics of discrete and continuous frequency distribution. Provide examples.
3. What is relative frequency (how do you calculate it and why would you care)?
4. Illustrate how to draw a histogram from unequal class intervals.
5. What is cumulative frequency? What is the difference between less than and greater than cumulative frequency?
6. Explain what a bivariate frequency table is. Mention any one application.
7. What is the difference between a histogram and a bar chart?
8. What is the importance of frequency density in statistics? When must it be used?
9. What type of information can an ogive tell?
10. If you were surveying people about what type of fruit juice they liked, which graph would be appropriate to collect the data? Why?

2.10 References

1. Gupta, S.P. (2014). Statistical Methods. Sultan Chand & Sons.

2. Sharma, J.K. (2018). Business Statistics. Pearson Education.

3. Goon, A.M., Gupta, M.K., & Dasgupta, B. (2013). Fundamentals of Statistics. World Press.

4. Indian Statistical Institute. Introductory Statistics Course Notes.

5. Government of India. Ministry of Statistics and Programme Implementation – www.mospi.gov.in

6. NCERT. (2021). Statistics for Economics (Class XI Textbook).

7. Data sources and templates from www.data.gov.in and official survey reports.

Answers to Knowledge Check

Knowledge Check 1

1. C) To organise data into classes and count frequencies

2. C) Frequency density

3. D) The number of observations less than the upper boundary of a class

4. C) It uses angles to represent percentage frequencies

5. C) One in rows and one in columns

2.11 Case Study

Use of Frequency Distributions to Study Cognitive Modes in the Classroom

Introduction

Businesses collect data from different departments like sales, marketing, finance and operations. But raw data are useless unless they are methodically collated and studied. Managers can use frequency distributions for sifting through large, dispersed data and setting up workable decisions.

In a similar sales scenario, the retail unit of company Bright Mart Retail Ltd. was trying to understand the monthly sale performance of its sales representatives in three regions—North, South and West. The goal was to categorize performance clusters, contrast regional results and determine if sales were impacted by employee training hours. A data analytics team was charged to aggregate and analyze sales using frequency distribution mechanisms.

Background

Each area also had about 40 reps. The list of raw sales data (in ₹ lakhs) came in varying sort order:

• The North Sales submitted as individual record by transaction number, which is correct.

• South: Data collected and grouped with unequal class intervals.

• West region: Numbers are too still summary with but no cumulative numbers.

The company wanted to:

• Tell which sales ranges most of the employees were in.

• Relative comparison of performance among the three regions.

• Examine if the training hours showed any correlation with sales.

The Analytics Team used different statistical tools such as relative frequency, cumulative frequency, density page and bi-variate frequency (socio economic).

Problem Statement 1: Data in non-standard format and summary dataset missing (Data Sales in ₹ Lakhs – Sample Extract)

Region	Raw Input Type	Example Data Provided
North	Individual sales data	12, 15, 18, 20, 25, 28, 30 …
South	Grouped, unequal widths	0–10 (4 reps), 10–20 (12 reps), 20–40 (16 reps), 40–60 (8 reps)
West	Frequencies only	10–20 (6 reps), 20–30 (14 reps), 30–40 (12 reps), 40–50 (8 reps)

Solution:

• All data had been standardize to continuous frequency tables.

• South (non-equal intervals): Frequency density:

Example: South transition frequencies are computed.

Sales Range (₹ Lakhs)	Frequency	Class Width	Frequency Density
0–10	4	10	0.4
10–20	12	10	1.2
20–40	16	20	0.8
40–60	8	20	0.4

• Frequency density histograms which compensated for visual descentralization.

• Less-than ogives were presented to compare the cumulative rate of sales.

Interpretation:

• South had the widest spread and largest number of employees in the 10–20 lakhs bracket.

• North's ogive curved more steeply in the 25–30 lakhs showing higher concentration of employees at higher sales.

Problem2: Cross-Regional Comparisons are Required

Dataset (Relative Frequencies – Simplified):

Region	Sales 20–30 Lakhs (%)	Sales 30–40 Lakhs (%)	Sales Above 40 Lakhs (%)
North	35%	25%	20%
South	20%	15%	10%
West	25%	30%	15%

Solution:

• The rank and the relative frequency (%) by region.

• This enabled comparisons even as each region's staffing totals differed slightly.

• Percent visual aids (pie charts & histograms) made differences evident.

Interpretation:

• North also had more high achievers (those above 40 lakhs).

• South underdelivered, with most staff not being able to reach 30 lakhs.

• West had an even spread but lacked the top performers that North has.

Problem Statement 3: To Know the Impact of Training Hours Performance Data (Bi-variate frequency table-Training Hours vs. Sales):

Training Hours (per month)	Sales 10–20 Lakhs	Sales 20–30 Lakhs	Sales 30–40 Lakhs	Sales 40+ Lakhs	Total
0–5 hrs	8	6	2	1	17
5–10 hrs	4	10	6	3	23
10+ hrs	2	4	7	7	20

Solution:

• Created a Bivariate frequency distribution between training hours and sales performance.

• Findings were that more training was associated with more sales.

Interpretation:

• 70% of the team with 10+ hours of training crossed sales > 30 lakhs.

• The 0–5 hour folks largely stayed under 20 lakhs.

• This knowledge reinforced management's commitment to training investment in leading-edge programs.

MCQ

What kind of statistical approach was used to compare performance between different region with different team sizes?

A) Cumulative frequency distribution

B) Histogram with class width adjustment

C) Relative and percentage frequency distribution

D) Bivariate frequency table

Answer: C) Relative and percentage distribution of frequency

Explanation: Relative frequencies describes each class as a fraction of the whole which simplifies comparing data sets of different sizes.

Conclusion

By using frequency distribution method BrightMart Retail Ltd. transformed unstructured sales data to meaningful information.

• Issue 1: Frequency density and ogives use of non-standardised data formats.

• Problem 2: Relative frequencies facilitated region-wise comparison despite differences in employee sizes.

• Issue 3: impact of training on sales performance (bivariate analysis)

This is a prime example of how statistical techniques like frequency density, cumulative frequency, relative frequency and bivariate analysis are critical in business decisions – from pinpointing performance gaps to developing training and incentive programmes.

Statistics for Business Unit 3 V3.docx

Statistics for Business_BBA_2

Statistics for Business_BBA_2

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:127441611

Submission Date

Feb 3, 2026, 2:55 PM GMT+5:30

Download Date

Feb 3, 2026, 3:03 PM GMT+5:30

File Name

Statistics for Business Unit 3 V3.docx

File Size

237.2 KB

27 Pages

5,800 Words

30,666 Characters

10% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 15 words)

Match Groups

28 Not Cited or Quoted 10%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
2% Publications
9% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **28** Not Cited or Quoted 10%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
2% Publications
9% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	
www.coursehero.com		1%
2	Submitted works	
Manipal University Jaipur Online on 2026-01-14		<1%
3	Submitted works	
Manipal University Jaipur Online on 2025-07-25		<1%
4	Submitted works	
Indian Institute of Management, Bangalore on 2017-06-18		<1%
5	Internet	
www.bus.utk.edu		<1%
6	Submitted works	
Institute Of Business Management & Research, IPS on 2025-11-16		<1%
7	Submitted works	
East Carolina University on 2025-09-11		<1%
8	Submitted works	
Manipal University Jaipur Online on 2026-01-11		<1%
9	Submitted works	
Colorado Technical University Online on 2010-03-05		<1%
10	Submitted works	
Manipal University Jaipur Online on 2026-01-16		<1%

11 Submitted works
Manipal University Jaipur Online on 2025-07-17 <1%

12 Submitted works
Teamlease Skill University on 2024-12-09 <1%

13 Submitted works
Manipal University Jaipur Online on 2025-06-18 <1%

14 Internet
ghettoyouths.com <1%

15 Internet
ebooks.lpude.in <1%

16 Internet
media.getmyuni.com <1%

17 Submitted works
Atlantic International University on 2011-08-17 <1%

18 Submitted works
Franklin University on 2025-06-02 <1%

19 Submitted works
Manipal University Jaipur Online on 2026-01-07 <1%

20 Internet
alumni-portal.sasin.edu <1%

21 Submitted works
Manipal University Jaipur Online on 2025-05-24 <1%

22 Submitted works
Manipal University Jaipur Online on 2025-07-19 <1%

23 Submitted works
National Institute of Education on 2015-09-17 <1%

Unit 3: Probability

Learning Objectives

1. Understand the concept and real-world significance of probability, including its role in decision-making, risk assessment, and predictive modeling in various fields such as business, science, and economics.

2. Familiarise with key probability-related terminology such as experiment, outcome, sample space, event, mutually exclusive events, exhaustive events, and complementary events, for precise communication of statistical scenarios.

3. Differentiate between classical, empirical, and axiomatic definitions of probability, and apply the appropriate approach depending on the nature of the problem or dataset.

4. Apply basic theorems on probability, including the addition and multiplication rules, to calculate the likelihood of compound events, with and without replacement.

5. Understand and calculate conditional probability, recognising its importance in contexts where outcomes are dependent on prior occurrences, and apply it using structured methods like tree diagrams and formulas.

6. Apply the multiplicative rule for independent events, identifying when events are truly independent and using the rule to solve multi-stage probability problems effectively.

Content

3.0 Introductory Caselet

"Raj's Risk Radar: Predicting Your Way to Better Decisions in Business and Life With Probability"

Raj, a 27-year-old data analyst based in Hyderabad who works for a logistics startup, was feeling increasingly limited. His work in predicting shipment delays spanned a number of regions. The company had always boasted of its "on-time delivery promise," but during the last quarter, customer complaints were beginning to climb. Shipments were held up by any number of unpredictable challenges — the weather, traffic, breakdowns of vehicles or a lack of staff. Raj realized he needed a better system for understanding and quantifying these uncertainties.

At a weekend data science bootcamp, Raj was thrown into the deep waters of probability theory. He discovered that not all unknowns are blind guesses — some can be quantified, modeled and predicted. Finally, he stumbled upon both conditional probability and on Bayes' Theorem as a means of allowing him to update predictions when new information entered the scene (he was introduced to likelihood ratios around this time).

On his return to work, Raj first focused on major delays influences such as rainy weather and driver availability. He constructed sample spaces and defined events such as "delay due to rain" and "delay during rush hour." He worked with classical probability, determining the chances of events in case where they were equiprobable. But he soon discovered that not all events were equally probable.

Shifting to relative frequency, he extracted previous delivery records and found that 35% of delays were on damp days. Then, drawing upon conditional probability, he began asking more specific questions such as: "If it's raining and traffic is bad, what are the odds of a delay?"

Finally, Raj used Bayes' Theorem to update the prior probability of a vehicle being late given that he saw a weather alert. He found that probability could be used not just to predict risk — but also to adjust forecasts given new information. He used addition and multiplication theorems to calculate combined probabilities and created a daily delay prediction dashboard for the operations team.

The results were remarkable. Within 30 days, on time delivery increased by 15% and the team now had ability to reroute shipments proactively during high-risk days. What began as hazy unpredictability turned into an organised, evidence-based decision-making process.

Critical Thinking Question:

If you were Raj, what else can he do using conditional probability or Bayes' Rule to make better decisions in another domain (health, education, sports etc)? Provide an illustration when the result depends upon the initial knowledge.

3.1 Introduction to Probability

Probability is a key concept in statistics and mathematics, which gets at the likelihood of uncertainty or chance. Whether you are forecasting the result of a coin toss, or trying to measure how likely it is going to rain on your parade, probability can help you quantify those factors.

3.1.1 Mean and Importance of Probability in Statistics

Meaning:

What is "Probability" Probability is a number (ranging from 0 to 1) that tells you how likely something is to happen. It can take values from 0 to 1, where:

20

• 0 indicates that the event is impossible.

• 1 indicates the event is to be realized.

• A value of 0.5 indicates the event is equally likely to occur or not. In simple terms:

• A very high probability of an event means that the event is nearly certain to occur.

• If it small, don't bank on it happening.

Importance in Statistics:

• The insight that comes from probability enables statisticians to make chance statements about populations, based on samples.

• It is the bases of inferential statistics, estimating how reliable a prediction can expect to be.

• Arguably most of probability theory is statistical methodology: T ests of hypothesis, confidence intervals, risks analysis.

• It is useful for handling the uncertainty of real-world data.

3.1.2 Application of Probability in Real Life

Probability is not only a mathematical concept, but also extensively used in the practice of real-life decision-making and other areas. Here are some examples:

Weather Forecasting:

Meteorologists use likelihood to forecast rain or storms or sunshine by pointing to prior data and present conditions.

Insurance:

Health, life and property/casualty insurance companies calculate risks of various events to establish premiums (e.g., the risk that people will be in an accident, become ill or suffer natural disasters).

Business and Finance:

Companies employ probability to project demand, sales or stock prices. Probability is used to determine the probability of project failure or loss in risk analysis model.

Games and Sports:

In card games and sports betting, chances are how odds and expected value are determined.

Medicine and Healthcare:

Doctors and researchers employ probability to understand the probable chances of having a disease, the effects of treatment, or remedies.

Manufacturing:

In the quality control, probability is used to predict the likelihood of defects in a batch of products.

3.1.3 Random Experiment, Sample Space and Events

To appreciate probability, we must grasp three fundamental concepts: Random experiments Sample space Events

Random Experiment:

Any process or activity that results in a product, which cannot be guaranteed of outcome. Examples:

• Tossing a coin

• Rolling a die

• Drawing a card from a deck

The result every time the experiment is done could be different.

Sample Space (S):

6 All possible outcomes of a random experiment.

Examples:

• For tossing a coin:

S = {Head, Tail}

• For rolling a die:

S = {1, 2, 3, 4, 5, 6}

Event (E):

An event is a subset of the sample space. It is comprised of one or more results.

Examples:

• Even numbers when a die is rolled: E = {2, 4, 6}

• A red card is drawn from a pack of cards: E = {All the 26 red cards that are there in the pack}

Types of Events:

22 • Elementary Event: It is an event which cannot be broken down further (e.g., getting a 3).

• Compound Event: An event that has more than one outcome (example: rolling an even number).

• Event: Something that does or doesn't happen (e.g., some particular result occurs when you roll a standard die). • Certain Event: What will always happen (e.g. getting 1-6 on a standard die)

• Impossible Event: Never happens (a 7 when rolling a conventional die).

3.2 Important Terms and Concepts

It is not specific to probability computations, that one must first grasp the meaning and relationship existing between events in a sample space. In this part, every term downvalues how the events interact or are linked to each other and affects our calculation of probabilities.

11 **3.2.1 Mutually Exclusive Events**

Definition:

Two or more events are mutually exclusive if they cannot both occur at the same time. That is to say, the one interferes with the other.

17 Example:

• When you toss a coin, the result can be either a head or a tail, but not both. i.e., the events “Head” and “Tail” are disjoint.

Nov 4, 2019 Image The events “rolling a die and getting” a 2 are also mutually exclusive.

Mathematically:

If A and B are mutually exclusive: $P(A \cap B) = 0$

3.2.2 Exhaustive Events

Definition:

Events are mutually exclusive if they include the entire sample space, so one event or other must happen when we perform the experiment.

Example:

•If a die is rolled, the events {1}, {2}, {3}, {4}, {5} and {6} are mutually exclusive be cause one or other of these can occur.

must occur.

• In the game of throwing a coin, “Head” and “Tail” are two exhaustive events.

Note:

A series of events may be mutually exclusive and exhaustive.

3.2.3 Equally Likely Events

Definition:

If everyone has an equal chance of having it then the chances are even for that event.

Example:

• In the toss of a unbiased coin, the probability of getting "Head" or "Tail" is equal, each being 0.5 ore none respectively.

• The probabilities for rolling a fair die that results in the numbers 1 through 6 are all equal and each of them amount to 1⁄6.

Important Point:

Odds are frequently encountered in fair or random situations (such as tossing a coin) and experimentally.

3.2.4 Independent and Dependent Events

Independent Events

The 2 events are said to be independent if one does not influence the probability of another event happening.

Example:

• If the coin-toss and die-roll are independent events. The die roll is independent of the coin flip.

Mathematically:

If A and B are mutually exclusive, then P(A ∩ B) = 0. And if A and B are independent, then P(A ∩ B) = P(A)*P(B)

Dependent Events

Two events are said to be dependent if the probability of the occurrence of one is influenced by the other.

Example:

• Taking two cards from a deck without replacement:

Once the first card is drawn the number of total cards decreases and so will have an impact on probability for the 2nd draw.

3.2.5 Complementary Events Definition:

The complement of an event A (notation: A, or A c) is the event that A does not happen. and A' are mutually exclusive & exhaustive.

Example:

• If event A is getting an even number on a die, then the complement of A, written as A' = "getting an odd number.

Mathematically:

P(A) + P(A') = 1

So, P(A') = 1 − P(A)

This identity is frequently applied in order to solve problems more easily by working out the complement of an event.

3.3 Definitions of Probability

Probability has several possible interpretations, depending on where it is applied. The three major approaches are:

• Classical (Theoretical) Probability

• Relative Frequency (Empirical) Probability

• Axiomatic Probability

Each has its own associated applications, limitations and assumptions.

3.3.1 Classical Definition

Definition:

The classical definition (or, theoretical probability) is founded on the premise of all outcomes being equally likely. It's used when every result in the sample space is equally likely to happen.

Formula:

P(E) = (Favourable outcomes) ÷ (Total of all the possible outcomes) Example:

• Flipping an unbiased coin: P(Head) = 1 ÷ 2 = 0.5

• Roller of a fair six-sided die: P(3) = 1 / 6

P(Even number) = 3 ÷ 6 = 0.5

Conditions:

• The results have to be mutually exclusive and equally likely.

• The range of the values must be a finite and enumerable set. Limitations:

• Inapplicable if the outcomes are not eqaully likely.

• Inapplicable to complex or real-life events as there is no symmetry.

3.3.2 Relative Frequency Definition Definition:

The relative frequency definition (also called an empirical probability) defines the probability of an event as: Probability[A] = Number of outcomes for which A happened Total number of possible outcomes. It is estimated by repeating an experiment numerous times and viewing how frequently the event happens.

3 Formula:

P(E) = (Number of times that the event happens) / (Total number of trials) E.g.

If it rained on 30 of the first 100 days:

P(Rain) = 30 ÷ 100 = 0.3

Uses:

• Applied in applied and observational investigations.

• Facilitates in estimating probabilities when theoretical probability is hard to impose.

Advantages:

• Reflects actual conditions.

• Works whether or not the events are all equally likely. Limitations:

• Getting more trials increases accuracy only.

• Not suitable for one-off or once-in-a-lifetime occurrences.

3.3.3 Axiomatic Definition Definition:

The modern and most general definition is that of Andrey Kolmogorov. It is to define what probability means in terms of a list of logical rules (axioms) rather than refer to experiments or equal chances.

In this interpretation, probability is a function (P) that numbers each event in a sample space S, subject to the following three axioms:

Kolmogorov's Axioms:

Non-negativity:

$P(E) \geq 0$ for any event E

Certainty:

$P(S) = 1$ (The probability of the sample space itself is 1)

5 Additivity (for mutually exclusive events):

If A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$

Advantages:

• Can handle infinite sample spaces.

• Generalizes to any type of events, including continuous probability.

• Prepares you for further study of statistical theory. Example Use:

•Axiomatic probability has applications in higher level math, finance, machine learning and more.

3.3 Definitions of Probability

Probability has several possible interpretations, depending on where it is applied. The three major approaches are:

• Classical (Theoretical) Probability

• Relative Frequency (Empirical) Probability

• Axiomatic Probability

Each has its own associated applications, limitations and assumptions.

3.3.1 Classical Definition

Definition:

The classical definition (or, theoretical probability) is founded on the premise of all outcomes being equally likely. It's used when every result in the sample space is equally likely to happen.

Formula:

P(E) = (Favourable outcomes) ÷ (Total of all the possible outcomes) Example:

• Flipping an unbiased coin: P(Head) = 1 ÷ 2 = 0.5

• Roller of a fair six-sided die: P(3) = 1 / 6

P(Even number) = 3 ÷ 6 = 0.5

Conditions:

• The results have to be mutually exclusive and equally likely.

• The range of the values must be a finite and enumerable set. Limitations:

• Inapplicable if the outcomes are not eqaully likely.

• Inapplicable to complex or real-life events as there is no symmetry.

3.3.2 Relative Frequency Definition

Definition:

The relative frequency definition (also called an empirical probability) defines the probability of an event as: Probability[A] = Number of outcomes for which A happened Total number of possible outcomes. It is estimated by repeating an experiment numerous times and viewing how frequently the event happens.

3

Formula:

P(E) = (Number of times that the event happens) / (Total number of trials) E.g.

If it rained on 30 of the first 100 days:

P(Rain) = 30 ÷ 100 = 0.3

Uses:

• Applied in applied and observational investigations.

• Facilitates in estimating probabilities when theoretical probability is hard to impose.

Advantages:

• Reflects actual conditions.

• Works whether or not the events are all equally likely. Limitations:

• Getting more trials increases accuracy only.

• Not suitable for one-off or once-in-a-lifetime occurrences.

3.3.3 Axiomatic Definition

Definition:

The modern and most general definition is that of Andrey Kolmogorov. It is to define what probability means in terms of a list of logical rules (axioms) rather than refer to experiments or equal chances.

In this interpretation, probability is a function (P) that numbers each event in a sample space S, subject to the following three axioms:

Kolmogorov's Axioms:

Non-negativity:

$P(E) \geq 0$ for any event E

Certainty:

$P(S) = 1$ (The probability of the sample space itself is 1)

5 Additivity (for mutually exclusive events):

If A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$

Advantages:

• Can handle infinite sample spaces.

• Generalizes to any type of events, including continuous probability.

• Prepares you for further study of statistical theory. Example Use:

•Axiomatic probability has applications in higher level math, finance, machine learning and more.

Did You Know?

"Did you know that the axiomatic definition of probability is the most universal and mathematically rigorous approach to probability? It doesn't depend on physical experiments or equally likely outcomes. Instead, it uses a set of logical rules called Kolmogorov's axioms, which apply to finite, infinite, and even continuous sample spaces. This is the foundation for most modern probability theories used in fields like machine learning, quantum mechanics, and financial risk modelling."

3.4 Theorems on Probability

In probability, events occur separately, together or in intersection. It is the addition theorem that aids us in determining the probability that either one event or another takes place. This is particularly helpful when events are not mutually exclusive (where they can occur simultaneously).

3.4.1 Addition Theorem

The Sum Rule Theorem offers a method to compute the probability of two events' union, i.e., that at least one of them occurs.

General Formula:

1 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Where:

• $P(A \cup B)$ is the probability that either event A or event B or both occur.

• $P(A \cap B)$ is the probability that A and B happen at the same time. This formula takes care of double counting the fraction that is common to A and B.

3.4.2 Special Case of Addition Theorem for Mutually Exclusive Events

If A and B are mutually exclusive (That is both can't happen at the same time). Then:

$P(A \cap B) = 0$

Hence the addition theorem reduces to:

$P(A \cup B) = P(A) + P(B)$

Example:

• Let A = drawing a red card

• Consider B = drawing a black card

because any one card cannot be red and black, A and B are disjoint.

If P(A) = 26⁄52 and P(B) = 26⁄52, then:

P (A ∪ B) = 26⁄52 + 26⁄52 = 1

It is reasonable to expect this, since the card must be red or black.

3.4.3 General Case of Addition Theorem

But if events A and B are not mutually exclusive, that is they can both happen at the same time, then we must use the more general formula:

P (A ∪ B) = P(A) + P(B) − P (A ∩ B)

Example:

• Let A be the event that a student takes calculus.

• Let B = science Is there anybody that takes Biology but not science?

If:

• P(A) = 0.6

• P(B) = 0.5

• P (A ∩ B) = 0.3

Then:

P (A ∪ B) = 0.6+0.5−0.3 = 0.8

Well, then the probability a student takes either math or science (or both) is 80%.

“Activity Applying the Addition Theorem to Event Probabilities”

Instruction to Student:

Consider a school where 70% of students like playing football, 50% like playing cricket, and 30% like both sports.

1. Use the general addition theorem to calculate the probability that a randomly selected student likes

either football or cricket.

2. Apply the formula:

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

3. After calculating, represent the situation using a Venn diagram, clearly showing the overlapping region.

4. Submit your calculations and the diagram and briefly explain why subtracting the intersection is necessary in this case.

3.4.4 Applications of Addition Theorem

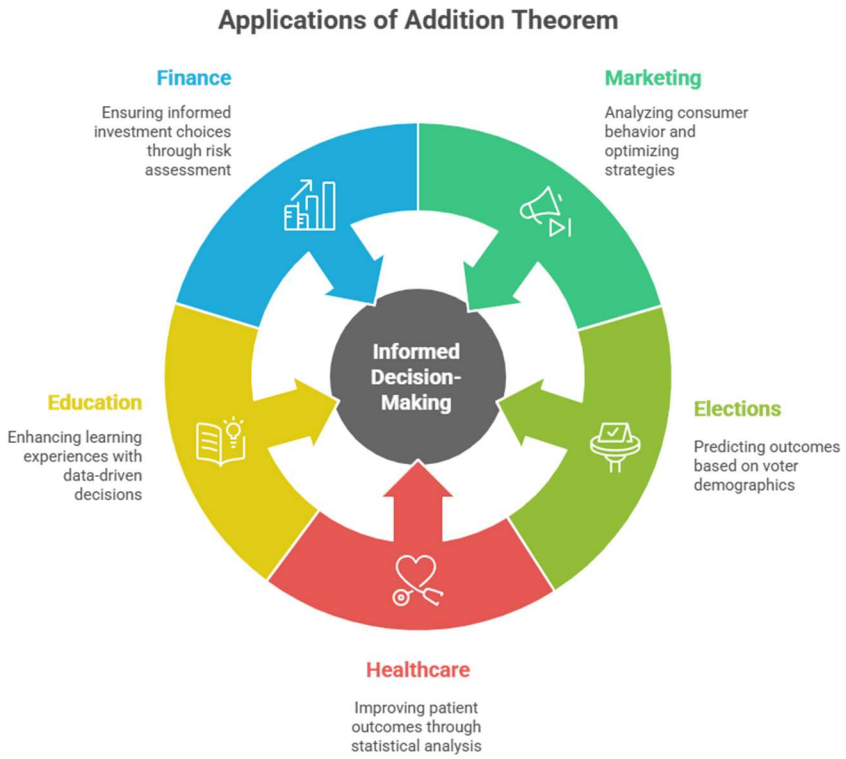


Fig.3.1. Applications of Addition Theorem

The summation rule is a tool that occurs frequently in probability problems as well as the real world.

Common applications include:

Marketing:

The likelihood that a customer purchases item A or B:

Elections:

Chance that a voter supports either of two candidates A, B.

Healthcare:

The chance that a patient exhibits the symptom X or Y.

Education:

How many students take at least one of two electives.

Finance:

Measuring exposure to risk when two events might both occur in the market.

Why is it important?

- Prevents overestimation of probability in cases where events are not mutually exclusive.
- Lays the groundwork for tackling compound probability questions.
- Frequently used in conjunction with Venn diagrams for overlay visualization.

Visual Tip:

In the Venn diagram case, that's the combined area of both circles minus the overlap. The latter is used to give an intuition of P (A ∪ B) for students.

3.5 Conditional Probability

In reality, in many real-life examples the probability that something (anything) will happen depends on whether some other event has happened or not yet occurred. Here's where conditional probability steps in.

3.5.1 Concept of Conditional Probability

Definition:

Conditional probability is the probability that one thing happens, given another situation. We represent it as P(A | B), which stands for: “the probability of event A when event B has occurred.” This concept is useful when:

- Events are not independent
- You have incomplete information about the result

You wish to modify probability after observing new evidence.

Example:

Let's say we have a student that we know has survived the course, who's traveled through mathematics. What is the chance that this student has passed science as well?

This is a type of conditional asked question, where an event (passing math) helps us estimate the other (passing science).

3.5.2 Formula and Examples

Formula for Conditional Probability:

9 If A and B are two events, a2) P(A|B). If A and B are two events and if P(B) ≠ 0, then:

P(A | B) = P(A ∩ B) ÷ P(B)

This means:

The probability of A occurring given that B has occurred =

The probability that both A and B occur divided by the probability of B.

Example 1:

A card is drawn from a deck. Let:

- A = that the card is a king
- B= red.

You have 2 red kings from 26 red cards, so:

P (A ∩ B) = 2⁄52 P(B) = 26⁄52

So:

P (A | B) = (2⁄52) ÷ (26⁄52) = 2⁄26 = 1⁄13.

Interpretation:

Since the card is red, the probability is.

Example 2:

There are 3 red and 2 blue balls in a box. One ball is drawn without replacement, then a second ball is drawn. Let:

- A = first ball is red
- B = second ball is red

We are looking for P (B | A): the probability of the second ball being red, with the consideration that the first one was red. If 1st red ball is not taken out, then there are 2 red and 2 blue balls left.

So,

P (B | A) = 2⁄4 = 0.5

"Activity Exploring Conditional Probability through Classroom Data"

Instruction to Student:

Your class recently conducted a quiz in which 30 students participated. Out of them:

- 18 passed in Science
- 20 passed in Mathematics
- 12 passed in both Science and Mathematics

1. Use this data to calculate:

a) P(Passed Science | Passed Math)

b) P(Passed Math | Passed Science)

2. Apply the conditional probability formula: P(A | B) = P(A ∩ B) ÷ P(B)

3. Write a short reflection (4–5 lines) on how the probability changes when we already know a student passed one subject.

3.5.3 Properties of Conditional Probability

Here are some important properties:

1 Multiplication Rule (Rewrite Joint Probability): P (A. B) = P(B) × P (A | B)

or

P (A ∩ B) = P(A) × P (B | A)

(Choose based on known condition)

1 If A and B are independent:

Then:

P (A | B) = P(A)

and

P (B | A) = P(B)

(Since the other does not influence one).

Conditional Probability of an Event Given Itself: P (A | A) = 1

Zero Conditional Probability:

If events A and B cannot occur simultaneously (mutually exclusive):

P (A | B) = 0

Total Probability from Conditional Events:

The total likelihood of an event taking place, considering it under various conditions (leading up to Bayes' Theorem in later sections).

3.6 Multiplicative Theorem for Independent Events

The theorem of products permits us to calculate the probabilities of several independent events together. It's a useful tool when one event does not affect the other — say, tossing a coin and rolling a die at the same time.

3.6.1 Statement of the Theorem

Definition of Independent Events:

Two events A and B are considered independent if the result of one does not have any impact on the result of other.

Statement of the Multiplicative Theorem:

If A and B are independent, so:

P (A ∩ B) = P(A) × P(B)

So, all we need to do is multiply the probabilities of A and B.

Example 1: Tossing a coin and rolling a die Consider the sample space S of the two step experiment when you toss a coin and then roll a die.

- A = Obtaining Head (P(A) = 0.5)
- B = rolling a 4 (P(B) = 1⁄6

Since these are independent events:

$P(A \cap B) = 0.5 \times 1/6 = 1/12$

3.6.2 Applications

Also, the multiplicative law can be convenient in many applications, particularly for repetitions or simultaneous experiments.

Common Applications:

Business and Finance:

Probability of two unrelated market forces happening, such as rise in oil price and fall in dollar price.

Quality Control in Manufacturing:

o If the probability that one product fails is 0.02, then the probability that 2 products independently chosen fail is:

$0.02 \times 0.02 = 0.0004$

Gaming and Probability Experiments:

o Rolling 2 dice and determining the chance of rolling 6 on both.

Weather Prediction:

• Probability that it rains in Delhi and at the same time in Mumbai (Weather in both the cities being independent).

Medical Studies:

o If the probability of side effect from Drug A is 0.1 and side effect from Drug B = 0.05 and drugs act independently, then probability of both occurring is:

$0.1 \times 0.05 = 0.005$

3.6.3 Extension to More than Two Events

The product principle extends to more than two independent events. Let A1, A2, A3,..., An be n independent events, then.

$P(A_1 \cap A_2 \cap A_3 \cap ... \cap A_n) = P(A_1) \times P(A_2) \times P(A_3) \times ... \times P(A_n)\}$

Example 2:

If you tossed three coins, your odds of getting all heads would be:

$P(\text{Head on coin 1}) = 0.5$

P (Head on coin 2) = 0.5

P (Head on coin 3) = 0.5 So:

P(All heads) = 0.5 × 0.5 × 0.5 = 0.125

Important Note:

And this theorem only works when they are independent. If there is dependence between two events (one event affects the other) then we have to use conditional probability.

Did You Know?

“Did you know that when you're tossing multiple coins or rolling multiple dice, the probability of a specific combined outcome is calculated using the extended multiplication rule for independent events? For example, the probability of getting three heads in a row is not 0.5—it’s 0.5 × 0.5 × 0.5 = 0.125. This principle is used in genetics, cryptography, and even digital communication systems to calculate the chances of exact sequences occurring.”

3.7 Bayes’ Theorem

Bayes’ Theorem is a central concept in probability that secures us the ability to update probabilities as new information is received. It is particularly useful for contexts in which the result depends on success of other conditions (e.g., traffic lights).

3.7.1 Statement and Explanation of Bayes’ Theorem

Statement:

Posting: Your name If $B_1 \cap B_2 \cap \cdots \cap B_n = \{\}$, B_1, B_2,..., B_n are mutually exclusive and exhaustive events and A is any event which has now occurred then the

probability of B_i occurs given A has happened as:

$P(B_i \mid A) = [P(A \mid B_i) \times P(B_i)] / \{ \Sigma[P(A \mid B_j)] \times [P(B_j)] \}$

Where:

• P(Bi | A) is the posterior probability of B_i given that event A has occurred (updated probability of B_i after A occurs),

•P(\(B_i\)) is the probability of B_i before A occurs

• $P(A \mid B_i)$ is the probability ($P(A|B_i)$: The probability of A occurring if B_i were true)

- The denominator is the sum of the product of the probability of A for all B events possible

Why Use Bayes' Theorem?

- It lets us revise our beliefs in the light of new evidence.
- And when we require reverse conditional probabilities.
- It answers: "A has happened, but how likely is B?"

Simple Example:

A diagnostic test for a rare disease that affects 1\% of the population.

- P(Disease) = 0.01
- P (No disease) = 0.99
- P (Test positive | Disease) = 0.95 (probability of test being positive given diseased)
- P (Test positive | No disease) = 0.05 (false positive rate)

13 What is the chance that a subject with a positive test truly has the disease?

Let:

- D = person has disease
- D' = person does not have disease
- T = person test positive We want P(D | T) =?

8 Apply Bayes' Theorem:

P(D | T) = [P(D) × P(T | D)] / [P(D) × P(T | D) + P(D') × P(T | D')]

Substitute values:

= [0.01× 0.95]/ [0.01× 0.95+ 0.99 × 0.05]

= 0.0095 ÷ (0.0095 + 0.0495)

= 0.0095 ÷ 0.059

≈ 0.161 or 16.1%

21 So even with a "positive" result, that person only has a 16.1 percent chance of actually having the disease (or condition), given its rarity and the rate at which this test falsely suggests miracles.

3.7.2 Applications of Bayes' Theorem in Real-world Problems

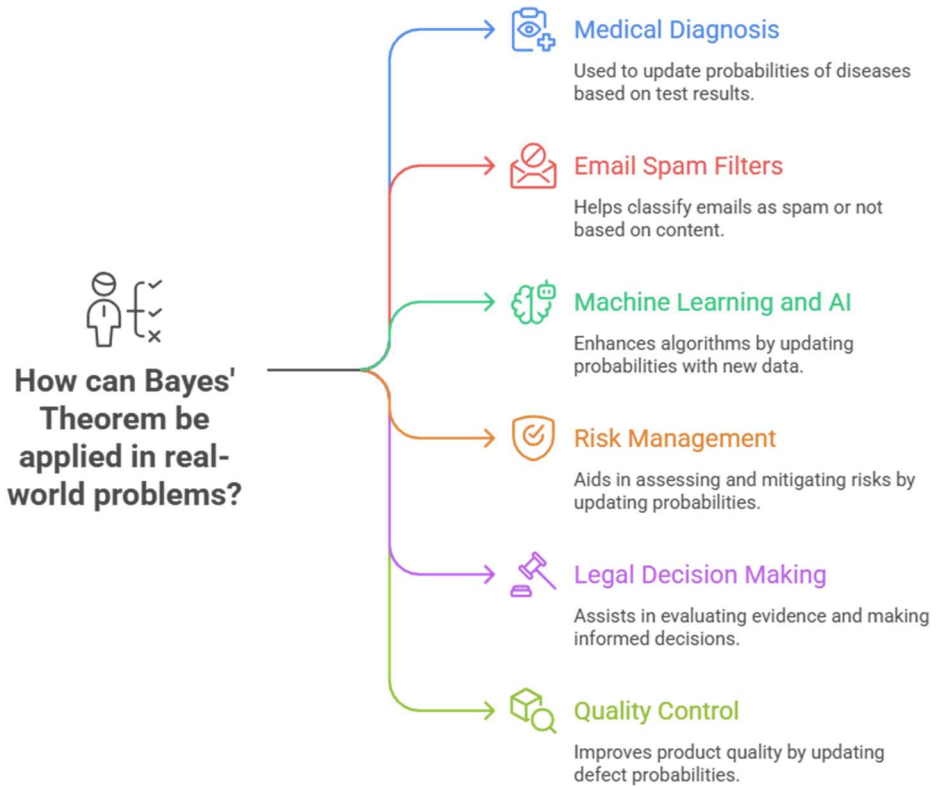


Fig.3.2. Applications of Bayes' Theorem in Real-world Problems

Bayes' Theorem routinely appears in many fields for evidence-based prediction and updating of prior beliefs. Here are some real-world applications:

Medical Diagnosis

Employed to determine the likelihood of disease on having a positive or negative test. Extremely important when diseases are uncommon and tests aren't perfect.

Email Spam Filters

A Bayesian spam filter calculates the probability that an email is spam by checking if certain words or phrases are there.

Machine Learning and AI

Naive Bayes classifiers are based on Bayes' Theorem:

- Sentiment analysis
- Document classification

• Image recognition

Risk Management

Applied in finance and insurance by considering the impact of new market data on the probability of risk events (e.g., crashes, defaults).

Legal Decision Making

Applied, for example, in court cases when evidence is measured of a subject being guilty or innocent (such as DNA which updates the prior probability).

Quality Control

In manufacturing, Bayes' Theorem is used for calculating the probability of defects in a product given a testing result and history of batches.

Knowledge Check 1

Choose the correct option:

1. Which definition of probability is most appropriate when outcomes are not equally likely, and data is collected from actual observations?

A) Classical probability

B) Relative frequency probability

C) Axiomatic probability

D) Experimental error model

15

2. If two events A and B are mutually exclusive, what is the value of $P(A \cap B)$?

A) $P(A) + P(B)$

B) $P(A) \times P(B)$

C) 1

D) 0

3. If the probability that a student passes in Maths is 0.7 and in English is 0.6, and the probability that the student passes in both subjects is 0.4, then what is the probability that the student passes in at least one subject?

A) 1.3

B) 0.7

C) 0.9

D) 1.0

4. Which of the following is the correct expression for conditional probability?

A) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

B) $P(A \mid B) = P(A \cap B) \div P(B)$

C) $P(A \cap B) = P(A) + P(B)$

D) $P(B \mid A) = P(B) \div P(A \cup B)$

5. Bayes' Theorem is mainly used to:

A) Measure the number of outcomes in a sample space

B) Predict future outcomes in completely random experiments

C) Calculate probability without any known prior information

D) Update a prior probability based on new evidence

3.8 Summary

- This module was to expose students to Basic Probability – the mathematics of measuring uncertainty, and estimating probabilities of events. It started by describing what probability is, why it's important and other applications in disciplines from statistics to healthcare, business etc.
- Key Terms: Mutually exclusive, exhaustive, complementary, independent and dependent events
- were introduced for relating events. The three primary theories of probability: classical, relative frequency and axiomatic offered different ways to compute and explain probabilities in different situations.
- The request focused on the complementary function and its use to determine the probability of the union or one or more events. Then, it added conditional probability to help refine estimates of probability when the cookbook had incomplete or previous information. The product rule of independent events and Bayes' Theorem were what allowed us to handle joint and revised probabilities, which is why probability was so useful in prediction and decision making.
- Using formulas and examples, students are prepared to address real-world-based problems that involve uncertainty employing the probabilistic methods of analysis.

3.9 Key Terms

1. Experiment – Action or process that results in an outcome.
2. Sample Space (S): The collection of all the possible outcomes of an experiment.
3. Event (E) - A subset of the sample space, containing one or multiple (or zero) outcomes of interest.
4. Disjoint (Mutually Exclusive) Events - Events that cannot occur simultaneously.
5. Complete Events - A collection of events that includes all possible outcomes.
6. Complementary Events - Two events for which one is the "not" of the other.
7. Mutually Exclusive Events - Two events such that occurrence or non-occurrence of one does not influence the other.
8. Conditional Probability - The likelihood of an event given that another occurred.
9. Classical Probability - Probability that comes from equally likely outcomes.
10. Relative frequency - Probability as determined by experimentation, data or observations.
11. Axiomatic Probability - Probability as defined by certain formal rules (axioms).
12. Theorem of Addition - Rule to determine the probability of union of events.

3.10 Descriptive Questions

1. Introduce probability and discuss its use in making decisions relevant to everyday life.
2. What are the mutually exclusive and exhaustive event? Give examples.
3. Describe the distinction between classical, relative frequency, and axiomatic definitions of probability.
4. State and prove the addition theorem of probability.
5. What is conditional probability and provide an example of it in real life?
6. What is the rule to multiply independent events? How is it applied?
7. Derive Bayes' Theorem for probability and explain its role in updating beliefs.
8. A card is drawn from a deck. What chances are there of it being a red card // or // a king /?
9. There are 3 red balls and 2 blue balls in a box. One ball is drawn and nothing is done as to its replacement, what's the probability that even this second ball will also be red?
10. Distinguish between independent and dependent with an example.

3.11 References

1. Gupta, S.P. (2014). Statistical Methods. Sultan Chand & Sons.

2. Goon, A.M., Gupta, M.K., & Dasgupta, B. (2013). Fundamentals of Statistics. World Press.

3. Hogg, R.V. & Tanis, E.A. (2010). Probability and Statistical Inference. Pearson Education.

4. NCERT (2021). Mathematics: Probability (Class XI).

5. Khan Academy. (2023). Probability and Statistics. [https://www.khanacademy.org]

6. Statistics How To. (2023). Bayes' Theorem Explained. [https://www.statisticshowto.com]

Answers to Knowledge Check

Knowledge Check 1

1. B) Relative frequency probability

2. D) 0

3. C) 0.9 → $P(A \cup B) = 0.7 + 0.6 - 0.4$

4. B) $P(A \mid B) = P(A \cap B) \div P(B)$

5. D) Update a prior probability based on new evidence

3.12 Case Study

Bayes at Work: Diagnosing with Data

Background:

Mumbai: A private hospital in the city began a screening test for a rare disease that affects 2 out of every 1000. The test is 95 percent accurate, in that it correctly detects 95 percent of people who have the disease (true positives) and mistakenly accuses 5 percent of those who do not (false positives). A patient, Arjun, tested positive. The doctor asked: "Given that Arjun has the disease, what are the chances that he actually has it?"

Problem:

The test result itself is not enough because the disease is so rare. The chances of false positives are huge, because the base rate of the disease is s low. Now, how likely is it that Arjun really has the disease in question, given that he tested positive?

Solution Using Bayes' Theorem:

Let:

- D = disease → P(D) = 0.002
- ¬D = does not have the disease → P(¬D) = 0.998
- $T^+ \mid D$ = test positive having disease → $P(T^+ \mid D) = 0.95$

• $T^+ \mid \neg D$ = test positive with no disease → $P(T^+ \mid \neg D) = 0.05$

Apply Bayes' Theorem:

$P(D \mid T^+) = [P(D) \times P(T^+ \mid D)] \div [P(D) \times P(T^+ \mid D) + P(\neg D) \times P(T^+ \mid \neg D)]$

$= [0.002 \times 0.95] / [0.002 \times 0.95 + 0.998 \times 0.05]$

$= 0.0019 \div (0.0019 + 0.0499)$

$\approx 0.0019 \div 0.0518$

≈ 0.0367 or 3.67%

Interpretation:

But since the disease is so rare and some false positives will occur, the probability that Arjun actually has the disease is only 3.67 percent1167 — barely higher than if he hadn't been tested at all.

Outcome:

The doctor does not want to send the patient home frightened and he encourages additional confirmatory testing. The hospital further evaluates the screening program to enhance decision-making through data-driven methods.

Statistics for Business Unit 4 V3.docx

Statistics for Business_BBA_2

Statistics for Business_BBA_2

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:127443390

Submission Date

Feb 3, 2026, 3:08 PM GMT+5:30

Download Date

Feb 3, 2026, 3:17 PM GMT+5:30

File Name

Statistics for Business Unit 4 V3.docx

File Size

160.5 KB

21 Pages

4,676 Words

23,646 Characters

6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 15 words)

Match Groups

16 Not Cited or Quoted 6%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
1% Publications
4% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **16** Not Cited or Quoted 6%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
1% Publications
4% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.coursehero.com 1%
2	Submitted works	University of Birmingham on 2009-12-09 <1%
3	Internet	kahedu.edu.in <1%
4	Submitted works	Iceland Consortium on 2025-12-17 <1%
5	Internet	mpra.ub.uni-muenchen.de <1%
6	Submitted works	Brisbane State High School on 2021-08-13 <1%
7	Submitted works	Fakultat fur Maschinenwesen der Technischen Universitat Munchen on 2015-06-26 <1%
8	Submitted works	Manipal University Jaipur Online on 2025-06-24 <1%
9	Submitted works	University of Adelaide on 2023-07-09 <1%
10	Submitted works	Westminster Academy School on 2016-07-10 <1%

11 Internet

dokumen.pub <1%

12 Submitted works

American Intercontinental University Online on 2006-12-16 <1%

13 Internet

exo7.emath.fr <1%

14 Submitted works

Hellenic Open University on 2023-04-26 <1%

Unit 4: Random Variables

Learning Objectives

1. Understand the concept of a random variable, distinguish between discrete and continuous random variables, and explain their roles in modeling real-world outcomes.

2. Define and interpret the Probability Mass Function (PMF) of a discrete random variable, and use it to calculate probabilities associated with specific values or ranges.

3. Understand and construct a Cumulative Distribution Function (CDF) for a discrete random variable, and explain how it reflects the accumulation of probabilities.

4. Apply rules of probability to solve problems involving discrete random variables, including expected value (mean), variance, and standard deviation.

5. Explore the concept of two-dimensional (joint) discrete random variables, and calculate joint, marginal, and conditional probabilities from a joint distribution table.

6. Compare and differentiate between PMF and CDF, and analyse how each is used in understanding distribution and cumulative behavior of discrete data.

7. Use discrete probability models to solve real-life problems in fields such as business, engineering, health sciences, and social sciences, making informed decisions under uncertainty.

Content

4.0 Introductory Caselet

"Simran's Sales Forecast: Making Sense of Daily Orders"

Background:

Simran, a young entrepreneur who owns and operates an online handmade crafts store, observed a lot of variation in her daily orders. On some days, she received no orders; on others, more than she could handle alone. She was looking to make better use of her inventory and time, so she appealed to probability for insights.

Discrete random variables were explained to her by a mentor. They combined, to come up with X as: Number of orders sim said she gets orders per day. Over the course of one month, Simran took note of her number of orders and worked out the PMF for these. She found that the most common number of orders was 3 per day.

She created CDF from this PMF help her calculate probability of having 2 orders or less which can be really helpful for her to be plan out low volume days. She also carried out a follow-up analysis, and she recorded the number of return customers (Y) on those very days. She built a joint probability table for (X, Y) and used marginal and conditional distributions to make sense of the effect of repeat customer patterns on daily sales.

With the help of probability tools such as PMF, CDF and joint distributions, Simran can now:

• Predict inventory needs,

• Identify peak order days,

• And put loyal customers at the front of your promotions line.

What used to be a shot in the dark became data-driven planning, just because she learned to represent her sales as a discrete random variable.

Critical Thinking Question:

If you were Simran, how would you use a joint probability distribution to make better marketing decisions? Provide an example of two variables you would monitor and describe your analysis.

4.1 Introduction

In probability and statistics we are frequently faced with the unknown and uncertain. To make sense of these results, what is needed is a way of systematically analyzing them: for this we have the notion of random variable. It provides a link between the results of an experiment and numbers that can be mathematically analysed.

4.1.1 Concept of Random Variables in Probability Theory

We call as a random variable to any numerical result of the realization of a random experiment. It's assigning a real number to every possible outcome in the sample space.

There are two categories of random variables:

- Discrete Random Variable: Assume countable values (i.e., 0, 1, 2, …).
- Continuous Random Variable: assumes values over a continuous domain (i.e., any real number between 0 and 1).

Definition:

A random variable is a function that assigns to each outcome of a random experiment a real number.

Example:

- Tossing a coin →suppose X = 1 if Head, X=0 if Tail
- A die is rolled →X = the face that lands up (1, 2,…,6)

The randomness comes in through the experiment, but once an outcome is observed, we know what value of the random variable was realized.

4.1.2 Importance of Random Variables in Statistics

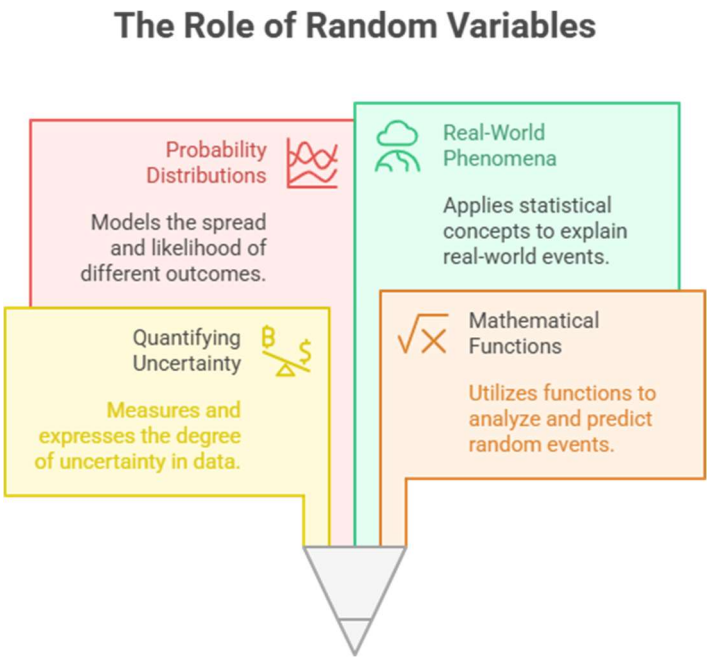


Fig.4.1. Importance of Random Variables in Statistics

Random variables are important entities in statistical inference since:

• Allow us to quantify uncertainty.

• Enable us to apply mathematical functions (e.g., mean, variance, probability functions) for the purpose of studying randomness.

• Are the heart of probability distributions, which represent how outcomes spread out.

• Assist in describing natural phenomena in terms of probabilities and numerical quantities.
Applications in Statistics:

• Predicting population means (using expected value).

• Measuring risk in finance (variance and standard deviation).

• Engineering and medical modelling (failure rates, waiting times, etc.).

4.1.3 Examples of Random Variables in Real Life

Some examples of real-life scenarios where random variables are employed:

Scenario	Random Variable (X)	Type
Number of goals scored in a football match	X = Number of goals	Discrete
Time taken for a customer to be served	X = Time in minutes	Continuous
Number of defective items in a batch	X = Count of defective items	Discrete
Daily rainfall in a city	X = Rainfall in mm	Continuous
Result of a multiple-choice quiz	X = Number of correct answers	Discrete
Number of heads in 3 coin tosses	X = 0, 1, 2, or 3	Discrete

Key Idea:

To each one of the random variables, it is associated an experiment or situation in which the result is uncertain but can be assigned equilibria and attach numbers to those outcomes and analyze their statistical behaviour.

4.2 Random Variables

In probability and statistics, we have the important concept of a random variable which can represent outcomes of a physical event as numbers. These quantifiers of uncertainty serve as the foundation for probability distributions, expectation and statistical inference.

4.2.1 Definition and Classification (Discrete and Continuous)

Definition:

A randoim variable is a real function that assigns to every outcome of a random experience a real number.

The reason it is called “random” is because its value is determined by chance. Let:

- X be a random variable
- S be the sample space

Then X: S → $\mathbb{R}$, so that X is a function that takes outcomes and returns real numbers.

Classification of Random Variables:

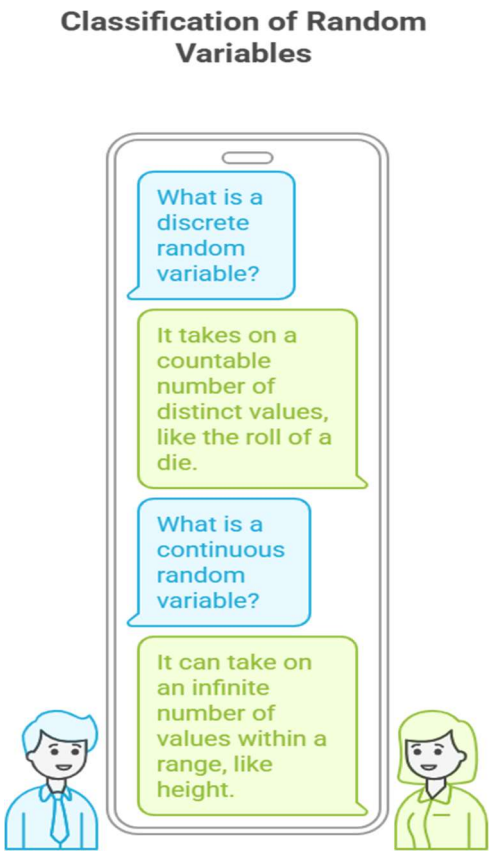


Fig.4.2. Classification of Random Variables

Random variables Types of random variable There are two types of random variables :

A. Discrete Random Variable

• Finite or denumerable number of distinct values.

• Often associated with counting.

• Examples:

o Heads count in 3 coin flips → X Є {0, 1, 2, 3}

o Cars in a parking lot

o Score on a 10-mark test

B. Continuous Random Variable

• Assumes infinite, uncountable number of values in a range.

• Usually associated with measuring.

• Examples:

o Finish time in race ($X \in [0, \infty)$)

o Heights of the students in a classroom

o The temperature in a city at 12 p.m.

4.2.2 Properties of Random Variables

Certain mathematical properties of random variables are valuable in statistical and probability modelling:

Probability Assignment

If X is a discrete random variable, then for each value x_i there is some probability $P(X = x_i)$ such that:

• $0 \leq P(X = x_i) \leq 1$

• $\Sigma P(X=x_i) = 1$ (summation over all possible values)

Expected Value (Mean)

The mean value of a random variable is what you expect to get if the experiment were repeated many times and averaged. For discrete X:

$E(X) = \Sigma x_i \times P(X = x_i)$

It is a kind of weighted average of possible values.

Variance and Standard Deviation

• The variance (Var(X)) gives a measure of this dispersion from the mean.

$Var(X) = E[(X - \mu)^2] = \Sigma (x_i - \mu)^2 \times P(X = x_i)$

• Now, Standard Deviation is the square root of the variation:

o $\sigma = \sqrt{Var(X)}$

These inform us of how reliable or variable the realizations of the random process are.

Linearity of Expectation

10 If X and Y are two randoms variables, a, b constants 12.

$E(aX + bY) = aE(X) + bE(Y)$

This always holds, regardless of whether X and Y are dependent.

Indicator Random Variables

A particular kind of random variable that can only assume the value 0 and 1. Symbol used to indicate of the occurrence of an event.

Example:

Let A = event student passes.

Define:

X = 1 (if student success in the test); X = 0 (otherwise)

Then E(X) = P (student passes)

4.3 Probability Mass Function (PMF)

Probability Mass Function (PMF): PMF is used to describe the distribution of a discrete random variable. That is, it informs us of the likelihood that a random variable assumes a given value.

4.3.1 Definition of PMF

14 The Probability Mass Function (PMF) of a discrete random variable X is the function that yields the probability that X assumes each of its possible values.

Formally, if X is a discrete random variable, then the PMF of X is:

6 $P(X = x_i) = p(x_i)$

Where:

• x_i is the value that X can assume

• $p(x_i)$ – probability that X takes on the value x_i

• $p(x_i) \geq 0$, for all i

• The total of all $p(x_i)$ on the sample space is 1

4.3.2 Properties of PMF

Properties of a proper PMF: For an accurate probability mass function, Below are the characteristics which it must satisfy.

Non-negative: $p(x_i) \geq 0$ for all values x_i.

Normalization (Total Probability = 1):

The probabilities for any possible value must add up to 1:

$\Sigma\, p(x_i) = 1$

Defined Only for Discrete Values:

The PMF is only defined for the actual possible values of this discrete random variable. For all other values, the probability is 0.

Probability of an Exact Value:

PMFs are for discrete variables, so you can calculate directly:

$P(X = x) = p(x)$

Did You Know?

"Did you know that the Probability Mass Function (PMF) can never assign a probability greater than 1 or less than 0 to any event? This rule applies strictly because PMF represents actual probabilities, not just relative frequencies. Also, the sum of all PMF values for a discrete random variable must always equal 1, even if the variable has many possible values—this ensures that all possible outcomes are accounted for."

4.3.3 Examples of PMFs Example 1: (Tossing a Fair Coin)

Let X = the number of heads when one coin is thrown. Possible values:

$X \in \{0, 1\}$ PMF:

- $P(X = 0) = 0.5$
- $P(X = 1) = 0.5$

This satisfies:

- $p(x) \geq 0$
- $\Sigma p(x) = 1$

Example 2: Rolling a Fair Die

Let X = result when a fair die is rolled. X: number of nodes visited $X \in \{1, 2, 3, 4, 5, 6\}$ Suppose we let f := n − r. Note that only the value(s) in the last line is (are) possible little m's > 'a'.

PMF:

2

- $P(X = 1) = 1/6$
- $P(X = 2) = 1/6$
- $P(X = 3) = 1/6$

• P(X = 4) = 1⁄6

• P(X = 5) = 1⁄6

• P(X = 6) = 1⁄6

Check:

• All probabilities are ≥ 0

• Sum = 6 × (1⁄6) = 1

Example 3: Custom PMF

Let X = the number of calls received at a customer service desk per hour. Assume probabilities are as follows:

x (Calls) 0 1 2 3

p(x) 0.1 0.3 0.4 0.2

Check:

• All values are ≥ 0

• 0.1 + 0.3 + 0.4 + 0.2 = 1

Thus, this is a valid PMF.

Visual Representation (Optional in class):

PMFs can be plotted as bar charts such that:

• X-axis = values of the variable happening in a random manner

Y-axis = corresponding probabilities p(x)

“Activity: Constructing a PMF and CDF Table”

Title: Daily Orders – PMF and CDF Practice

Instruction to Student:

Your team manages a small e-commerce store. Over 10 days, you recorded the following number of orders:

{2, 3, 3, 1, 4, 2, 2, 3, 1, 2}

1. List all unique values of the random variable X (number of orders).

2. Compute the frequency and relative frequency (PMF) for each value.

3. Using the PMF, calculate the Cumulative Distribution Function (CDF) for each value of X.

4. Present your final table with three columns: X, P(X), and F(X).

5. Submit a short reflection: What is the probability of receiving 3 or fewer orders on a given day?

4.4 Cumulative Distribution Function (CDF)

Time-to-event results A Cumulative Distribution Function (CDF) specifies the cumulative probability that a random variable assumes a value less than or equal to an observation. This can be very efficient in summarizing the behaviour of a random variable over a wide range instead of at one point.

4.4.1 Concept and Definition

The CDF of a random variable X is the probability that it takes on some value less than or equal to x.

Mathematically:

$F(x) = P(X \leq x)$

This implies that F(x) is the sum of the probabilities for all X less than or equal to x.

4.4.2 CDF for Discrete Random Variables

For a discrete RV, the CDF value is obtained by summing up the PMF values at and below that point. If X is a discrete random variable with possible values $x_1, x_2, x_3, \ldots, x_n$ and PMF p(x), then:

$F(x) = \Sigma\, p(x_i)$ for each $x_i \leq x$

The CDF is:

- Stepwise increasing
- Always between 0 and 1
- Right continuous (value jump at each x defined)

Example:

Let X represent the number of heads from two coin tosses. Allowed values: X = {0, 1, 2}

x P(X = x)

1 0 0.25

1 0.50

2 0.25

Then the CDF F(x) is:

• F(0) = P(X ≤ 0) = 0.25

• F(1) = P(X ≤ 1) = 0.25 + 0.50 = 75 % Some Python Features and Keywords: • A x is used to store an intermediate value of a calculation – When using \(a \) to represent the unknown, its value up to that point in the expression can be replaced in the python code by using \(a(x)\).

F (2) =P(X ≤ 2) = 0.25 + 0.50 + 0.25=1.0

For values between these points:

4 • F(x) = 0 when x < 0

• F(x) = 0.25 when 0 ≤ x < 1

• F(x) = 0.75 when 1 ≤ x < 2

• F(x) = 1 when x ≥ 2

4.4.3 Relationship between PMF and CDF

There is a close relationship between the PMF and CDF:

• The PMF predicts the chance of any one value:

p(x) = P(X = x)

• The CDF represents the cumulative probability below a value:

F(x) = P(X ≤ x)

From PMF to CDF: You add up the PMF values. From CDF to PMF, you simple subtract:

p(x) = F(x) − F(x−1)

With this you can build one from the other when necessary.

4.4.4 Graphical Representation of Discrete Distributions

In the case of a discrete CDF this graph is the step function. Key characteristics of the graph:

• X-axis = realization of the random variable

• Y-axis = cumulative probabilities F(x)

• The graph goes higher at every X.

• The height of a step = probability from the PMF

• It is piecewise-flat at the x-values (both on steps and jumps between points).

Example:

Using the coin toss example:

• Mark the points (0, 0.25), (1, 0.75) and (2, 1.0)

• Connect horizontal lines between the steps This graph makes it easy to visualise:

• How much probability has accumulated until these values

• Where the values are most likely

• Whether range is cantered or asymmetrical

Did You Know?

"Did you know that a Cumulative Distribution Function (CDF) graph for a discrete variable looks like a step function? Unlike continuous functions which flow smoothly, CDF graphs for discrete variables "jump" at each possible value of the variable. The size of each step corresponds to the probability assigned by the PMF, and the final step always reaches 1 (or 100%)"

4.5 Two-Dimensional Discrete Random Variables

Frequently, for populations studied in actual applications of statistics, there will be two or more random variables on interest. For instance in a family, the number of boys and girls; in business, sales and marketing costs; on a survey, level of education and employment.

When we are dealing with two random variables together, we have 2-dimensional (bivariate) discrete distributions.

Let X and Y be a pair of discrete random variable. The collective behaviour is analysed with a joint probability distribution.

4.5.1 Joint Probability Mass Function (Joint PMF)

The Joint PMF denotes the probability that $X = x_i$ and $Y = y_j$ happen at the same time. It is denoted as:

$P(X = x_i, Y = y_j) = p(x_i, y_j)$

The joint PMF is given by:

• $p(x_i, y_j) \geq 0$ for all i, j

• The sum of all joint probabilities is 1:

$\Sigma \Sigma\ p(x_i, y_j) = 1$

These values are commonly organized in a joint probability table with each row and column corresponding to a particular value of X and Y.

4.5.2 Marginal Distributions

The marginal distribution of a variable is its separate distribution obtained from the joint by summing over the other variable.

• Marginal PMF of X:

$P(X = x_i) = \Sigma\ p(x_i, y_j)$ (sum over all y_j)

• Marginal PMF of Y:

$P(Y = y_j) = \Sigma\ p(x_i, y_j)$ (sum over all x_i)

These graph the behavior of each of the two variables independent of the other Variable.

"Activity: Joint and Marginal Distribution Analysis"

Title: Tracking Sales: Bread vs Milk

Instruction to Student:

You are given the following joint probability table showing the number of customers who bought bread

(X) and milk (Y) on a given day:

3

X\Y	0	1	2
0	0.05	0.10	0.05
1	0.10	0.20	0.10
2	0.05	0.15	0.20

1. Compute the marginal distribution of X and Y.

2. What is the probability that a customer bought at least 1 unit of either product?

3. Calculate P(X = 2 | Y = 2) — the probability that a customer bought 2 breads given they bought 2 milks.

4. Submit your calculations along with a brief interpretation of what this data suggests about buying behavior.

4.5.3 Conditional Distributions

The conditional distribution of one variable with respect to the other is a posterior probability when it represents the probability after partial information.

- $P(X = x_i \mid Y = y_j) = p(x_i, y_j)/P(Y = y_j)$

- $P(Y = y_j \mid X = x_i) = p(x_i, y_j) \div P(X = x_i)$

It enables us to observe the behaviour of a variable for a given value of the other. This is particularly helpful if events are not independent.

4.5.4 Examples on Marginal and Conditional Distributions

Example: Joint Probability Table

Let X = Number of books sold in a day (0, 1) Let Y = Number of customer complaints (0, 1)

	Y = 0	Y = 1	Total
X = 0	0.1	0.2	0.3
X = 1	0.4	0.3	0.7

Marginal PMFs:

- P(X = 0) = 0.1 + 0.2 = 0.3
- P(X = 1) = 0.4 + 0.3 = 0.7
- P(Y = 0) = 0.1 + 0.4 = 0.5
- P(Y = 1) = 0.2 + 0.3 = 0.5

Conditional PMFs:

- P(X = 0 | Y = 1) = 0.2/0.5 = 0.4
- P(X = 1 | Y = 1) = 0.3 ÷ 0.5 = 0.6

- $P(Y = 0 \mid X = 1) = 0.4 \div 0.7 \approx 0.571$
- $P(Y = 1 \mid X = 1) = 0.3 / 0.7 \approx 0.429$

Knowledge Check 1

Choose the correct option:

1. What is the essential condition for a valid Probability Mass Function (PMF)?

A) It must have at least 10 values

B) All values must be negative

C) The sum of all probabilities must be 1

D) The cumulative probability must be 0

2. Which of the following statements is true about a Cumulative Distribution Function (CDF)?

A) CDF decreases as the value of the random variable increases

B) CDF always remains constant

C) CDF is always equal to PMF

D) CDF is a non-decreasing function that reaches 1

1

3. A discrete random variable X has the following PMF: P(0) = 0.2, P(1) = 0.5, P(2) = 0.3

What is the $P(X \leq 1)$?

A) 0.5

B) 0.7

C) 0.9

D) 1.0

4. In a joint probability distribution, the marginal probability of variable X is obtained by:

A) Dividing joint probabilities by totals

B) Multiplying all joint probabilities

C) Subtracting Y's probabilities from X's

D) Summing over all values of Y for each value of X

5. What does the value P(X = 2 | Y = 3) represent?

A) Probability of both X and Y being 2

B) Probability that Y equals 3, regardless of X

C) Probability that X is 2, given that Y is 3

D) Total probability of X and Y being less than 5

4.6 Summary

- This unit addressed the main idea in discrete probability distribution which is essential for random behavior mathematician statistics. It started with random variables and how they were together divided into discrite/continuous types.
- So after that we discussed about Probability Mass Function (PMF) which depicts the probability for each value of a discrete random variable. Following on this, the Cumulative Distribution Function (CDF) originated as a cumulative count of the probability up to a given value.
- ϖ The unit then expanded into bivariate discrete random variables, namely, how to create and describe joint, marginal and conditional distributions. Students were given the power to investigate the relationship between two variables simultaneously, through worked-through examples and step-by-step tables.
- These fundamental concepts enable statisticians, data scientists, and practitioners to understand uncertainty, pattern identification and data-driven decision making in real-world scenarios.

4.7 Key Terms

1. Random Variable (RV) – A variable that assumes numerical values based on the outcome of a random experiment.
2. Discrete Random Variable - A random variable that has a finite number or countable number of distinct possible values.
3. 2,6-PMF (Probability Mass Function) – A function that provides the probability for each discrete value of random variable.
4. CDF (Cumulative Distribution Function) - A function that provides the total probability up to a value x,1 i.e., P (X ≤ x).
5. Joint Probability Mass Function (PMF) - Their probability distribution together with anther discrete random variable.
6. Marginal Distribution- The probability distribution of one variable, derived from the joint probabilities over the other variable by summation.

7. Conditional Distribution – Probability of one variable given another has a particular value.

4.8 Descriptive Questions

1. What is a random variable that takes on only discrete values? Give an example.
2. Describe the PMF clearly and state its properties.
3. Show how we obtain the CDF by using the PMF.
4. Distinguish between continuous and discrete random variables.
5. What is the method to build a joint probability distribution table?
6. What is the distinction between marginal and conditional distributions?
7. Describe the connection between PMF and CDF with an example.
8. A die is rolled once. Determine an appropriate random variable and its PMF.
9. Students were surveyed in the classroom for time dedicated to homework and number of subjects. How could you represent this with two dimensional random variables?
10. (Since) i.e. why is the sum off all the PMF 1?

4.9 References

1. Hogg, R.V. & Tanis, E.A. (2015). Probability and Statistical Inference. Pearson.

2. Sheldon Ross (2014). Introduction to Probability Models. Academic Press.

3. Walpole, R.E., Myers, R.H., et al. (2012). Probability and Statistics for Engineers and Scientists. Pearson.

4. Goon, A.M., Gupta, M.K., & Dasgupta, B. (2013). Fundamentals of Statistics, Vol I.

5. NCERT (2021). Statistics Textbook, Class XI and XII

6. Khan Academy. (2023). Random Variables and Probability Distributions. [https://www.khanacademy.org]

Answers to Knowledge Check

Knowledge Check 1

1. C) The sum of all probabilities must be 1

2. D) CDF is a non-decreasing function that reaches 1

3. C) 0.9 (i.e., P(0) + P(1) = 0.2 + 0.5)

4. D) Summing over all values of Y for each value of X

5. C) Probability that X is 2, given that Y is 3

4.10 Case Study

"Discrete Probability Distributions in Inventory Management: How to model your inventory and plan for the future".

Introduction

And in the intensely competitive world of retail, where thin margins are endemic, companies depend on data to make their operations run optimally. A major managerial problem of the retail industry is the stock management or how much to produce and sale per day in order not run out-of-stock and avoid overstock. Such a decision is frequently made depending on sales data which vary under demand ambiguity. Managers can make better estimates and save money by knowing discrete probability distributions.

In this caselet, we look at how a retail store manager leverages random variables and probability mass functions (PMF) including cumulative distribution functions (CDF) to make the best plan for stocking. It shows the application of two-dimensional discrete variables and how joint and conditional probabilities can help companies make smarter decisions about inventory.

Background

Anjali, operator of one mid-sized convenience store in Nagpur, noticed that daily sales of some perishable items such as bread and milk were not consistent. Some days left the store out of product by the evening; other days brought waste. This imbalance was discouraging sales and profitability.

Anjali decided to tackle this by keeping track of the count of bread packets that were getting sold daily (random variable X) for over a month. She labeled the results and computed the PMF and observed that some values (e.g., 4 or 5 packets) were more likely. She subsequently built a CDF that she can use to compute the probability of her selling ‘k’ or fewer packets on any given day.

As she developed her analysis, Anjali introduced a second random variable Y to signify the number of bottles of milk sold. She constructed a joint probability table for X and Y to investigate their relationship, as well as marginal and conditional probabilities. That, she says, let her answer questions like:

“Four packets of bread were sold, what is the probability that milk sales are higher?”

Issue 1: Inaccurate Demand Forecasting of BreadCause As per the article I read Topp, E.N.But Then Again, Maybe Amazon's "Make-You-Buy" Grocery Store Wasn't Such a Good Idea After All they pointed out that with its new Amazon Go grocery store, has had great difficulty sourcing enough bread to make sandwiches.

Unpredictable demand each day caused Anjali to over-order (wastage) or under-order (lost revenue) on bread.

Solution:

Anjali let X = number of packets of bread sold in one day, a random variable. With her recorded results, she created a PMF and then moved on to creating a CDF that showed cumulative probabilities, such as $P(X \leq 3)$ or $P(X \geq 5)$. That has helped her more efficiently determine reorder points and safety stock levels.

MCQ 1:

What is this cumulative distribution function (CDF) in stock planning Anjali wants to try?

A) Total profit on bread sales is to be determined

B) To determine the likelihood of making accurate demand predictions

C) To obtain the cumulative probability of sold packet up to a certain number of packets.

D) To measure average shelf life

Answer: C) Needed to calculate the cumulative probability of selling from 1 upto a certain number of packets.

Issue 2 :Unable to Map the Bread and Milk Sales

Anjali was wondering if there existed an association between sales of bread and milk but she had no well-defined mechanism to approach the problem.

Solution:

Now she has another thing, Y = number of bottles of milk sold, and a new joint probability distribution table for (X, Y). She then calculated:

• Marginal probabilities (e.g., $P(X = 3)$)

• The probability of events based on other events (e.g., $P(Y = 2 \mid X = 3)$)

This study revealed that high purchases of bread were also collocated with moderate purchases of milk, which supported

coordinated inventory planning. MCQ 2:

Why do we need Joint probability distribution?

A) To calculate only average sales

B) To estimate the sales of other merchandise

C) To examine associations between two variables such as sales of bread and milk

D) To randomly allocate inventory

Answer: C) To study the correlation of two variables such as bread sales and milk sales

3 Problem Statement: Cannot Plan due to No Prescient from Daily Sales Data

Prior to her analysis, Anjali had been treating daily sales as mere numbers were taken isolatedly and not part of a statistical pattern.

Solution:

She also considered sales as independent random variables, allowing her to use statistical functions like PMF (probability mass function), mean and variance. She also plotted these distributions to understand what the most probable outcomes were and how spread out they were. This enabled her to move from reactive to predictive inventory management.

MCQ 3:

What do we gain by considering sales counts as discrete random variables?

A) It helps randomize supply

B) It alleviates requirement of data

C) It allows for probabilistic analysis of outcomes

D) It simplifies purchase orders

Answer: C) It allows for analytic estimation of probability of event

Conclusion

Using discrete probability distributions, Anjali went from relying on guesswork to forecast her inventory… 12. Through an analysis of PMF, CDF and joint distributions she was in a position to improve her decision on the inventory stocking levels, save money on waste, and serve demand more efficiently. This example demonstrates how simple statistical tools can assist in enhancing operations and efficiency, and guide smarter business decisions in retail management.

Statistics for Business Unit 5 V3.docx

Statistics for Business_BBA_2

Statistics for Business_BBA_2

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:127447831

Submission Date

Feb 3, 2026, 4:13 PM GMT+5:30

Download Date

Feb 3, 2026, 4:17 PM GMT+5:30

File Name

Statistics for Business Unit 5 V3.docx

File Size

213.6 KB

24 Pages

5,574 Words

28,523 Characters

7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 15 words)

Match Groups

22 Not Cited or Quoted 7%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
0% Publications
5% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **22** Not Cited or Quoted 7%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
0% Publications
5% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.coursehero.com <1%
2	Submitted works	Manipal University Jaipur Online on 2025-07-11 <1%
3	Submitted works	Manipal University Jaipur Online on 2025-05-12 <1%
4	Submitted works	Manipal University Jaipur Online on 2025-06-01 <1%
5	Internet	11.realinfo.tv <1%
6	Submitted works	Manipal University Jaipur Online on 2025-07-01 <1%
7	Internet	shikshanation.com <1%
8	Submitted works	University of Florida on 2025-04-19 <1%
9	Submitted works	Manipal University Jaipur Online on 2025-06-09 <1%
10	Internet	mafiadoc.com <1%

11 Internet

solr.bccampus.ca:8001 <1%

12 Internet

youthforpakistan.org <1%

13 Submitted works

Manipal University Jaipur Online on 2025-07-02 <1%

14 Submitted works

University of Hertfordshire on 2025-12-06 <1%

15 Internet

www.gucdoe.in <1%

16 Submitted works

Manipal University Jaipur Online on 2025-06-25 <1%

17 Submitted works

Wayne State University on 2024-10-19 <1%

18 Internet

brightideas.houstontx.gov <1%

19 Internet

krmangalam.edu.in <1%

Unit 5: Measures of Central Tendency

Learning Objectives

1. Define and distinguish between different measures of average, including arithmetic mean, median, and mode.
2. Calculate the arithmetic mean for both ungrouped and grouped data sets.
3. Determine the median and mode for various data types and interpret their significance in real-world contexts.
4. Compare and contrast the characteristics, strengths, and limitations of arithmetic mean, median, and mode.
5. Apply empirical formulas to estimate the mode and understand the relationship between the three central tendency measures.
6. Interpret and analyze statistical data sets using appropriate measures of central tendency for decision- making.
7. Evaluate case studies to identify the most suitable average measure and justify its application with evidence and reasoning.

Content

5.0 Introductory Caselet

"Rahul's Rental Records: Discovering the Typical Customer"

Background:

Rahul is an executive in a city-based mid-range car rental company. Over the last year, he had observed an important variability in the availability of days for renting a car depending on seasons, client segment profiles and ongoing promotions. To strike the right balance between price, fleet availability and customer targeting, Rahul knew he would first have to determine what "a standard" rental looked like.

He started keeping track of the number of days his borrowers were renting. Once I had data on 150 rentals, Rahul noticed a lot of variance—some customers only rented the item for one day; others used it for up to ten days or more. A data consultant recommended studying the central tendency measures to get a better grasp on rental patterns. Together, they calculated:

• The mean, which represented the average rent that applied to all customers.

14 • The median, which is the middle value of a set.

• The mode, which is the rental length that appears with the highest frequency.

Rahul noted that although the average rental time was 4.8 days, median rental time was 4 days and mode of rental time = 3 days. This demonstrated that while a certain fraction of the customers rented for long time instances, most of them rented images just for short instances. Interpreting this data, Rahul achieved:

• Price plans by the most frequent periods of rental,

• Build predictive models on availability of vehicles,

• And create promotions that target specific segments of your customers.

Knowledge and application of central tendency measures had helped Rahul to shift his decision-making from reactive to data-oriented, resulting in increased operational efficiency and customer satisfaction.

Critical Thinking Question:

If you were Rahul and noticed that there was a significant difference between the mean and median time until rental, what might this discrepancy say about the distribution of their data? How would that affect your business strategy: pricing, promotion?

5.1 Measures of Average

In statistics, an average is the value that measures or summarises data. What this does is give us, in a single number, something that describes the centre of the data set or its typical value. Averages allow us to organize reams of data by allowing us to turn a bunch of numbers into one number that serves as the best representative for the entire group.

Averages This video is a tutorial on mathematical averages, which are used in statistics. The most popular averages are:

• Mean (Most Commonly Used, often simply referred to as "the mean")

11 • Median

• Mode

They are called measures of central tendency as they tell us where the center of the data is.

The concept of average is related to other statistical tools, including when and how to use them, and is beneficial for various disciplines: simple in daily life but also complex as economics or social science. For example, the average temperature offers a prediction of weather, the average income reflects an assessment of economic conditions and the average score shows us how well students are doing.

5.1.1 Concept and Significance of Central Tendency

Concept of Central Tendency

The central value of a set of data is an idea about estimating the central or middle value around which other values are distributed. It represents a typical value in a group of numbers.

For instance, if we are to take the ages of 50 students in a class, it is not practical to tabulate all the ages each time. How do we state "average age" is 19 years? This single number provides some good perspective of what the aave age of the students is."

1 There are 3 basic measures of central tendency:

Mean: The average of the values in a set.

Median: The middle value of a data set when it has been arranged in numerical order.

Mode: The number that occurs most often in the data.

Each of these quantities provides a different measure for what may be called "typical" in the set of data.

Importance of Central Tendency

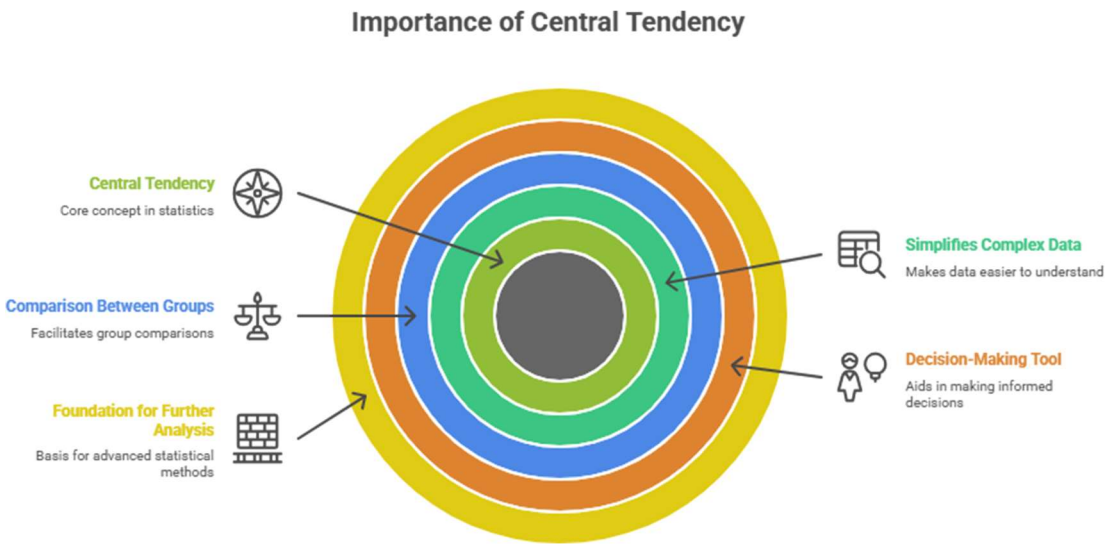


Fig.5.1. Importance of Central Tendency

Distills Complex Data: Instead of hundreds, or thousands, of individual numbers, you get one average that sums up the entire data set.

Comparison Between Groups: Measures of central tendancy which make it easier to compare one group with another. For example, comparing the average incomes of two cities to find out which is wealthier.

Business Tool: An average helps business owners or managers make a decision, namely what to expect in terms of an average sales number (for example) or profits, and how to please the most amount of customers if they don't have data about different preferences.

Basis for Additional Statistical Analysis: Measures such as variance and standard deviation are based on the mean. Most statistical models are based on measures of central tendancy.

Universal Application: from a government looking at population statistics, a teacher computing average scores to a doctor examining the average recovery periods – all apply central tendency.

Pattern Recognition: You can use central tendency to spot trends or patterns in the context of time, like the average annual rainfall over a region—information that can impact anything from agriculture to planning.

Central tendency is a basic concept to consider when you are working with data. It is the first part which takes up numerical data and attempts to interpret.

5.1.2 Characteristics of a Good Measure of Central Tendency

To be a useful measure of central tendency, it should satisfy the following important requirements which makes its representation as a data efficient and successful. These characteristics include:

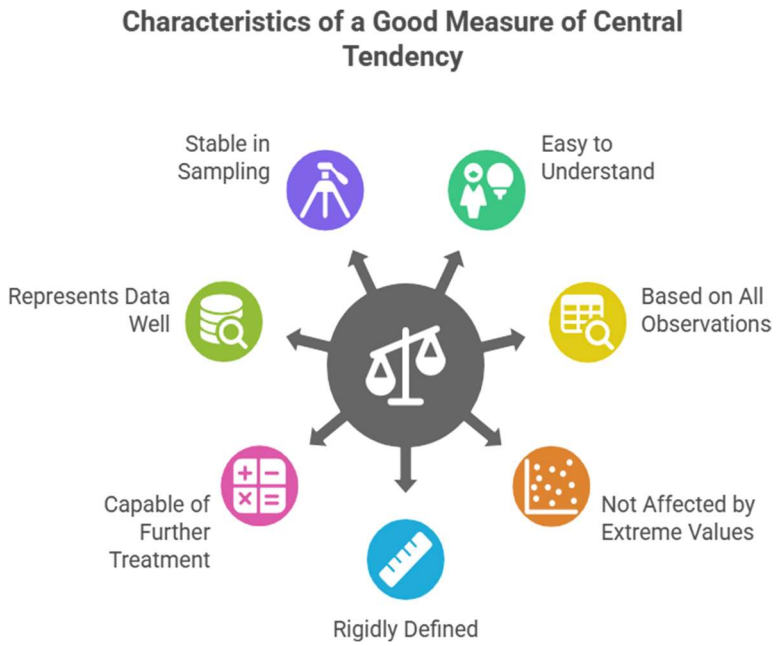


Fig.5.2. Characteristics of a Good Measure of Central Tendency

Easy to Understand and Calculate

The calculation should be easy and the outcome intuitive. As per real-life applications, this offers while being a more realistic value.

Based on All Observations

A good measure of dispersion should consider each value in the data set. This guarantees that no part of the data is overlooked, so that the measurement becomes more accurate.

Little Impacted by Outliers

If there are some extremely large or small values that exist in a dataset (these are sometimes referred to as outliers), then an ideal measure is rarely overly influenced by them. There are, for instance, techniques that render a measure more robust to outliers than others, such as the median versus the mean.

Rigidly Defined

The sense in which the measure is being calculated needs to be specified explicitly or at least very clearly, both for a measurement carried out by different individuals on equally should yield same final value.

Capable of Further Mathematical Treatment

A reportable result should be obtained, which may be subjected to further statistical analysis. For example, the average may be used in formulas to compute standard deviation, correlation and regression.

They Also Ought to Fit the Data Set Nicely

The selected criterion must be intermediate between the extremes of the data set and should conform to the general trend or distribution profile of that case.

Should Be Stable in Sampling

Value should not fluctuate much between samples taken from the same population, so good central tendency measure.

5.1.3 Types of Averages (Mean, Median, Mode, etc.)

Averages come in all different types, and the type of average to be used depends on what kind of data you have. The mean, median, and mode are the three most widely used. Other such means are the geometric mean and harmonic mean.

Arithmetic Mean (or simply "Mean")

The average is the one we use the most. It is the mean of all the values.

15 Formula (for ungrouped data):

Mean = (Σx) ÷ n Where:

• Σx = sum of all values

• n = number of values

Example:

Just say you have 10,20,30,40 and 50 in the list.

So, Mean = (10 + 20 + 30 + 40 + 50) ÷ 5 = 150 ÷ 5 = 30

Note: The mean is sensitive to extremely large or small values (outliers).

18 Median

The median is the value that separates the data set into two equal halves, when it is ordered in increasing or decreasing order. It splits the data in half.

How to find the median (ungrouped Data) – 4 step process:

5

• Arrange the data in order.

• If there is an odd number of values (n), then the median is the middle value.

• If n is even, the median is the average of the two middle numbers.

Example:

Data: 15, 20, 25, 30, 35

Median = 25 (middle of sorted list) Data: 12, 16, 20, 24

Median = (16 + 20) ÷ 2 = 36 ÷ 2 = 18

Note: Medians are not influenced by extremes.

Mode

Mode is the number that occurs most often in a set of data.

Example:

Data: 5, 8, 8, 10, 12

Mode = 8 (appears twice)

• A data set may have:

o None (if all values are distinct)

o One mode (unimodal)

o Two modes (bimodal)

o More than two modes (multimodal)

Note: Mode is very helpful in the case of categorical data and not get's affected by extreme values as well.

Geometric Mean

Geometric mean is applied when you will like to obtain average of scores that are products and percentages.

Formula (for n values):

GM = $\sqrt[n]{x_1 \times x_2 \times x_3 \times \ldots\ldots x_n}$ (n is the number of numbers)

Example:

For values: 2, 4, 8

GM = $\sqrt[3]{(2 \times 4 \times 8)} = \sqrt[3]{64} = 4$

Applications: Compound Interest, Exponential growth of population and Exponential growth rates.

Harmonic Mean

Harmonic mean can be applied whenever data are in rates (speed, efficiency etc).

Formula (for n values):

$HM = n / (1/x_1 + 1/x_2 + 1/x_3 + \ldots + 1/x_n)$

Example:

eg 2,4,6 Hm = 3 / (1/2 + 1/4 + 1/6)

= 3 ÷ (0.5 + 0.25 + 0.1667)

= 3 ÷ 0.9167 ≈ 3.27

Examples: Average velocity, price per unit, % yield.

5.2 Arithmetic Mean

8 Meaning Arithmetic mean (also called the mean) is one of the most important measures of central tendency. It is the total of all values in a given set, divided by the number of values. It provides a rough sense of the "typical" or "average" value.

5.2.1 Simple Arithmetic Mean (Ungrouped and Grouped Data)

A. Ungrouped Data

For ungrouped data, each value stands alone not organized by ranges.

Formula:

Mean ($\bar{x}$) = (Σx) ÷ n Where:

• Σx = addition of all observations

• n = number of observations

Example:

If marks of five students in a class are: 50, 60, 70, 80 and 90

Mean = (50 + 60 + 70 + 80 + 90) ÷ 5 = 350 ÷ 5 =7o

B. Grouped Data

The grouped data is given in frequency distribution. There are two primary ways to calculate the mean of grouped data:

Direct Method Formula:

Mean ($\bar{x}$) = (Σ f × x) ÷Σf Where:

• f = frequency of each class

• x = mid-point of each class

• Σf = total frequency

Steps:

Determine the midpoint (x) for each class: (Lower Limit + Upper Limit) ÷ 2

Multiply each midpoint by it corresponding Frequency(7 x 3) (10 x 4)+ Do the calculations =42 40 Then add up all these figures.

Now take the sum of all of these: Σf × x

Divide by the overall frequency (Σf)

Example:

Class Interval	Frequency (f)	Mid-point (x)	f × x
10 – 20	3	15	45
20 – 30	5	25	125
30 – 40	7	35	245
40 – 50	5	45	225
Total	20		640

Mean = 640 ÷ 20 = 32

“Activity: Finding the Class Average”

You are provided with the following frequency distribution, representing the marks scored by students in a statistics examination. The class intervals are 0–10, 10–20, 20–30, 30–40, and 40–50, with corresponding frequencies of 2, 5, 8, 12, and 3. Your task is to calculate the mid-point for each class interval and then use the direct method to compute the arithmetic mean of the marks. After calculating the mean, write a short interpretation explaining what this average tells you about the performance of the class. Does the mean accurately reflect a “typical” score, or do you observe any signs of skewed distribution? Include both your calculation steps and a brief reflection in your submission.

5.2.2 Weighted Arithmetic Mean

The weighted mean is used in many cases where items in a data set have varying levels of importance, such as the means analysied from different reaction times, shock intensities etc.

Formula:

Weighted Average ($\bar{x}$) = sum of the w × x) /sum of the w and, where :

- x = value
- w = weight of each value
- Σw = total of weights

Example:

6 A student obtained 80 in theory (weight = 3) and 90 in practical (weight =2). \- Weighted Mean = {{80×3+90×2}} {3 + 2 } \)= {(240+180)}5 = 420(5) =84

Such an approach is helpful in grading systems, economic indicators and scenarios where some elements are more important.

Did You Know?

“The weighted mean is used in stock market indices like the S&P 500, where each company’s weight in the index is based on its market capitalization. This means larger companies (like Apple or Microsoft) have a bigger impact on the index value than smaller ones. Unlike the simple mean, the weighted mean reflects the real-world significance of each component.”

5.2.3 Applications and Properties of Arithmetic Mean

Properties:

The Sum of the Deviations From the Mean is Zero

$\Sigma(x - \bar{x}) = 0$

So the positive and negative values on both sides of a mean cancel one another.

Mean is Sensitive to All Values

By changing any of my data values, I change the mean.

Average lies between the smallest and biggest values

An arithmetic mean is always in the middle remainder of data.

Mean is Unique

There is one mean for a data set.

We Can Add the Mean of Several Groups

The mean of several groups can be combined as follows:

$\bar{x}$ (pooled) $= (n_1\bar{x}_1 + n_2\bar{x}_2) \div (n_1 + n_2)$

Applications:

- Business: Average, sales and revenue with cost analysis
- Schooling: Average scores, grade sheet analysis
- Economics: Income, prices and growth average
- Engineering: Measurement and quality control
- Everyday Life: Medium spend, medium speed

5.2.4 Merits and Demerits of Arithmetic Mean

Merits:

Simple to Calculate and Understand

The formula is simple and involves some basic arithmetic.

Uses All Observations

All of the values in a data set influence the mean.

Mathematically Useful

The average is a key variable in numerous statistical calculations prior and following the analysis.

Consistent and Rigidly Defined

There's no interpretation, the outcome is unique.

Suitable for Further Statistical Operations

It is amenable to algebraic manipulation, as in the case of standard deviation and correlation.

Limitations:

Affected by Extreme Values

The mean can be distorted by extremely high or low values.

Not Always a Realistic Value

The mean can be a value that isn’t even in the dataset (e.g., family size of 4.2 members).

Not Suitable for Qualitative Data

You can't take an average on categories like gender or eye color.

May Mislead if Data Is Skewed

The mean is misleading when we have skewed distribution.

Requires Numerical Data

Not applicable to non-numeric or ordinal data without transformative process into numerical value.

5.3 Positional Averages: Median and Mode

Positional averages are averages, in the statistical sense, involving the location and not the arithmetic of numbers in a distribution. Unlike the average, they are not distorted by extreme values. The most commonly used positional averages are the median and mode, and other measures associated with them such as quartiles, deciles, and percentiles.

5.3.1 Median for Ungrouped and Grouped Data

A. Median for Ungrouped Data

Median is the value obtained by arranging all the data in ascending or descending order and then selecting which falls in the middle of the list.

Steps:

2 Arrange the data in order.

Find the number of observations (n).

o If n is odd:

Median = value at position (n + 1) / 2

o If n is even:

Median = mean of the numbers at positions n ÷ 2 and (n ÷ 2 + 1)

Example (odd number of values):

Data: 25, 30, 35, 40, 45

n = 5 → Median = value at (5 + 1) ÷ 2 position = 3rd position → Median = 35

13 Example (even number of values):

Data: 20, 30, 40, 50

n = 4 → Median = (30 + 40) / 2 = 35

B. Median for Grouped Data

4 The median of continuous grouped data is determined using:

Formula:

Median = L + [(N/2? F) / f] h Where:

• L = lower limit of the median class

• N = total frequency

• F = frequency less than median class

• f = frequency of the median class

• h = class width

Steps:

Calculate N = Σf

Find N ÷ 2

Determine the class in which the cumulative frequency is greater than N ÷ 2 → this is the median class

Apply the formula

Example:

7

Class Interval	Frequency (f)	Cumulative Frequency
0 – 10	4	4
10 – 20	6	10
20 – 30	10	20
30 – 40	5	25
40 – 50	5	30

• N = 30, N ÷ 2 = 15

• Median class = 20 – 30

• L= 20,* F = 10* Surf. waves = L/2f (surf.) The only minimalistic value for the object depicted is surf.

Median = 20 + {15 - 10}{10}]× 10 = 20+ (5/10) × 10 =20 +5=25

5.3.2 Quartiles, Deciles, and Percentiles

These are percentile measures break down the data into equal parts:

A. Quartiles

Data are divided into four equal parts using quartiles.

• Q_1(First Quartile) = the value below which 25% of the data falls.

• Q_2 (2nd Quartile) = median 50% • Q_3 (Third Quartile).

•Q_3 (Third Quartile) = the value below which 75% of the data lies

Formula of Q_1, Q_3 for grouped data:

$Q_1 = L + [(N/4 - F)/f] \times h$ $Q_3 = L + [(3N/4 - F)/f] \times h$

B. Deciles

Deciles create 10 equally-sized cuts of the data.

• $D_1, D_2, ..., D_9$

• D_5 = Median

Formula:

$D_k = L + [(kN \div 10 - F) \div f] \times h$

Where k = 1 to 9

C. Percentiles

Percentiles are measures which divide the data set into 100 equal parts.

• $P_1, P_2, ..., P_{99}$

• P_{50} = Median

Formula:

$P_k = L + [(kN \div 100 - F) \div f] \times h$ (3)

Where k = 1 to 99

Such measures are often found in test score data, income distributions and rankings.

Did You Know?

“In competitive exams like SAT, GRE, or national aptitude tests, your score is often reported in percentiles, not just marks. A percentile score of 90 means you scored better than 90% of the test-takers—a positional statistic used to rank performance without disclosing raw scores.”

5.3.3 Median for Ungrouped and Grouped Data

1

A. Mode for Ungrouped Data

The mode is the value that occurs most frequently in a data set.

Example:

Data: 12, 14, 14, 18, 19

Mode = 14 (since its occured twice) The possible values are:

• Raw (no mode; all values occur once)

• One mode (unimodal)

• Two modes (bimodal)

• More than two modes (multimodal)

B. Mode for Grouped Data

For continuous data, the mode is estimated using the following formula:

3

Formula:

Mode = $L + [(f_1 - f_0) \div (2f_1 - f_0 - f_2)] \times h$ Where:

• L = lower limit of the modal class

• f_1 = frequency of the mode class

• f_0 = frequency of the class before modal class

• f_2 = frequency of the class after the modal class

• h = class width

Steps:

Determine which class has greatest frequency → this is the mode class.

Plug values into the formula

Example:

Class Interval	Frequency
10 – 20	5
20 – 30	12
30 – 40	18 ← Modal class

40 – 50	10
50 – 60	5

9 with L = 30, f_1=18, f_0=12, f_2=10 and h=10.

Mode = 30 + [(18 − 12) ÷ (2×18 − 12 − 10)] × 10

= 30 + (6 ÷ 14) × 10

= 30 + 4.29 ≈ 34.29

5.3.4 Comparison between Mean, Median, and Mode

Basis of Comparison	**Mean**	**Median**	**Mode**
Definition	Sum of values ÷ Number of values	Middle value of ordered data	Most frequently occurring value
Use of All Data	Uses all values	Uses position only	Uses only frequent values
Affected by Outliers	Yes	No	No
Type of Data	Quantitative	Quantitative	Both quantitative and categorical
Mathematical Use	Useful in further analysis	Limited mathematical use	Not used in advanced calculations
Stability	Stable across samples	Less stable	Can be unstable
Real-World Example	Average salary	Median household income	Most common shoe size

Key Insight:

• Mean is**best for symmetric distributions without outliers.

• If the data is skewed with a few outliers, then median should be preferred.

• Mode is very good for finding the most frequent value in your data, such as for categorical data.

"Activity: Choosing the Best Average"

Select a dataset containing at least ten numerical values from any real-life scenario of your choice—such as daily expenses, step counts from a fitness tracker, monthly rainfall, or scores from a recent cricket or football series. Calculate the mean, median, and mode for this data. Carefully observe how similar or different these three values are. Based on your results, analyze whether the dataset is symmetrical, positively skewed, or negatively skewed. In a paragraph, explain which measure of central tendency best represents your data and justify your reasoning. Make sure your answer reflects a clear comparison among the three measures and demonstrates understanding of when each is most appropriate.

5.4 Empirical Analysis of Central Tendency

The assessment framework for central tendency addresses how the mean, median and mode are related to each other in different types of distributions. It also discusses which measure is suitable in what context, and demonstrates how these measures are applied in the broad fields of business and economics.

5.4.1 Relationship among Mean, Median, and Mode

Mean = Median, and Mode in a symmetrical distribution In a perfectly symmetrical (normal) distribution the mean is the same as the median and that is also same as mode.

But for the case of nonsymmetric (skewed) distribution, their relationship is distorted as:

• For a right (or positive) skewed distribution: Mean > Median > Mode

•Negative Skewed Distribution (tail on the left): Mean < Median < Mode

Knowledge of the direction of skewness aids in selecting the best measure of central tendency. The median tends to be more i.e. "better" if the data is not symmetric or if the data has outliers that are several times farther away from other observation than the rest of this observations in a dataset.

5.4.2 Karl Pearson's Empirical Formula

When the mode or median is lacking, or hard to be determined, Karl Pearson developed an empirical formula that can be used to estimate one based on the relationships established.

A. Empirical Relationship Formula:

Mode ≈ 3 × Median − 2 × Mean

This formula is the rare case for when one has a moderately skewed distribution and while you do not know the exact mode, but you know both median and mean.

B. Rearranged Forms:

• Median ≈ (Mode + 2 Mean) ÷ 3

• Average ≈ (3 × Median − Mode) ÷ 2

These are approximations, not exact values; they should be used only when the distribution is far from being highly skewed.

Example:

If Mean = 60, Median = 55,

Then Mode ≈ 3 × 55 − 2 × 60 = 165 − 120 = 45

This means the data is positively skewed because the mean is greater than the mode.

Did You Know?

“Karl Pearson’s empirical formula—Mode ≈ 3 × Median − 2 × Mean—was developed before the widespread use of computers. It provided a quick mental estimate of skewness in distribution and was often used in early social and economic research to interpret census or survey data with limited tools.”

5.4.3 Situational Use of Different Measures

The appropriate measure of central tendency will depend on the type of data and its application. Here are some scenarios and the best actions:

Situation	Preferred Measure	Reason
Data with extreme values (outliers)	Median	Not affected by outliers
Categorical data (e.g., favorite brand)	Mode	Only measure that applies to qualitative data
Symmetrical distribution	Mean	Uses all data and supports further calculations
Skewed income or wealth data	Median	Provides a more realistic picture
Most common product size or choice	Mode	Shows the highest frequency

Financial or scientific calculations	Mean	Required for mathematical analysis

5.4.4 Practical Applications in Business and Economics

Measures of central tendency are frequently used in the businesses and economics fields for statistical analysis, prognostication, trends planning and decision-making.

A. Applications of Mean

• Marketing: Average customer spend is useful for determining pricing strategy.

• Finance: Generic returns on investment are taken into account for risk monitoring.

• Production: Average production per worker can be used as a measure of efficiency.

• Sales: Monthly average sales, which can help to establish and monitor goals.

B. Applications of Median

• Statistics of Income: The median income is more indicative than the mean when raised to the power one does not get a better sense, statistically.

• Real Estate: Median price of a home is used to represent the typical price of a home.

• Policy Making: Knowledge of the typical consumption levels of families may be useful for policy makers.

C. Applications of Mode

• Inventory management – Mode allows you to determine the bestselling product size or colour.

• Fashion Industry: Is used to calculate what the most frequent clothing size sold is.

• Health: Mode can display commonly diagnosed diseases at a hospital.

Each measure serves different purposes. Knowing when to use them and in what format is vital for interpreting your data.

Knowledge Check 1

Choose the correct option:

1 1. Which of the following measures is most affected by extreme values (outliers)?

A) Median

B) Mode

C) Mean

D) None of the above

2. In a positively skewed distribution, the correct order of central tendency measures is:

A) Mean < Median < Mode

B) Mode < Median < Mean

C) Mean = Median = Mode

D) Median < Mode < Mean

3. The arithmetic mean of 10 numbers is 45. If one number is removed and the new mean becomes 40, what is the value of the number removed?

A) 45

B) 40

C) 95

D) 100

4. Which of the following is the correct formula to find the mode in grouped data?

A) Mode = $L + [(N/2 - F) \div f] \times h$

B) Mode = $L + [(f_1 - f_o) \div (2f_1 - f_o - f_2)] \times h$

C) Mode = $L + [(3 \times \text{Median} - 2 \times \text{Mean})]$

D) Mode = $L + [(F - f) \div h] \times N$

5. Which of the following best describes the median?

A) The value that occurs most frequently

B) The average of all values

C) The middle value when data is ordered

D) The sum of frequencies divided by the number of classes

5.5 Summary

- This module covered central tendency, a basic concept of descriptive statistics that refers to the use of one value (out of many) to summarize the information contained in a dataset.

- Average measurements present a view of the central tendency of a data set, including mean (average), median and mode.
- The mean, simply put, is the sum of all values divided by the number of values and is useful in interpreting symmetrical distributions and fits into further statistical considerations.
- Relational averages (median, mode, geomean) are based on the position or frequency and are more appropriate to be used if we have uneven distributions of data non-normally distributed or qualitative variables.
- The three measurements relationship can also represent the distribution distortion. The estimates one using the others and helps us to break free of this circular reasoning.
- Application areas include business, economy, healthcare and social sciences, where the choice of a proper measure is crucial for informed decision-making based on data.

5.6 Key Terms

1. Central Tendency - A statistical value which is representative of a set of data
2. The "Average" - The total of whatever you're studying divided by the number of examples rces.
3. Median - The middle value of data when written in order.
4. Mode – The number that appears most often in a list of numbers
5. Grouped Data: When data is arranged in class intervals alongside their corresponding frequencies.
6. Weighted Mean - Mean where values are given a weight representing their importance or how frequently they occur
7. Skewness - A statistic that measures the asymmetry of a distribution of data.
8. Modal Class - The class containing the maximum frequency
9. Empirical formula: (Approximate relationship) Mode ≈ 3 × Median − 2 × Mean
10. Quartiles -The four equal parts in which the data is divided.
11. Percentiles – Values that partition the data into 100 equal distinctions

5.7 Descriptive Questions

1. What is central tendency and discuss its significance in statistical analysis?
2. Explain the process of finding the arithmetical mean for ungrouped data.
3. Distinguish between the arithmetic mean and the median giving appropriate examples.
4. Describe the formula and procedure required to find out mode for grouped data.
5. When should I use median versus mean for reporting data?

6. Express the empirical relationship among mean, median, and mode as a statement of Karl Pearson.
7. Write short notes on:
    - Quartiles
    - Weighted Mean
    - Modal Class
8. What is the impact of skewness on measure of central tendency values (mean, median and mode)?
9. Illustrate one example from your experience where mode is the best measure of central tendency.
10. Why are we happy to treat the arithmetic average as a number?

5.8 References

1. Gupta, S. C. (2014). Fundamentals of Statistics. Himalaya Publishing House.
2. Levin, R. I., & Rubin, D. S. (2012). Statistics for Management. Pearson Education.
3. Sharma, J. K. (2018). Business Statistics. Vikas Publishing House.
4. Spiegel, M. R., & Stephens, L. J. (2018). Schaum's Outline of Statistics. McGraw Hill.
5. UGC e-Pathshala. (n.d.). Descriptive Statistics Modules. Retrieved from https://epgp.inflibnet.ac.in/
6. Government of India. (n.d.). National Statistical Handbook. Ministry of Statistics and Programme Implementation.

19

Answers to Knowledge Check

Knowledge Check 1

1. C) Mean
2. B) Mode < Median < Mean
3. C) 95
4. B) Mode = $L + [(f_1 - f_0) \div (2f_1 - f_0 - f_2)] \times h$
5. C) The middle value when data is ordered

5.9 Case Study

"El papel de la centralización en la optimización da estrategias comerciales al detalle"

Introduction

In the rapid environment of high-street retail, you need to be making data-led decisions simply to survive. Central tendency is one of the fundamental tools employed by retail analysts and managers to transform the customer, sales, and pricing data that they collect into useful information. Averages, such as mean, median and mode enable businesses to recognize normal values and patterns in data which can drive strategic planning and resource efficiency.

This caselet discusses how a national retail chain used central tendency concepts to improve store operations and marketing. It illustrates real-world problems like deceptive averages from outliers, the demand for positional and order-based averages on skewed data sets as well as the significance of selecting correct statistical measure based on business inference.

Background

Case Example: Retailer A retail chain named RetailMart with 200 national outlets noticed huge variety in customer buying behaviors by regions. Some urban stores were able to achieve a very high monthly sales average, but rural ones were trailing them by far. Its corporate headquarters' leadership used arithmetic average to determine how each store was performing in the early days. But this resulted in unachievable targets for many branch managers who did not have big metro stores.

A more detailed breakdown found the average was distorted by a handful of high-performing outlets. To counter this, the data analytics team started to use median and mode value analysis alongside mean. The monthly median sales gave a better idea of what an "average" store accomplished, and the footfall mode helped managers know when they could expect common traffic rates — or how frequently to staff.

While RetailMart's decision to use all three central tendency measures enabled them to:

- Modify regional goals from average sales, not just the mean

- If you want to find popular product categories, using the mode: That way of getting a sense of what were popular items.

- Shift more of its promotional spending to stores that are closer to the average performance levels

Problem 1: Misleading Benchmarking of Store Performance

The dependence of RetailMart on the arithmetic mean rendered unviable sales targets, for instance, for small stores in tier-2 and tier-3 towns.

Solution:

For performance comparisons use the median not the mean. Median is a great measure of the central value in a skewed data set because it mostly disregards extreme values.

The staff have underestimated the crowd based on average footfall.

On average, footfall data indicated even traffic distribution across the stores, but in the reality, most of the stores experienced crowdedness in weekends and extremely low crowd during weekdays.

Solution:

The mode can be used to determine the most common footfall pattern. Store managers would then know when to schedule its workers based on the day of peak traffic, making operations more efficient.

3) Single promotion strategy Problem: A strategy will be forced oblivious of customers' needs.

With the sole information of mean sale, RetailMart started promoting to high mean performing stores instead normal one's on a national level.

Solution:

Use average, median and specific statistical measures to segment the stores and set promotions. High performing stores got other incentives and median ones were given promotional support to improve performance.

Conclusion

The realization of the usefulness of various measures of central tendency in a business environment is demonstrated through this case study. Going beyond the one true metric and selecting the appropriate average based on data distribution enabled RetailMart to enhance their planning, fair treatment in performance reviews, and operational decisions. Using statistics tools such as mean, median, mode properly is also a good way to turn raw data into knowledge which can be applied in business decision making.

Statistics for Business Unit 6 V3.docx

Statistics for Business_BBA_2

Statistics for Business_BBA_2

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:127454069

Submission Date

Feb 3, 2026, 6:15 PM GMT+5:30

Download Date

Feb 3, 2026, 6:23 PM GMT+5:30

File Name

Statistics for Business Unit 6 V3.docx

File Size

38.3 KB

22 Pages

4,836 Words

26,313 Characters

1% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 15 words)

Match Groups

2 Not Cited or Quoted 1%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

1% Internet sources
0% Publications
0% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **2** Not Cited or Quoted 1%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

1% Internet sources
0% Publications
0% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1 Publication

Younghoon Kwak, Sun-Hye Mun, Chang-Dae Park, Sang-Moon Lee, Jung-Ho Huh. "... **<1%**

2 Internet

www.assignmenthelp.net **<1%**

Unit 6: Measures of Dispersion

Learning Objectives

1. Explain the purpose and importance of measuring dispersion in statistical data, and distinguish it from central tendency.
2. Calculate the range and interpret it as a basic measure of variability within a dataset.
3. Understand and compute quartile and percentile measures, and use them to analyze data distribution across different segments.
4. Calculate and interpret mean deviation for both ungrouped and grouped data, using both actual and assumed means.
5. Compute standard deviation and variance, and understand their role as key indicators of spread and consistency in a dataset.
6. Compare and contrast different measures of dispersion, and evaluate their appropriateness based on data characteristics and analytical needs.
7. Apply the concepts of dispersion to real-life business and economic scenarios, interpreting results for better decision-making and risk assessment.

Content

6.0 Introductory Caselet

"Ananya's Garment Analytics: Approach to Spread of Customers' Spends"

Background:

Ananya owns a mid-sized clothing brand that has multiple online and offline stores in three cities. And after a year of operating, she realized that customer spending per visit was, on average, ~₹1,200 ($20), yet sales varied with no rhyme or reason. On some days, small purchases by most customers accounted for the majority of sales; on others, a few big-spending buyers drove sales up considerably.

While she had already figured out what the average purchase value was, Ananya found that didn't tell the whole story. For her to truly understand this, she had to know how far customer spending deviations were spread around that average. Her marketing consultant had introduced her to the idea of dispersion, which quantifies the spread or variability in a set of data.

Ananya started with the range and found a big gap between the lowest and highest purchase. She then graduated to more reliable measures such as standard deviation and variance, in order to gauge how consistent customer spending was.

She learned:

• A small standard deviation implied that the majority of customers spent approximately around the mean,

• A high standard deviation indicated that the spending patterns are unsystematic,

• Quartile analysis suggested that nearly half of the total revenue was driven by top 25% of customers. By analyzing dispersion:

• For the top spending quartile, she revamped loyalty offers,

•Generating price-targeted promotions for the lowest 25 percent,

• And establish sales goals more realistically centered on variable spending rather than just averages.

For Ananya, the spread of information gives her more value than averages.

Critical Thinking Question:

If you were Ananya, and you noticed that your average sales figures remain around the same every month but that the standard deviation keeps climbing, what might have gone wrong? What decisions would you make with respect to marketing or customer differentiation?

6.1 Objectives of Measuring Dispersion

Measures of central tendency such as the mean, median and mode give a good idea about where the data is concentrated (the spread), but you don't know anything about how much it varies. Two sets of data can have the same average and presen t very different cohesive, dispersive or expansive trends. And this is where dispersion comes in.

Dispersion gives us an indication of how much the values differ from one another and from the central value. It is a measure of consistency or variability in a given data set, which enables us to analyze the data without any biases.

6.1.1 Meaning of Dispersion

Dispersion is a term that conveys how much the values of a data set spread out or scatter. It indicates the

diversity in the data.

The dispersion can be low if the values in the data set are all similar/close to each other. If values are much spread out dispersion is also high.

Example:

Consider these two sets:

• Set A: 45, 46, 47, 48, 49

• Set B: 20, 35, 50, 65, 80

Both potentially have similar averages, but there is no doubt that Set B's has more variation. This difference is picked up by dispersion.

6.1.2 Importance of Dispersion in Statistics

Dispersion is valuable to understand for several reasons:

Reveals Data Consistency:

Dispersion indicates the variance or reliability of data. Lesser dispersion means consistency which could be a good thing for quality control say.

Enhances Interpretation of Central Tendency:

Averages alone can be misleading. Dispersion gives the meaning to how well the mean or median can be interpreted.

Helps in Risk Analysis:

In finance or business terms, higher dispersion in returns means higher risk. Standard deviation and variance are what investors use to evaluate risk.

Supports Comparison Between Data Sets:

Two datasets with identical average may vary greatly. Dispersion reveals which one is the more stable or predictable.

Serves as Foundation for Sophisticated Statistical Algorithms:

Dispersion is a fundamental notion in numerous statistical approaches, such as regression, correlation or hypothesis testing.

6.1.4 The Applications of Dispersion in Business and Economics

Dispersion matters in decision-making:

Finance and Investment:

Volatility of stock return is also measured by standard deviation. With higher investment risk comes a higher standard deviation.

Production and Quality Control:

If production output are being generated well within a low range spread, it is an indication of process stability. Too much variation may result in waste or customer unhappiness.

Human Resources and Payroll:

An understanding of wage dispersion contributes to an analysis of income inequality, satisfaction among workers or the public policy-relevance of wage structures.

Marketing and Customer Analysis:

The dispersion in customer spending tell us if a business is dependent on a few high-paying customers, or whether it has broad and consistent base of customers.

Economic Policy and Planning:

Governments observe regional disparities in terms of, for example, income or employment dispersion to inform policy making and resource allocation.

6.2 Range

The range is the most basic of all measures of dispersion. It depicts the range of values in a data set. Simple but fast measure of spread, particularly helpful when comparing the spreads of more than two sets of data.

6.2.1 Definition and Formula for Range

The range is a dispersion measure which signals the distance between highest and smallest values in an observation. It lets us know the distance between the largest and smallest values.

Formula (for ungrouped data):

Range = Maximum value − smallest value Example:

If the temperatures in that week are:

28Â°C, 32Â°C, 35Â°C, 30Â°C, 27Â° C, 29Â° C and was established.

Then,

Range = 35 − 27 = 8°C

Coefficient of Range (Relative Range):

In the case of a pair or datasets that has varying units or is on the other direction, we employ range as a ratio using coefficient of range: Coefficient of Range = (L − S) ÷ (L + S)

Where:

L = Largest value of all S = Smallest value of all Eg:

If L = 80 and S = 20

Range Coefficient = (80 − 20) ÷ (80 + 20) = 60 ÷ 100 = 0.6

6.2.2 Merits and Limitations of Range

Merits:

Simple and Easy to Calculate

Only two values required—the maximum and the minimum. Useful for quick comparisons.

Gives a Quick Sense of Spread

Useful with early data exploration or if you are only interested in their extreme values.

Relevant for Qualitative and Quantitative Data

For numeric quantities (such as sales or temperature) or ordered categories that can be ranked.

Useful in Quality Control

Particularly for manufacturing, where identification of variation between the extreme readings is essential.

Limitations:

Ignores All Intermediate Values

It does not take into account what happens to the data in between the minimum and maximum.

Highly Affected by Outliers

Only one outlier may cause range distortion.

Unreliable for Large and Grouped Data

Doesn't give any information on the distribution / flatness of most values.

Not Suitable for Statistical Analysis

Range does not suffice to trigger more statistics, such as standard deviation or statistical hypothesis testing.

Did You Know?

“Although range is the simplest measure of dispersion, it was historically one of the first statistical tools used in meteorology. Scientists used it to identify climate variability in different regions long before more advanced measures like standard deviation were developed.”

6.2.3 Applications of Range

It is one of the most popular measures among different areas of application, especially when there is a demand for rapid evaluations of variability or risk.

Weather Forecasting:

Meteorologists apply the range to compare how much temperatures vary among cities or from season to season.

Stock Market Analysis:

Volatility is judged by the better as high − low of the day. A broad spread means a more volatile asset.

Sports Performance:

In games – cricket or football for example – a range in scores/distances might be indicative of consistency/variability among players.

Quality Control in Manufacturing:

Product dimensions (e.g., length, weight) are measured by engineers and the range is checked for conformance with standards.

Education and Test Analysis:

Teachers might check the range of scores on an exam to see how much students in a class differ in their performance.

Business Decision-Making:

By comparing the sales spectrum of branches, one may want to identify variation in performance.

6.3 Quartile and Percentile Measures

Whereas the minimum and maximum pertain to the value, quartiles and percentiles are about position within the data thus allow for a more intricate exploration on how values are distributed. The specific dividing lines used for the N-tiles are splitting the data into equal integer values and are very effective at identifying similarly located blocks of data, and outliers or skewness.

6.3.1 Quartile Deviation (QD)

Definition:

Quartile Deviation (SD) is another measure of dispersion which is based on the middle 50 per cent of the values. It measures the amount of spreading in the middle of a set of data values and is based on Q 1 (the first quartile) and Q 3 (third quartile).

Formula:

Quartile Deviation (QD) = $(Q_3 - Q_1) \div 2$ Coefficient of QD =

QD value = $(Q_3 - Q_1) \div (Q_3 + Q_1)$

Example:

If $Q_1 = 45$ and $Q_3 = 65$,

Therefore QD = $(65 - 45) \div 2 = 20 \div 2 = 10$

Coefficient = $(65 - 45) \div (65 + 45) = 20 \div 110 \approx 0.18$

Note: QD is robust to outliers and especially useful for skewed distributions.

6.3.2 Interquartile Range (IQR)

Definition:

The Interquartile Range (IQR) is a range of the middle 50% of your dataset. It is the difference between the first and third quartiles.

Formula:

$IQR = Q_3 - Q_1$

Use:

IQR is a resistant measure of spread since it does not take outliers into account. This is frequently used in box plots and for identifying outliers (values being more than 1.5 × IQR's away from Q_1 or Q_3).

Example:

If $Q_1 = 40$ and $Q_3 = 80$

Then $IQR = 80 - 40 = 40$

"Activity: Analyzing Income Segments Using Interquartile Range (IQR)"

Instruction to Student:

You are provided with the monthly income data (in ₹) of 15 families in a neighborhood:

₹18,000, ₹22,000, ₹24,000, ₹25,000, ₹27,000, ₹28,000, ₹30,000, ₹32,000, ₹33,000, ₹34,000, ₹35,000,

₹37,000, ₹40,000, ₹42,000, ₹48,000

1. Arrange the data in ascending order (if not already).
2. Identify Q_1 (first quartile), Q_2 (median), and Q_3 (third quartile).
3. Calculate the interquartile range (IQR).
4. Based on the IQR, analyze the income inequality in this neighbourhood.
5. Briefly comment on whether the central 50% of families have closely grouped incomes or not. Submit your quartile calculations and a 4–5 line interpretation.

6.3.3 Percentile Measures of Dispersion

Definition:

Percentiles split a data set into 100 equal portions. Used to analyze relative position and the distribution.

within large datasets.

• 10th percentile (P_{10}) - where at least 10% of the observations are lower than this value.

• P_{90} is the 90th percentile, which is the value at or below which 90% of measures are found.

• The differences between any two percentiles can quantify spread.

Formula (for percentile range):

Percentile Range = $P_{90} - P_{10}$

This may also be used instead of IQR (eg if the focus is on a wider range about the central value).

Example:

If $P_{10} = 20$ and $P_{90} = 90$,

Hence Percentile Range = $90 - 20 = 70$

34 Quartile-like rankings are commonly used in educational testing and demographic studies and health indicators.

Did You Know?

“In large-scale educational testing (like SAT, GRE, and India’s NEET), students are not ranked by marks but by percentiles. A student with a percentile of 95 didn’t score 95 out of 100, but instead scored better than 95% of all candidates. This makes percentile-based dispersion measures essential for interpreting competitive exam results.”

6.3.4 Uses of Quartiles and Percentiles

Quartiles and percentiles are very common in several scientific disciplines for analysis and comparison:

Education and Testing:

Percentile ranks indicate the percentage of students a child scored better than (“Your child scored in the 85th percentile”).

Human Resources and Salary Analysis:

Quartiles are utilized to compare the distribution of wages and establish pay standards. For instance, firms might zero in on 75th percentile top performers.

Medical and Health Research:

Percentiles are commonly used in growth charts (child's height or weight at the 50th percentile).

Market Research and Customer Segmentation:

Commonly used to analyze spending or customer value by separating customers into quartiles or deciles.

Outlier Detection:

In data visualization / cleaning, we often consider values beyond Q_1 −1.5×IQR and Q_3 + 1.5×IQR as outliers.

Data Summarization:

Box plots are a powerful exploratory data analysis (EDA) tool, and quartiles are useful because they assist in creating them.

6.4 Mean Deviation

Mean Deviation Mean deviation is a measure of dispersion which tell us on the average, how much each value deviates from the central value (mean, median or mode). It does not square the deviations like standard deviation, and is more intuitive and less sensitive to outliers.

6.4.1 Definition and Computation of Mean Deviation

Definition:

Mean absolute deviation (once known as the average absolute deviation) is the mean of the absolute deviations from a point, i.e.,av |(x i − c),)/(n − 1.

Formula (Ungrouped Data):

MD = (Σ|x − A|) ÷ n

Where:

• x = individual observations

• A = the central value (mean, median or mode)

• n = number of observations

• |x − A| = Non-negative difference – absolute deviation If you have forgotten what "absolute value" is, refer back to the Math Review of Absolute Value – Explanation.

Example:

Data: 5, 7, 9

Mean ($\bar{x}$) = (5 + 7 + 9) ÷3 =21 ÷ 3=7

MD = (|5−7| + |7−7| + |9−7|) ÷ 3 = (2 + 0 + 2) ÷ 3 =1.33

Grouped Data Formula:

MD = (Σf × |x − A|) ÷ Σf

Where:

• f = frequency of each class

• x = mid-point of class

• A = value around which the observations cluster (mean, median or mode)

6.4.2 Mean Deviation about Mean, Median, and Mode

The value of the mean deviation may be computed from any one of the three central values thus:

A. Mean Deviation about Mean

Until now mainly seen for Mathematics/Statistics.

Formula:

$MD_{\bar{x}} = (\Sigma|x - \bar{x}|) \div n$

B. Mean Deviation about Median

This is commonly used when the data is highly skewed since median getting affected by those outliers will be relatively less. Formula:

$MD_{me} = (\Sigma|x - M|) \div n$

C. Mean Deviation about Mode

Applied from time to time, replaced by categorical or modal distributions.

Formula:

$MD_{mo} = (\Sigma|x - Mo|) \div n$

Note:

Average absolute deviation about the median (which is defined as MAD for around the median) necessarily generates smallest value among all central points, so it can be utilized to reduce total deviations.

6.4.3 Merits and Limitations of Mean Deviation

Merits:

Easy to Understand and Interpret

Results are then easier to understand because deviations are computed in absolute values.

Based on All Observations

The full range of values are used in the computation.

Better than Range

Unlike range, it provides a clearer picture of average spread rather than just extremes.

Useful for Skewed Distributions

When calculated from median, it is a trustworthy measure for non-symmetrical distributed data.

For Single and Grouped Data

Flexible across data types.

Limitations:

Ignores Direction of Deviation

By taking absolute values, it does not differentiate between data points above and below the center value.

Of No Further Value Algebraically

Unlike variance and standard deviation, it does not have any properties that are interesting for further statistical purposes.

Less Common in Inferential Statistics

It’s not commonly used in modelling, hypothesis testing or prediction.

Manual Calculation Can Be Time-Consuming

Particularly when we have grouped data and hence mid points are involved like scores / frequencies.

6.5 Standard Deviation and Variance

Deviation standing and variance are the ones which are most commonly used as well as reliable measures of dispersion in statistics. Not only are they an indicator of the common rough size, but they are the foundation for many powerful statistical techniques and analyses. These points are useful for evaluating the stability, risk and variation aspects of data.

6.5.1 Concept of Standard Deviation

Definition:

The SD is the sqrt of the mean of the squares deviations from the mean. It indicates how much each data point differs from the mean and provides a sense of the spread of the data.

1

• A low standard deviation indicates that the values tend to be close to the mean of the set (the closer, the better).

• If the standard deviation is high, then the values are widely spread apart.

Symbol:

• The Standard Deviation: σ (population), s (sample)

Conceptual Formula:

Standard Deviation (σ) = √[Σ(x − x̄)² ÷ n] where, Σ : Sum of the all elements (x − x̄)²: Difference between each element and its mean raised to the power of 2.

Where:

• x = each data point

• x̄ = mean of the data

• n = total number of observations.

6.5.2 Calculation of Standard Deviation (Ungrouped and Grouped Data)

A. Ungrouped Data (Raw Data) Steps on How to do it:

Find the mean (x̄).

Find the difference between each of the values and their mean (x − x̄).

Square each result.

Take the average of these squared differences.

Take the square root.

Formula:

$\sigma = \sqrt{[\Sigma(x - \bar{x})^2 \div n]}$

Example:

Data: 4, 6, 8

Mean $\bar{x} = (4 + 6 + 8) \div 3 = 6$

Squared deviations = $(4-6)^2 + (6-6)^2 + (8-6)^2 = 4 + 0 + 4 = 8$ Variance = $8 \div 3 = 2.67$

Standard deviation = √2.67 1.63 (n.a.).5 degrees of freedom for S.E.

B. Formula for Grouped Data (Direct Method):

$\sigma = \sqrt{[\Sigma f(x - \bar{x})^2 \div \Sigma f]}$

Where:

• x = mid-point of class

• f = frequency

• $\bar{x}$ = mean

• Σf = total frequency

Shortcut (Assumed Mean) Method:

If numbers are large, use:

$\sigma = \sqrt{\{[\Sigma f(d^2) \div \Sigma f] - (\Sigma f \times d \div \Sigma f)^2\}} \times h$ where: σ = standard deviation of method for pellets, h = height.

Where:

• d = (x − A)/h deviation from assumed mean / total number of deviations=3/4 (University of Mumbai) 43 SEMESTER-III a.

• h = class width

“Activity: Exploring Output Consistency through Standard Deviation”

Instruction to Student:

You are given the following daily output data (in units) for two machines over 7 days:

Machine A: 120, 118, 122, 121, 119, 120, 121

Machine B: 105, 115, 140, 90, 130, 95, 135

1. Calculate the mean output for each machine.

2. Compute the standard deviation for both machines using the appropriate formula for ungrouped data.

3. Interpret the results: Which machine is more consistent?

4. Write a short paragraph explaining why standard deviation is more informative than just comparing the mean output.

Submit your full calculations and interpretation.

6.5.3 Concept and Calculation of Variance

Variance is simply the average of the squared differences from the mean. It is the square of the standard deviation and it used to measure how widely spread out are all the points.

Formula for Variance:

Relationship: $\text{Var}(\sigma^2) = \Sigma(x-\bar{x})^2 \div n$ Where:

Standard Deviation = √Variance

Example (continued):

From earlier: Variance = 2.67 So SD = √2.67 ≈ 1.63

Note: Variance is helpful in statistical modelling, but SD is more intuitive.

Did You Know?

“The term "variance" was first introduced by Ronald A. Fisher in 1918, not as a theoretical idea, but as a practical tool for analyzing agricultural crop experiments. Variance is now one of the most important tools in risk analysis, genetics, psychology, and finance, where it's used to calculate volatility in stock returns and variability in genetic traits.”

6.5.4 Properties and Applications of Standard Deviation

Properties :

Non-negative

SD is always greater than or equal to zero.

Minimum Value

If all X values are equal, SD = 0.

Uses All Observations

It is all-inclusive, because it's based on each data point.

Mathematically Treatable

Appropriate for algebraic operations in additional statistical processing.

Affected by Extreme Values

Because the deviations are squared, major differences matter a lot.

Applications:

Finance:

As a measure for investment returns volatilities (risk assessment).

Business:

Aids in the control of costs and variations in consistency and quality.

Education:

Aids in comparing student performance across varied subjects or assessments.

Healthcare:

Allows you to understand how much variability there is in the amount treatment effects are going to have, lab test results or response rates.

Operations Management:

Recognizes the perturbations in demand forecast, inventory, and process control.

6.5.5 Merits and Limitations of Standard Deviation

Merit:

Most Accurate Measure of Dispersion

The SD as opposed to range or mean deviation exhibits true variability with accuracy.

Based on All Observations

Optimally utilizes the available data.

Useful in Inferential Statistics

Basis for z-scores, confidence intervals, regression, etc.

Mathematically Efficient

is easily handled using equations.

Limitations:

Affected by Extreme Values

Because it is taking the squares of deviations outliers matter more.

Complex for Beginners

Slightly more complicated and requires you know what something squared venoms.

Less Intuitive Interpretation

SD is more difficult to comprehend at a glance than range or mean deviation.

Sensitive to Scaling

The SD changes directly when the units or scale of measurement are changed.

Knowledge Check 1

Choose the correct option:

1. Which of the following measures is most affected by extreme values?

A) Interquartile Range

B) Standard Deviation

C) Median

D) Mode

2. What does a low standard deviation indicate about a dataset?

A) The data is skewed to the left

B) The data values are widely spread

C) The data values are tightly clustered around the mean

D) The dataset has outliers

3. Which formula is used to calculate quartile deviation (QD)?

A) $(Q_1 + Q_3) \div 2$

B) $Q_3 - Q_1$

C) $(Q_3 - Q_1) \div 2$

D) $Q_3 \div Q_1$

4. If the mean of a data set is 60, and the variance is 25, what is the standard deviation?

A) 5

B) 12

C) 35

D) 8

5. In which of the following scenarios would interquartile range (IQR) be the most appropriate measure of dispersion?

A) When you want to include all values

B) When the data has extreme outliers

C) When data is normally distributed

D) When the mean is equal to the median

6.6 Summary

- This chapter covered dispersion, a fundamental idea in descriptive statistics that represents how much data values are spread out around a central value. Central tendency is to determine the center of a data set, dispersion will help us get a sense for how consistent, or spread out (or even variable), the data are.
- • Range is the most basic measure, a view into how wide the highest differences and lowest values spread.
    - Quartile and Percentile: Each of these measures splits the data into equal portions, which makes it easy for you to compare the position and spot outliers.
    - Mean Deviation averages the difference between values and center and is thus a more moderate measure of diversity.
    - Standard Deviation and Variance are both mathematically sound measures, which are often applied in corporations, banks, provisions of service companies as well as controlled laboratory experiments for the measurement of risk levels, consistency level and degree of variability.
- Taken together, these tool enable analysts and decision makers to better know the reliability of data, its spread and basic pattern.

6.7 Key Terms

1. Dispersion - The amount of the spread of a set of data values around an average or middle value.

2. Range - The difference between the highest and lowest value of a set of values or numbers.
3. Quartile - Values that divide the data set into four equal parts
4. IQR - Difference between the third quartile and first quartile ($Q_3 - Q_1$)
5. Percentile - The value below which falls a certain percentage of observations
6. Mean Deviation: It is average between absolute differences to a central value.
7. Standard deviation is like the average distance of a data point from the mean.
8. Variance - average of squared differences from the mean
9. Dispersion coefficient (CoD)- Dimensionless ratio of variability, not expressed in units
10. Outlier - Any observation very far away from the rest of the observations

6.8 Descriptive Questions

1. Define dispersion. Why do we have to learn dispersion in statistics?
2. What is range? How is it measured, and whom does it leave out?
3. Define with example the quartile deviation and interquartile range.
4. Differentiate between quartiles and percentiles.
5. What is the method of finding mean deviation from ungrouped and grouped data?
6. Compare the mean deviation about mean, median and mode.
7. Define standard deviation. Then, using the short-cut method for discrete grouped data.
8. Differentiate variance from standard deviation.
9. Describe two business applications of standard deviation in everyday life.
10. Describe advantages and disadvantages of range, mean deviation, and standard deviation.

6.9 References

1. Gupta, S. C. (2014). Fundamentals of Statistics. Himalaya Publishing House.

2. Levin, R. I., & Rubin, D. S. (2012). Statistics for Management. Pearson Education.

3. Sharma, J. K. (2018). Business Statistics. Vikas Publishing House.

4. Spiegel, M. R., & Stephens, L. J. (2018). Schaum's Outline of Statistics. McGraw Hill Education.

5. Ministry of Statistics and Programme Implementation (MoSPI), Government of India

6. UGC e-Pathshala: Online modules on Statistics and Data Analysis

Answers to Knowledge Check

Knowledge Check 1

1. B) Standard Deviation
2. C) The data values are tightly clustered around the mean
3. C) $(Q_3 - Q_1) \div 2$
4. A) 5
5. B) When the data has extreme outliers

6.10 Case Study

"Detect Variability in Factory Production: Rohan's Dilemma"

Introduction

In all such production-based industries, it is equally crucial to measure the variance of output as a measure of dispersion as that of means used for describing output trends. For Rohan, who works as an operations manager at a medium sized electronics manufacturers using average Daily production data was not enough to guide his decision making. His factory had several machines making LED circuit drivers. First attention was trained on average daily output — typically 120 to a printer. But quality problems with the product and an inability to count on how many would ship suggested that something deeper was going wrong.

So when Rohan moved beyond averages, he moved to statistical measures of dispersion range, mean deviation and standard deviation. These enabled him to measure output variance, and determine whether each machine's performance was uniform or erratic. That high variation, even if the average was stable, could have an effect on quality and delivery speed. This case discussion is about how Rohan leveraged dispersion metrics to expose deeper inconsistencies in production and adopt operational improvements based on data.

Background

Rohan took readings from two machines (Machine A and Machine B) for a 10 day production run. While the two machines advertised a daily output of about 120 units, those from Machine B were much more volatile from day to day (90–150 units). Meanwhile, the outputs from Machine A hovered consistently within a band of 115–125.

He calculated:

• State: Machine A had a state that varied by 10 units; Machine B's state varied by 60.

• Standard Deviation: Machine A = 4.2 units; Machine B = 18.5 units • Item to Item Variation for Machines has been defined as within the same plant per machine and was a common index used in industry for relative comparison of different machines.

• Average deviation: Machine A = 3.9 units; Machine B = 17.2 units.

Rohan found out that it was because of the irregularity of Machine B which had an impact on the general flow of Production causing certain shifts to face stock outs and in ability to pack. He reported the results to the maintenance group where they saw signs of intermittent heating on a component within Machine B.

Problem 1: The Peril of Overgeneralizing from Averages

Consideration of the mean output alone made both machines equally productive. But it masked serious variations in thickness.

Solution:

Rohan included regular standard deviation calculations alongside the average in daily reports. This enabled management to spot deviations early and schedule maintenance.

(Attention: specific problems or needs are identified based on your discussion.) P2: Variability-induced inventory planning challenges.

The large variance and attenuation in Machine B output were the obstacle to designing an efficient purchasing of raw materials and a timing for labour shifts.

Solution:

Rohan collaborated with a supply chain team to utilize IQR and standard deviation for setting safety stock levels and buffer time, cutting down on last-minute inventory hiccups.

Problem Statement 4: Inefficient Benchmarking Across Machines

Machines were judged by production alone, with no consideration of reliability and variance.

Solution:

A new criterion was defined based on mean and variance of the output values. Machines with a high average and low variability were rated higher.

MCQ:

Which of the following is best to detect inconsistent output when mean is equal?

A) Mode

B) Median

C) Range and Standard Deviation

D) Total Output

Answer: C) Range and S.D.

Explanation:

These metrics explain the data spread and reveal how steady is a machine's performance, even when the average performance doesn't change.

Conclusion

This case emphasizes the need to understand not only what is being produced, but how consistently it is produced. Rohan upvoted: "I Decision Rohan's decision to add dispersion measures unearthed hidden inefficiencies that the mean alone could not uncover. For that reason, he was able to minimize production problems, enhance quality control and manage resources more efficiently. Dispersions helped the team make more informed data-backed decisions to drive operational efficiency.

Statistics for Business Unit 7 V3.docx

Statistics for Business_BBA_2

Statistics for Business_BBA_2

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:127498765

Submission Date

Feb 4, 2026, 10:29 AM GMT+5:30

Download Date

Feb 4, 2026, 10:33 AM GMT+5:30

File Name

Statistics for Business Unit 7 V3.docx

File Size

103.6 KB

21 Pages

4,461 Words

24,657 Characters

6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 15 words)

Match Groups

15 Not Cited or Quoted 6%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

4% Internet sources
0% Publications
5% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **15** Not Cited or Quoted 6%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

- 4% Internet sources
- 0% Publications
- 5% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	brightideas.houstontx.gov <1%
2	Submitted works	Myanmar Imperial University on 2025-10-29 <1%
3	Internet	info.daviscollege.edu <1%
4	Submitted works	Myanmar Imperial College on 2025-07-04 <1%
5	Submitted works	Nanyang Polytechnic on 2024-08-18 <1%
6	Submitted works	CTI Education Group on 2025-03-11 <1%
7	Submitted works	Myanmar Imperial University on 2025-10-30 <1%
8	Submitted works	Trident University International on 2024-11-23 <1%
9	Submitted works	King's College on 2023-12-14 <1%
10	Submitted works	Melbourne Institute of Technology on 2025-10-03 <1%

11 Submitted works

University of Durham on 2010-04-28 <1%

12 Internet

myweb.uiowa.edu <1%

13 Internet

towardsdatascience.com <1%

14 Internet

methods.sagepub.com.suss.remotexs.co <1%

Unit 7: Probability Distributions

Learning Objectives

1. Explain the concept of probability distributions, including the distinction between discrete and continuous distributions.

2. Understand the characteristics, assumptions, and probability structure of the Binomial Distribution, and apply it to real-life problems involving success/failure outcomes.

3. Define and apply the Poisson Distribution, identifying suitable conditions (such as rare events over fixed intervals of time or space) and computing relevant probabilities.

4. Recognize the key features of the Normal Distribution, including symmetry, bell-shaped curve, mean- variance relationship, and the empirical rule.

5. Use the Standard Normal Distribution (Z-distribution) to compute probabilities and understand standardization of raw scores.

6. Differentiate between Binomial, Poisson, and Normal Distributions, and select the appropriate distribution model based on data conditions.

7. Apply distribution models in business, economics, operations, and quality control, interpreting outcomes to support decision-making under uncertainty.

Content

7.0 Introductory Caselet

"Maya's Delivery Dilemma: Predicting with Probability Distributions"

Background:

Maya operates the operations for a local same-day courier company. Her firm has individual and business clients throughout the city. Eventually, Maya started observing a mysterious pattern — some days her drivers would be swamped with delivery requests and other days they wouldn't have anything to do. Even though the monthly average for deliveries was steady at around 200, the daily numbers were erratic.

Maya first attempted predicting deliveries based on averages, but its predictions were erratic and unhelpful. Even though she's comfortable with spreadsheets, she wanted a more sophisticated way to model the uncertainty and variability of how many robots might have to be deployed in daily delivery volumes and other factors, so she turned to a data analyst. The analyst opened Maya's eyes to the world of probability distributions.

They started by applying Binomial Distribution, which models the probability of failure for a given number of trials in each scenario – whether packages would be delivered on time or not, based on delivery performance history. They next used the Poisson Distribution for the number of delivery requests per day, especially where such delivery requests were randomly and independently distributed.

For example:

• It averaged 12 orders an hour, but the number varied. Applying Poisson allowed her to predict the likelihood that she would receive 15 or more orders during any given hour.

• When the analyst wanted to forecast generalized customer behavior across a month, they turned to the Normal Distribution, fitting neatly for capturing the distribution of total monthly delivery distances.

By applying probability distributions:

• Maya could meanwhile better predict driver effort,

• Involve backup vehicles in peak Poisson-projected times,

• And predict risks of long-late delivery using binomial probability models.

What appeared to be capriciousness in operations fit statistical patterns, she found, once she imposed the right distribution models.

Critical Thinking Question:

As Maya, and you discover that the arrival of customer orders are random on an hourly basis, how would applying a Poisson Distribution assist in scheduling staff and fleet availability? Is there another business process where the Poisson model could be used?

7.1 Introduction

In probability and statistics theory, we often face that the outcome is uncertain and random but follows predictable pattern. These patterns are described byprobability distributions — mathematical functions that indicate the probability of different outcomes of a random experiment.

There are two primary types of probability distributions:

• Distributions for discrete variables, in the sense that values are counted (e.g., number of defective items from a production batch).

• Continuous distributions, in which values may be drawn from the entire range (height of people; time to carry out work).

These distributions are of primary importance for predictions, risk assessment, and decision making under uncertainty conditions. Random variables that are distributed according to certain well-established distributions appear in many practical business, industrial and scientific problems. Of these, three are of particular significance:

• Binomial Distribution (discrete)

• Poisson Distribution (discrete)

• Normal Distribution (continuous)

All of these distributions are appropriate for certain types of data and situations. Understanding when and where to add them are fundamental principles of probability theory and statistical modelling.

7.1.1 Introduction and Applications of Some Special Probability Distributions

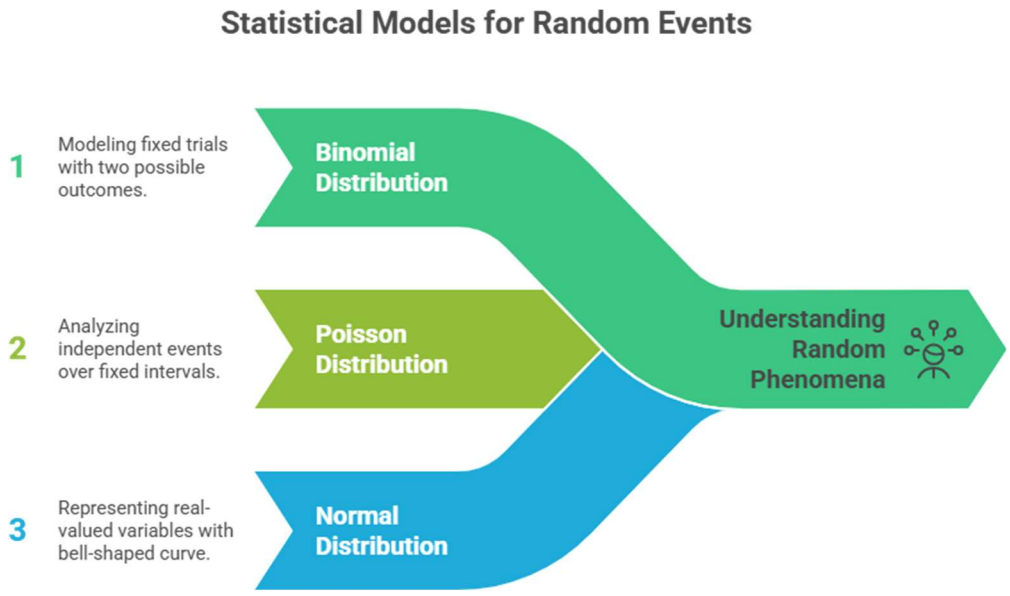


Fig.7.1. Introduction and Applications of Some Special Probability Distributions

There are certain probability distributions which appear so frequently in problems that they have been given their own names, these special probability distributions. These include:

Binomial Distribution

• Nature: Discrete

• Applies to: The result of an experiment is binary (success or failure), repeated under identical conditions.

• Example Applications:

o Foretelling the amount of malfunctioning items in a lot.

o Predicting how many successful sales calls one can make in a day.

o Modelling pass or fail to quality control status.

Poisson Distribution

• Nature: Discrete

• If: The event is rare (it happens randomly and independently over some fixed time or space period).

• Example Applications:

o Estimating the number of customers calling in an hour.

o The mean number of accidents which occur at a junction within a week.

o Forecasting arrivals to a point of service (such as patients in a clinic).

Normal Distribution

• Nature: Continuous

• When to use: Can be used if the data is symmetric around an average and the majority of values are near that average.

• Example Applications:

o Evaluating performance test scores or job performance ratings.

o Forecasting demand and supply.

o Manufacturing in-line quality control (e.g., product dimension measurement). These are remarkable distributions in statistical analysis. They help businesses:

• Estimate probabilities

• Set control limits

• Model uncertainties

• Optimize resource allocation

In this detailed examination of each distribution, let's understand the assumptions, parameters and formulas participated., and the kinds of problems for which each is best suited.

7.2 Binomial Distribution

2 One of the most common discrete probability distributions is the Binomial Distribution. It represents the number of successes in a fixed number of independent trials, each with only two possible outcomes: success or failure.

Concept and Properties of Binomial Distribution Concept: If the probability distribution follows a binomial, then it is called a binomial distribution.

There is a binomial distribution whenever a random experiment is:

• N times (fixed number of trials),

• The product of and is used as the likelihood for successive trials, where each trial leads to a success (S) or failure (F),

11 • There exist constant probability of success (p) in every trial,

• The test instances are all mutually independent.

If the number of successes in n trials is given by X, then we say that X has a Binomial distribution with parameters n and

p, written as:

X ~ B(n, p)

Key Characteristics:

• Distribution: X is distributed discretely on the values 0, ..., n.

• Two parameters: n (number of independent trials) and p (probability that event is a success).

• Failure probability q = 1 − p.

• Symmetric if p = 0.5; otherwise, it will be skewed.

12 • When n is large, and p isn't too close to 0 or 1, the binomial distribution looks like a normal distribution.

7.2.2 Probability Mass Function (PMF) of Binomial Distribution

3 The PMF provides the probability of getting exactly k successes in n trials:

$P(X = k) = C(n, k)p^k q^{n-k}$

Where:

• C(n, k) = n! ÷ [k! (n − k)!] (binomial coefficient)

• p = probability of success

• q = 1 − p = probability of not success

• k = number of successes ($0 \le k \le n$)

Example:

If a coin is flipped 4 times (n = 4 and probability of heads per trial wasp = 0.5), then the probability of getting exactly 2 successes is:

$P(X = 2) = C(4, 2) \times (0.5)^2 \times (0.5)^2 = 6 \times 0.25 \times 0.25 = 0.375$

Did You Know?

“The Binomial PMF is not just used in statistics—it's also used in genetics. In Mendelian inheritance, the probability of inheriting dominant or recessive traits follows a binomial pattern. For example, if two heterozygous parents cross (Aa × Aa), the probability of their child inheriting a recessive gene (aa) follows the same logic as calculating binomial probabilities.”

7.2.3 Mean and Variance of Binomial Distribution

In terms of statistics, these statistical properties are part of the binomial distribution:

13 • Mean (μ) = $n \times p$

• Variance (σ^2) = $n \times p \times q$

• Standard Deviation (σ) = $\sqrt{(n \times p \times q)}$ Here, n is the number of trials.

Interpretation:

- The average is the number of successes you expect.
- The variance and the standard deviation tell how much variation there is about the average.

Example:

For a binomial distribution of n = 10 and p = 0.4:

- Mean = 10 × 0.4 = 4
- Variance = 10 × 0.4 × 0.6=2.4
- √Variance = √2.4 ≈ 1.55

7.2.4 Applications and Examples of Binomial Distribution

The binomial distribution is common in independent repeated binary experiments. Common applications include:

Business and Quality Control:

- Finding the probability that a batch contains defective items up to a specified number.
- Counting the successful sales calls, out of all attempts.

Healthcare:

- Simulating the odds that a specified number of patients will recover from treatment.

Education:

- Predicting the likelihood that a certain number of students will pass an exam.

Finance:

- The number of successful investments or days in the market. Example:

The probability of a factory producing defect is 5%. And so, in 10 items(n=10), what is the probability that exactly one of itc is defective?

Here,

- p = 0.05 (defective),
- q = 0.95 (non-defective),
- n = 10,
- k = 1

$P(X = 1) = C(10, 1) \times (0.05)^1 \times (0.95)^9 \approx 10 \times 0.05 \times 0.630 = 0.315$

7.3 Poisson Distribution

The Poisson distribution is a for the number of events having an occurrence at random intervals in a given space/time interval and assuming a known average rate of events per interval.

7.3.1 Concept and Characteristics of Poisson Distribution

Concept:

7 A Poisson's distribution is a probability that provides the likelihood of the number of events that occur in a fixed interval of time, distance, space, or volume given:

These events are uncommon, sporadic occurrences.

- The rate parameter (λ) is independent of time,
- Two things cannot happen at the same time.

Notation:

Where X is the number of occurrences, we can express this as:

X ~ Poisson(λ)

Where:

λ (lambda) = mean number of events in interval

Characteristics:

- X is a discrete distribution (X = 0,1,2,3,…).
- It models count data (e.g., number of emails per hour).
- In theory, the number of occurrences is unlimited.
- The events are mutually independent and exponentially distributed.
- The mean and the variance are (λ), where λ > 0.

7.3.2 Probability Mass Function (PMF) of Poisson Distribution

The PMF of the Poisson distribution can be expressed as:

$P(X = k) = (e^{(-\lambda)} \times \lambda^k) \div k!$

Where:

- λ = average number of event occurrences in the interval

• k = occurrence number (k = 0, 1, 2,...)

• e ≈ 2.71828

• k! = factorial of k

Example:

(e) When λ = 4 (mean number of calls per hour at a call centre).

What is the chance of receiving only 2 calls in an hour? P (X = 2) = (e^(−4) × 4^2) ÷ 2!

= (0.0183 × 16) ÷ 2 = 0.1464

Did You Know?

“The Poisson distribution was actually developed as a way to approximate the binomial distribution for rare events. In fact, it was first used in 1837 by Simeon Denis Poisson to analyze the number of soldiers accidentally killed by horse kicks in the Prussian army—making it one of the earliest data-based applications of probability theory.”

“Activity: Modelling Customer Support Calls with Poisson Distribution”

Instruction to Student:

A customer support centre receives an average of 5 calls per hour. Using the Poisson distribution, calculate the probability that:

- a) Exactly 3 calls are received in an hour
- b) 5 or more calls are received in an hour

Use the formula:

P(X = k) = (e^(−λ) × λ^k) ÷ k!, where λ = 5

Show all steps and round your final probabilities to four decimal places. Comment on how useful this method would be for planning the number of support staff needed per shift.

7.3.3 Mean and Variance of Poisson Distribution

The Poisson distribution has a mean = variance = λ.

• Mean (μ) = λ

• Variance (σ^2) = λ

• Standard deviation (σ) = √λ

Example:

If λ = 5, then:

• Mean = 5

• Variance = 5

• S.D = √5 ≅ 2.24

It is this property which makes the Poisson distribution so easy to interpret, particularly when used in queueing systems and inventory analysis.

7.3.4 Relationship between Poisson and Binomial Distribution

The Poisson distribution is the limit of a binomial distribution under certain conditions:

• The value of n → Δ (for very many test series),

• The probability of hitting your target p → 0 (wee),

• The product n× p = λ is still finite.

9

This is the relationship which justifies using the Poisson distribution as an approximation to the binomial.

when:

• n is large (often > 20), and

• p is low (often < 0.05)

Example Use Case:

If we wish to model the number of typing errors in a given page, and only few pages contain many words (n is large) but the probability per work is very low (p is small), the Poisson distribution will be more efficient than the binomial.

Applications and Illustrations of Poisson Distribution

Poisson distribution is broadly employed across fields to model random, independent events over time or space.

Business & Operations:

• Number of calls to customer service per hour

• Defects per 100m in the cable

Healthcare:

• Caseload per night in an emergency room

• Rareness of diseases in a community

Manufacturing & Quality Control:

• Flaws per sq. mtr of fabric

• Defective parts per shipment

Traffic and Transport:

• Cars passing through a toll plaza in 1 minute

• Weekly train delays

Example:

I get 3 hits per minute to my web server. What is the likelihood that it will receive exactly 5 hits in a minute?

$\lambda = 3, k = 5$

$P(X = 5) = (e^{(-3)} \times 3^5) \div 5!$

= 0.1008 (Another rounding A value): 0.1008

So 10.08% of the tome server gets exactly 5 hits in 1 minute.

7.3.5 Application/Examples of Poisson Distribution

The Poisson distribution is commonly used in many industries to model random independent events over time or space.

Business & Operations:

• Customer service calls per hour

• Defects per 100 m of cable

Healthcare:

• Visits per night of patients to an emergency room

• Prevalence of rare diseases within a population

Manufacturing & Quality Control:

• Defects per square meter on fabric

• Number of defective parts within a shipment

Traffic and Transport:

• Flow of vehicles at a toll plaza in cars per minute

• Frequency of train delays per week

Example:

There are 3 hits per minute to a web server. So, what is the chance that it hits exactly 5 time in any one minute?

$\lambda = 3, k = 5$

$P(X = 5) = (e^{(-3)} \times 3^5) \div 5!$

$= (0.0498(243)/120) \approx 0.1008$

Well, a 10.08 % chance that the server gets exactly (in one minute) five hits."

7.3.5 Applications and Examples of Poisson Distribution

The Poisson distribution is commonly used in many industries to model random independent events over time or space.

Business & Operations:

• Customer service calls per hour

• Defects per 100 m of cable

Healthcare:

• Visits per night of patients to an emergency room

• Prevalence of rare diseases within a population

Manufacturing & Quality Control:

• Defects per square meter on fabric

• Number of defective parts within a shipment

Traffic and Transport:

• Flow of vehicles at a toll plaza in cars per minute

• Frequency of train delays per week

Example:

There are 3 hits per minute to a web server. So, what is the chance that it hits exactly 5 time in any one minute?

$\lambda = 3, k = 5$

$P(X = 5) = (e^{(-3)} \times 3^5) \div 5!$

$= (0.0498(243)/120) \approx 0.1008$

Well, a 10.08 % chance that the server gets exactly (in one minute) five hits."

7.4 Normal Distribution

The Normal Distribution is a statistical distribution of continuous probability, that describes how the values of a random variable are distributed. It is also known as the Gaussian distribution and its use in probability and statistics has made it one of the most important functions for natural phenomena.

7.4.1 Concept and Characteristics of Normal Distribution:

Concept:

The Normal Distribution The normal (Gaussian) distribution is used to model continuous data whose values are distributed symmetrically around a central mean. The normal 'bell-shaped curve' is used to describe data in which:

• The majority of values are concentrated around the average,

• Values are less likely to occur as we go away from the average. It is characterized by two parameters:

• μ (mu) = Mean (center of the distribution),

• σ (sigma) = Standard deviation of the Gaussian kernel (controls spatial spread).

Key Characteristics:

• Symmetrical about the mean (μ),

• Mean = Median = Mode,

• Area under the curve (Total) = 1,

- The curve never touches its asymptote,

Defined for all real x(−∞ to ∞),

- The form is entirely specified by μ and σ.

7.4.2 Probability Density Function (PDF) of Normal Distribution

10 The Probability Density Function (PDF) for a normal distribution is:

$f(x) = (1 / (\sigma\sqrt{2\pi})) * e^{(-(x - \mu)^2 / (2\sigma^2)}$

Where:

- x = any real number,
- μ = mean,
- σ = standard deviation,
- e is approximately equal to 2.71828 (natural base),
- π ≈ 3.14159

This is the bell-shaped curve. The curve is centred at x = μ, and the width of the curve is governed by σ.

Did You Know?

"The Poisson distribution was actually developed as a way to approximate the binomial distribution for rare events. In fact, it was first used in 1837 by Simeon Denis Poisson to analyze the number of soldiers accidentally killed by horse kicks in the Prussian army—making it one of the earliest data-based applications of probability theory."

7.4.3 Properties of Normal Curve

Symmetric: Curve is symmetric with respect to the mean. There is left-right symmetry of the two halves.

Mean = Median = Mode: All the averages are placed in the center.

Bell Curve: Uni-Modal, smooth curve of distribution.

Infinite Tails: the curve never touches the x-axis, but it gets closer and closer to 0.

Area under the Curve:

o Approximately 68.27% of the data fall within ±1σ of the mean,

o About 95.45% lies within ±2σ,

o About 99.73% lies within ±3σ.

Empirical Rule: These percentages (68–95–99.7 rule) are helpful for estimating probabilities.

Did You Know?

"Over 99% of human heights, test scores, and even errors in manufacturing follow a normal distribution. That's why it's often called the "natural law of error." The fact that the mean, median, and mode are all the same in a normal distribution makes it uniquely useful for modelling balanced systems and populations."

7.4.4 Standard Normal Distribution (Z-Distribution)

A Standard Normal Distribution is simply a normal distribution where the mean(μ) is 0 and the

standard deviation (σ) is 1. This distribution is known as a Z-distribution.

Z-Score Formula

Here the Z-score is simply quantifies how many standard deviations a value (X) deviates from the mean (μ):

$Z = (X - \mu) \div \sigma$

Where:

• X = Individual value

• μ = Mean of the distribution

• σ = The standard deviation of the distribution

Interpretation of Z-scores

•A Z-score greater than 0 indicates a score above the mean

• A Z-score below zero would represent a value less than the mean

• A Z-score of 0 corresponds to the value being the mean

Example

Test has the following statistics:

• Mean (μ) = 70

• Standard deviation (σ) = 10

• For one student, say he scores X = 85 Using Z-score formula:

Z = (85 − 70) ÷ 10 = 1.5

What this says is that the student was 1.5 standard deviations above the mean.

Using the Z-Table

14 A Z-table (also known as a Standard Normal Table) provides the probability that a standard normal random variable, Z, is less than or equal to -z or z.

−∞) up to the given Z-score.

For Z = 1.5, the standardized cumulative is about 0.9332. This means:

• Below 85 sees 93.32% (83 of them) of the scores

• The student is in the 93rd percentile

Example:

If you want to find the cumulative probability for Z = 1.53, from before by looking at the row that intersects 1.5 and column which has 0.03 → Value = 0.9370

Applications of Standard Normal Distribution

• Contrast of scores from several SD with the SDs compared to each other.

• Finding percentiles of individual scores

• Calculating probabilities for hypothesis testing

• Identifying outlier(s) by, for example Z-score other than ±2 or 3.

"Activity: Finding Probabilities Using Z-Scores"

Instruction to Student:

6 The scores on a standardized test are normally distributed with a mean of 500 and a standard deviation of 100.

1. Calculate the Z-scores for students who scored:

o a) 620

o b) 450

o c) 700

2. Use a Z-table (or standard normal distribution table) to find the probability of scoring:

- a) Less than 620
- b) More than 700
- c) Between 450 and 620

Write a short interpretation of each result. Discuss how Z-scores can help identify performance levels (e.g., above average, average, below average).

7.4.5 Applications and Examples of Normal Distribution

The normal distribution is widely used in theoretical statistics and real-world applications.

Applications:

Business and Finance:

o Stock price fluctuations,

o Demand forecasting,

o Return on investment (ROI) modelling.

Manufacturing and Quality Control:

o Parts measurement (diameter, weight),

o Tolerances and control charts.

Education:

o Standardize scores from an exam (i.e., SAT, IQ),

o Grading on a curve.

Healthcare:

o Blood pressure and cholesterol levels,

o Patient recovery times.

Social Sciences:

o Human behavior measurements,

o Population studies.

Example:

The heights of men in a city follow the normal distribution with mean 170 cm and standard deviation 6 cm for adult men.

• What is the proportion of men that are taller than 182 cm? First, calculate the Z-score:

Z = (182 − 170) ÷ 6 = 2.0

P(Z 2.0) = 1 − 0.9772 = 0.0228 or that men who are taller than 182 cm constitute, of course, about 2.28% of all men.

7.5 Summary

- This chapter discussed three of the most important proba- bility distri butions in statistics:
  - Binomial Model: Describes the number of successes that occur in a fixed number of independent trials, with each trial having exactly two possible outcomes. It is applicable for discrete data where the number of experiments is constant.
- Poisson Distribution is a good fit for modelling the number of events occurring within a fixed time or space period if these events happen with a known constant mean rate and independently. It is widely used in queues systems, traffic problems and quality control.
- The Normal Distribution A continuous, symmetrical distribution which is bell-shaped and commonly employed to represent data from nature. It can be characterized by its mean and standard deviation, and it is the basis of many statistical techniques such as hypothesis testing and confidence intervals.
- Each of these distributions comes with its own assumptions, formulas, and use cases – and knowing where and when to apply them is critical for successful data analysis and decision-making.

7.6 Key Terms

1. Probability Distribution - A function that predicts the probability of different results
2. Binomial Distribution - discrete distribution that represents the number of successes in n fixed trials
3. Poisson Distribution- The discrete probability distribution for the number of events in a fixed interval.
4. Normal Distribution – A continuous symmetric distribution that is bell-shaped
5. Probability Mass Function (PMF) - For a discrete distribution, this is a function that describes the probability of each specific value

6. PDF - Probability Density Function: For continuous distributions, a function whose integral over some interval is a probability.
7. Z-Score - How many standard deviations a datum is from the mean.
8. Mean (μ) - The average value of the distribution
9. Variance (σ^2) – Variability in a dataset as to how it is spread out.
10. Random Variable - one whose possible values can be subject to randomness (1)NORMAL DISTRIBUTION - a mathematical representation of frequency which is symmetrical above and below the mean, as indicated by the bell curve (2)Mean (μ) - average value of a set of numbers Variance Standard Deviation Variance © 2003 McGraw-Hill Ryerson Limited Page 859 All Rights Reserved Page:704 Standard Deviation Square root of variance; denotes consistency or fluctuation

7.7 Descriptive Questions

1. Explain the assumptions under which a binomial distribution is appropriate.
2. What is the relationship between Poisson and binomial distribution? Give an example.
3. Write down the formula for the probability mass function of binomial distribution.
4. Describe how the standard normal distribution is obtained and applied.
5. State the 68–95–99.7 rule and explain how it is related to the normal curve.
6. Compare and contrast the binomial and Poisson distributions. When would you use each?
7. Provide two real-life uses of the normal distribution in business or science?
8. A machine has a 90% chance of functioning properly. What is the chance of having 2 out of five defective items?
9. What are the complexities of human life when normal distribution is applied?
10. The number of customers arriving in an hour follows a Poisson process with an average of 4 arrivals. (I) What is the chance that exactly 2 people will arrive during a single hour?

7.8 References

1. Gupta, S. C. (2014). Fundamentals of Statistics. Himalaya Publishing House.

2. Levin, R. I., & Rubin, D. S. (2012). Statistics for Management. Pearson Education.

3. Sharma, J. K. (2018). Business Statistics. Vikas Publishing House.

4. Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2016). Statistics for Business and Economics. Cengage Learning.

5. UGC e-Pathshala. Modules on Probability Distributions and Statistical Theory.

6. MoSPI – Ministry of Statistics and Programme Implementation, Government of India.

7.9 Case Study

"Demand Forecasting with Distributions: Priya's Inventory Planning"

Background:

Priya is a supply chain manager of a grocery distribution company that provides perishable commodity to 60 retail stores. One of her assigned duties is to predict the need for specific items, such as milk packets, fruit and bread, that must be restocked each day. Even though she could rely on historical pricing and sales data, her team was plagued by understocking and overstocking.

She started looking at probability distribution models to refine the predictions.

•For products that are only sold or not sold (success/failure), such as an individual milk packet, she applied the binomial distribution to calculate how many outlets would sell out in a day.

• To model the number of surprise orders arriving during the day, she used a Poisson distribution because these arrivals were random and low in volume.

• For daily total sales volume that was linear with anomalies, she also used the normal distribution to calculate the mean demand and have safety stock because of standard deviations.

Problem description 1: Demand Patterns are not uniform Among the capacity planning factors, one issue is that of inconsistent demand profile.

Although all average sales were stable, stockouts were common at the various outlets.

Solution:

Priya applied binomial probability to predict when an outlet might sell out, and dynamically moved stock across stores according to the changing success probability.

Problem 2: The Confounding Same-Day Orders

Midday retailer order requests were sporadic and unpredictable.

Solution:

With the Poisson model, Priya calculated how many extra orders to expect in a given hour and updated delivery schedules accordingly.

Issue 3: Overestimation of the Safety Stock

The problem is that stockpiling becomes waste, and also inflates cost.

Solution:

Priya used a normal distribution to represent demand, and the forecast Z-score to determine how much safety stock they should order (~95% of the demand range).

MCQ:

Which type of distribution should Priya use to model the probability that 3 shops will sell out of stock on a day, if each has a 20% chance of selling out?

A) Poisson Distribution

B) Binomial Distribution

C) Normal Distribution

D) Uniform Distribution

Answer: B) Binomial Distribution

Conclusion:

By using the correct probability distribution to model a business situation, Priya managed to cut down wastage, increase her odds of having stock available and create an agile supply chain. What initially appeared as chaos magically became tractable through formal statistical modeling.

Statistics for Business Unit 8 V3.docx

Statistics for Business_BBA_2

Statistics for Business_BBA_2

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:127500307

Submission Date

Feb 4, 2026, 11:10 AM GMT+5:30

Download Date

Feb 4, 2026, 11:45 AM GMT+5:30

File Name

Statistics for Business Unit 8 V3.docx

File Size

183.9 KB

24 Pages

5,048 Words

29,103 Characters

3% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 15 words)

Match Groups

10 Not Cited or Quoted 3%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

2% Internet sources
1% Publications
3% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **10** Not Cited or Quoted 3%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

2% Internet sources
1% Publications
3% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	
repository.au.edu		<1%
2	Submitted works	
Myanmar Imperial College on 2025-07-04		<1%
3	Publication	
Tingliang Zhang, Haiwang Zhong, Zhenfei Tan, Xinfei Yan. "Misaka: Interactive S...		<1%
4	Internet	
www.upgrad.com		<1%
5	Submitted works	
University of Wales Swansea on 2022-11-25		<1%
6	Internet	
studyhippo.com		<1%
7	Submitted works	
Manipal University Jaipur Online on 2025-07-09		<1%
8	Submitted works	
University of Bolton on 2025-10-21		<1%
9	Submitted works	
University of Keele on 2025-11-13		<1%

Unit 8: Correlation

Learning Objectives

1. Define correlation and explain its significance in measuring the strength and direction of a relationship between two variables.

2. Identify and differentiate between various types of correlation (positive, negative, zero; linear and non- linear), using both graphical and numerical methods.

3. Interpret scatter diagrams and simple correlation graphs to visualize data patterns and preliminary relationships between variables.

4. Calculate Karl Pearson's Coefficient of Correlation (r) for both ungrouped and grouped data, and interpret the result in terms of strength and direction.

5. Understand the properties and limitations of the correlation coefficient, including its range, sensitivity to outliers, and lack of causality.

6. Apply Spearman's Rank Correlation method for ranked or ordinal data and evaluate correlation where data is not suitable for Pearson's method.

7. Use correlation analysis in real-world contexts such as economics, business, psychology, and social sciences to study relationships (e.g., income vs. expenditure, advertising vs. sales).

Content

8.0 Introductory Caselet

"Ravi's Research on Return Rates: Finding Relations in Unexpected Places"

Background:

Ravi is a data analyst working for an e-commerce company, and he was requested to research higher return rates among hands of goods. His early findings didn't point to a clear pattern — some customers bought back pricey electronics, others cheap clothing. The marketing committee suspected that ad campaigns and steep discounts might be driving everyone crazy, but they had no proof.

Ravi wondered whether the return rates were related to other observable variables and he investigated correlation. He began by gathering information on:

• Product return rate (%),

• Price of the product,

• Discount offered (%),

• Satisfaction score of customers (from feedback surveys).

He started by creating scatter plots and saw, in certain product categories, the higher the discount, the greater returned percentage. He then measured the strength of these associations by using Karl Pearson's coefficient (r). For some categories, such as Clothing and Accessories for which r >+0.75, there was a strong association between high discounts with return rates.

In order to examine the relation between customer satisfaction and returns, he employed Spearman's Rank Correlation and discovered a strong negative correlation ($r_s \approx -0.85$), which means that merchandise with low satisfaction scores were likely to be returned.

With the help of correlation tools, Ravi could now:

• Determine what marketing tactics were causing supersize returns,

• Suggest product quality improvement for products with low-ratings,

• Assist the company in minimizing its return fees all the while preserving customer loyalty.

What was once viewed as a potpourri of consumer behaviour became a kind of data-informed perception layered on top of correlation analysis.

Critical Thinking Question:

If you were Ravi and found strong correlation of discount to return rates in only one product category how would you guide the marketing team? What other factors could you look at to be sure that your recommendation is valid?

8.1 Introduction

Correlation is a method used in statistics which can be used to quantify and explain the strength and direction of an association between two variables. Two variables are said to be correlated when they have a tendency to move together in a predictable way. We can use correlation analysis to answer questions like:

• Do advertising expenses influence sales?

• Does employee satisfaction impact productivity?

• How are inflation and interest rates related?

It is also essential to keep in mind that correlation does not imply causation. It measures how closely two variables move in tandem, but does not suggest that one causes the other.

8.1.1 Meaning and Importance of Correlation

Meaning:

5 Correlation measures the strength of the linear association between two quantitative variables. It is measured using a

and is a value between −1 and +1, most often referred to as Karl Pearson's r.

• A perfect positive state is denoted by +1

• −1 denotes a perfect "negative" correlation

• 0 indicates no linear relationship

Importance:

Identifies Relationships:

Correlation is useful for determining the relationship between two variables (if any) and its direction.

Supports Decision-Making:

In business, an understanding of the relationship between figures such as sales and marketing spend can aid strategic planning.

Reduces Uncertainty:

When relationships are being understood, the guesswork is eradicated and data-driven insights follow.

Forms Basis for Further Analysis:

Correlation form the basis for more advanced statistics, such as regression.

8.1.2 Distinction between Correlation and Causation

There is all sorts of nonsense that can be made to appear serious with correlation instead of causation. This is not true.

Feature	Correlation	Causation
Definition	Measures degree of association	Indicates that one variable affects another
Direction	Can be positive, negative, or zero	Has a specific direction of influence
Proof Needed	No proof of cause-effect is needed	Requires experimental or theoretical proof
Example	Ice cream sales ↑ and drowning cases ↑	Ice cream doesn't cause drowning

Key Point:

It doesn't follow that if two variables are correlated, one causes the other. There might be a plight They might be Both respond to a

third (confounding) variable or be spurious.

Did You Know?

"The famous example of ice cream sales and drowning incidents often cited in statistics classes is used to explain why correlation ≠ causation. Although these two variables may show a strong positive correlation during summer months, one does not cause the other. Instead, a lurking variable—hot weather— influences both."

8.1.3 Applications of Correlation in Business and Economics

Correlation occurs in many areas of business and economic analysis. Some practical applications include:

In Business:

• Marketing Analysis: Knowing if increased marketing efforts correlate to greater sales.

• People Management: Researching the relationship between worker commitment and productivity.

• Operations Management: Investigating the connection between manufacturing volume and error rate.

In Economics:

• Macroeconomic Analysis - The relationship between inflation and interest rates, for example, or unemployment.

• Consumer Behavior: Understanding how income affects what consumers spend their money on.

• Stock Market Analysis:(This is done by the correlation between two stock returns to construct an investment portfolios) With the power of correlation, decision makers can make intelligent guesses about how they should allocate their resources and optimize strategic actions.

8.2 Types of Correlation

Depending on the direction, nature and number of variables involved the correlation can be categorized by several types. Knowing these differences aids in the choice of which method to use for analysis and how to interpret the results.

8.2.1 Positive and Negative Correlation

The connexions of which I am now speaking are all determined by the relation between one variable and another.

A. Positive Correlation

Two variables have a positive correlation when one increases as the other does, and likewise one decreases while the other decreases.

Examples:

• Height and weight

• Advertising spends and sales revenue

• Education level and income

Graphically: The data points in a scatter diagram move upward from left to right.

B. Negative Correlation

Two variables are negatively correlated if an increase in one result in a decrease of the other.

Examples:

• Price and quantity demanded (law of demand)

• Fuel efficiency and car weight

• Interest rate and investment levels

Graphically: The points move downward from left to right.

8.2.2 Linear and Non-linear Correlation

This is a categorization based upon the shape or form of the relationship.

A. Linear Correlation

A relationship is linear if a change in one variable yields a fixed/constant change in another. The relationship is linear and can be described by a straight line.

Example:

• Sales growing linearly with salespeople

Equation Form: y = a + bx

B. Non-linear (Curvilinear) Correlation

The relationship is non-linear when the change in one of the variables is not a constant multiple of changes in the other. The data points follow a curve.

Examples:

• Learning: Faster initially, but slower later fugure 6.learning speed increase during the inductive phase, and followed byi a plateau.

• Age vs Income relationship: Income may increase, then slow down after retirement

In graphics: Instead took the scatterplot shape being uneven (curved) instead of a straight line.

8.2.3 Simple, Partial, and Multiple Correlation

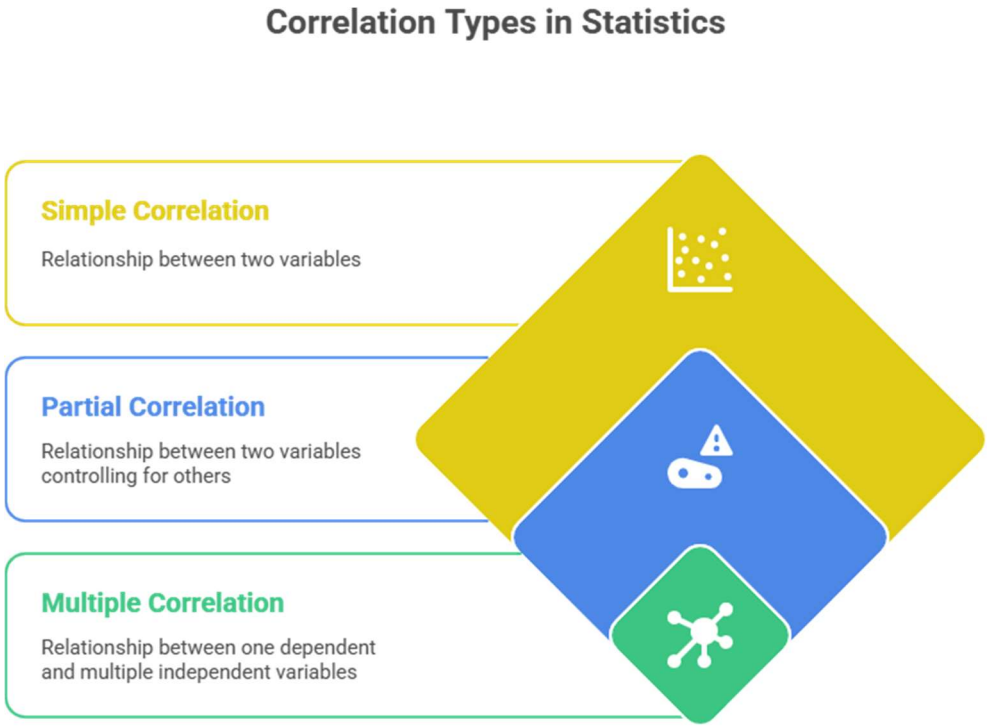


Fig.8.1 Simple, Partial, and Multiple Correlation

This classification is determined by the quantity of variables being investigated.

A. Simple Correlation

Implicates only two variables, examine how one independent variable relates to one dependent variable.

Example:

• Ratio between temperature and electric energy use

B. Partial Correlation

It's a technique that tests which resources explain the most amount of variance in another resource, when we control for one or more other resources.

Example:

• Analyzing the impact of how long a student studies on their grades, controlling for amount of sleep

When to use: We are trying to control for the effect of some variables, while studying the effect of others.

C. Multiple Correlation

Investigates how one dependent variable responds to multiple independent variables simultaneously.

Example:

• Estimating sales as a function of advertising spending and market size

• Comparing job performance across experience, educational attainment and age.

Math formula: $y = a + b_1x_1 + b_2x_2 + ... + b_nx_n$

Summary Table: Types of Correlation

Basis	Types	Example
Direction	Positive / Negative	Sales & Ads / Price & Demand
Form	Linear / Non-linear	Salary & Experience / Age & Income
No. of Variables	Simple / Partial / Multiple	Rainfall & Yield / Yield & (Rain, Fertilizer)

8.3 Scatter Diagram and Simple Graph

Graphical methods are always one's first exploration of the form of relationship among variables like these. Scatterplots and simple graphs also help to physically see if there is an association, and what type of (if any) relation is present.

8.3.1 Concept of Scatter Diagram

3 A scattergraph (also known as a scatter plot) is a type of display using Cartesian coordinates to display values for typically two variables for a set of data.

• Any point on the graphic corresponds to (x, y).

6 • The independent variable is on the X-axis, and the dependent variable is on the Y-axis.

Purpose:

• To identify patterns or relationships,

• To assess whether the trend is linear or not,

• To observe outliers or clusters.

8.3.2 Method of Plotting Scatter Diagram

Here is how to make a scatter graph:

Collect bivariate data: for example, marks of a set of students in two subjects.

Mark your axes, first variable on the X axis and second variable on the Y.

Select the appropriate scales: according to their magnitudes.

Plot the points: Plot a point at the intersection of each pair (x, y).

Consider the pattern: This is useful to see the kind of relationship.

8.3.3 Interpretation of Scatter Diagrams

From the configuration of the points, correlation can be understood as follows:

Example:

For example, if you plot the number of hours studied and scores on exams, and points tend to the right, we can say that there is a positive correlation between study hours and exam grades - more study hours tend to get better final scores.

8.3.4 Simple Graph Method of Studying Correlation

Paired values are plotted by scatter diagrams, whereas simple graphs plot two separate curves over the same axes to display movement with time or some other shared factor.

Steps:

Obtain time series data on two related variables (for example, sales at a weekly level and advertising expenses at monthly level).

Graph the two variables on the same scale, typically with respect to a common X-axis (time).

Plot on two lines or curves, one for each variable.

Compare the trends visually. If the curves rise or fall in tandem, it indicates that there is a positive correlation. With correlation bake test there is negative-if opposite or the same direction.

Advantages:

• Compare trends over time,

• May assist in pinpointing lagging relationships (e.g., the advertising effect on sales being delayed).

Comparison: Scatter Diagram vs. Simple Graph

Feature	Scatter Diagram	Simple Graph
Data Structure	Paired observations	Time-series or trend-based data
Visualization	Points on a graph	Curves or lines
Correlation Type	Linear / Non-linear	Directional movement
Use Case	Measuring degree of association	Comparing trends over time

8.4 Karl Pearson's Coefficient of Correlation

Karl Pearson's Coefficient of Correlation (r) Karl Pearson's r is the most common statistical method used to show both the strength and direction of a linear relationship. It is a quantitative technique that combines with graphical techniques such as the scatter plot.

8.4.1 Definition and Formula

Definition:

The Karl Pearson coefficient of correlation measures the strength that two variables X and Y. |r| −1 to +1.

• r = +1 Perfect positive linear relationship.

• -1 ≤ r < 0: Negative linear (but not perfect because it's less than 2) correlation • r = -1: Perfect negative linear correlation

• r = 0: No linear relationship exists between the variables.

Formula (for ungrouped data):

$r = \Sigma[(x - \bar{x})(y - \bar{y})] \div \sqrt{[\Sigma(x - \bar{x})^2 \times \Sigma(y - \bar{y})^2]}$ where • r is the correlation coefficient,• xi is each individual score of X,• yi is the corresponding individual score of Y,• n is the number of elements in a group or category, and• that bar thing over the X and Y means "average" (or "mean") so you calculate those first.

Or using raw scores:

(57.26) $r = n\Sigma xy - (\Sigma x)(\Sigma y) / \sqrt{[n\Sigma x^2 - (\Sigma x)^2] \times [n\Sigma y^2 - (\Sigma y)^2]}$

Where:

• x and y are single data points,

• $\bar{x}$ and $\bar{y}$ are the averages of X and Y,

• n is the number of observations.

8.4.2 Assumptions of Pearson's Correlation

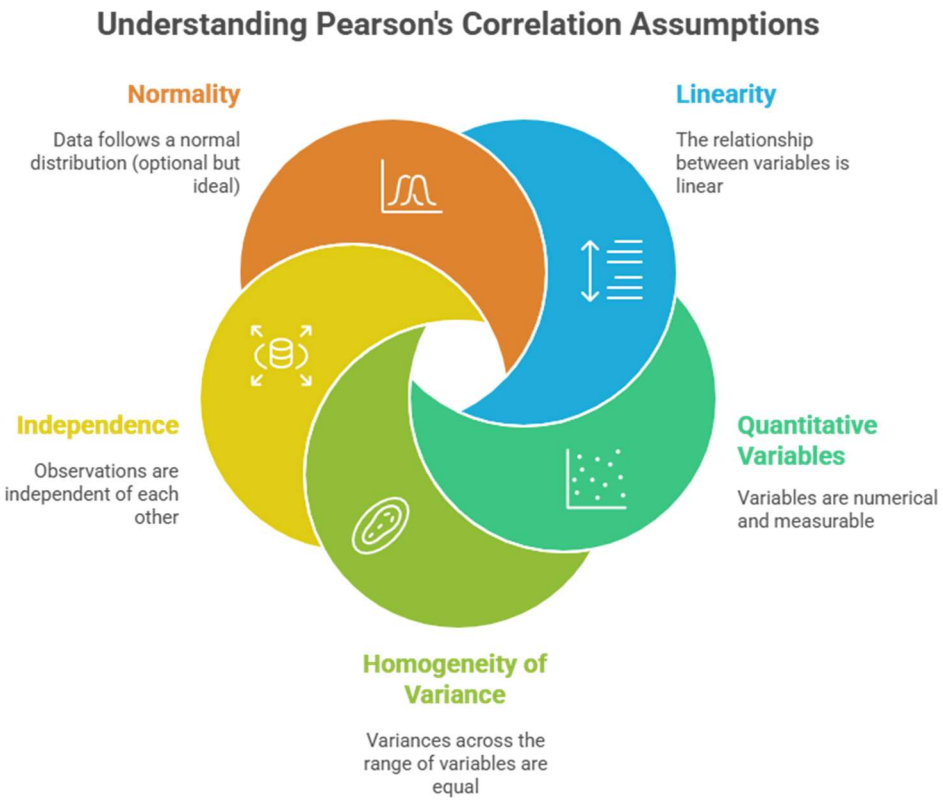


Fig.8.2. Assumptions of Pearson's Correlation

Assumptions of Pearson's Correlation For the use of Pearson's correlation as a valid and reliable statistical measure, three assumptions must be kept in mind: Data distribution.

Linearity: The two variables should be linear.

Quantitative Variables: The two variables must be numeric (interval or ratio scale).

Homogeneity of Variance: The variance of the values is similar in all parts of the range.

Independence: Observations should be independent of each other.

Normality (not required, but useful): both variables would be normal for inference.

8.4.3 Computation of Pearson's Correlation

A. For Ungrouped Data

Use the raw score formula:

Step-by-Step:

Find the sum of x,y, x^2, y^2 and xy.

Substitute into the formula:

$r = [n\Sigma xy - (\Sigma x)(\Sigma y)] / \sqrt{\{[n\Sigma x^2 - (\Sigma x)^2] \times [n\Sigma y^2 - (\Sigma y)^2]\}}$

Example:

For example if we have 5 students who got marks in Math (say X) :40, 50,60,70,80

and in Science (Y): 42, 49, 65, 68, 75

Calculate r by equation above (substituting actual numbers in calculation).

B. For Grouped Data

If the data is presented in terms of class intervals, you can employ the following formulae:

$r = \Sigma f(dx \times dy) \div \sqrt{[\Sigma f(dx)^2 \times \Sigma f(dy)^2]}$

Where:

- dx = deviation of X from assumed mean (A)/class width,
- dy = deviation of Y from assumed mean (B),
- f = frequency of class.

Steps:

Calculate midpoints, dx and dy,

Multiply dx and dy by f,

Compute all required sums,

Substitute into the formula.

Did You Know?

"If all data points lie exactly on a straight line with a positive slope, Karl Pearson's coefficient r becomes +1, indicating a perfect positive correlation. However, such perfect correlation is extremely rare in real-world business data, due to noise and variation in measurements."

8.4.4 Merits and Limitations of Pearson's Correlation

Merits:

Numerical Relationship: Reports a specific value which expresses strength and direction.

Mathematically tight: It is derived from an exact formula—the results are stable and repeatable.

Applicable Widely: Works across various disciplines – economics, psychology, business and so on.

Analysis Groundwork: Can be used for regression, prediction and testing hypotheses.

Limitations:

Ignores Non-linearity: It is not suitable to see the non-linear relations.

Influenced by Outliers: It is largely affected by the extreme values.

Only Measures Association, Not Causation A high correlation does not prove that one factor causes the other.

Interval or Ratio Data Required: Not applicable to ordinal/nominal variables.

Scale Sensitive: R is invariant to metric changes, but covariance isn't however r (although it may be changed in interpretation).

8.5 Properties of Coefficient of Correlation

The correlation coefficient(r), usually calculated by the method of Karl Pearson, is a statistical measure describing linear relationship between two variables. Although it is very simple to calculate, its meaning and proper use depend on its critical properties and limitations.

8.5.1 Range of Correlation Coefficient

The Pearson correlation coefficient's value is in the range:

−1 ≤ r ≤ +1

• r = +1: Positive linear correlation is perfect • r = 0: There is no tendency for a linear relationship between the variables.

• r = −1: Perfect negative linear relationship

• r = 0 : No linear correlation • -1 ≤ r &""" #$% &$!

These limits are useful to determine the strength of linear relationship between two variables.

8.5.2 Interpretation of Values of 'r'

The size and direction of the correlation is given by r.

+0.90 to +1.00	Very strong positive correlation
+0.70 to +0.89	Strong positive correlation
+0.40 to +0.69	Moderate positive correlation
+0.10 to +0.39	Weak positive correlation
0	No linear correlation
−0.10 to −0.39	Weak negative correlation
−0.40 to −0.69	Moderate negative correlation
−0.70 to −0.89	Strong negative correlation
−0.90 to −1.00	Very strong negative correlation

Note: These categories are guidelines. The meaning also depends on the context (social scientists, for example, may consider r = 0.3 as moderate).

8.5.3 Mathematical Properties of 'r'

Symmetry Property:

Correlation is symmetric in nature:

r(x, y) = r(y, x)

This means the correlation between X and Y is equal to the correlation of Y and X.

Unit-Free Measure:

The value of the coecient r does not depend on the unit with which the measurement is made, since it is a standardized value. Note that converting from kilograms to grams, etc., or from rupees to dollars does not change the value of r.

Covariance-Based:

r is calculated from the covariance and standard deviations:

$r = Cov(X, Y) \div (\sigma_x \times \sigma_y)$

Where Cov(X, Y) is the covariance between variables X and Y.

Linear Transformation Invariance:

If you transform the variables (e.g., to X' = aX + b), r doesn't change, as long as a > 0.

No Directional Causality:

The coefficient r only denotes association, it does not indicate direction or cause.

8.5.4 Common Misinterpretations

Despite being so simple, the correlation coefficient is also often mispoen or used inappropriately. Here are common pitfalls:

Correlation Implies Causation (False Assumption):

Correlation does not imply causation.

Example: There may be a positive relationship between ice cream sales and drowning deaths because it's hot outside, not because one is causing the other.

Zero Correlation Equals No Relationship (But There Are Exceptions!

r = 0 simply means no linear association between the two variables. Its relationship with fasting insulin is also likely to be non-linear.

Ignoring Outliers:

Correlation coefficient can be distorted substantially when we have a few extreme values, and these outliers would mask the true nature of the association.

Assuming Linearity Always Holds:

Pearson's r is based on a linear relationship. Applying it on non-linear data produces garbage.

Correlation vs. Agreement:

High correlation does not imply good agreement. For instance, two measuring instruments could have a high r value but different mean values.

8.6 Spearman's Rank Correlation

The Spearman Rank Correlation Coefficient is a nonparametric measure of the magnitude and direction of association between two ranked variables. It is particularly helpful in the presence of ordinal data, or when assumptions for Pearson's correlation (i. e., linearity, normality) are violated.

8.6.1 Concept and Formula

Concept:

The Spearman's Rank Correlation (ρ or sometimes denoted as r_s) gives us an idea of how the relationship between two variables can be characterised by a monotonically increasing or decreasing form (it doesn't have to be linear!).

It measures the count of ranks (positions) but not directly as absolute values.

The formula (in case there are no tied ranks) is:

$r_s = 1 - [6 \times \Sigma d^2 \div n(n^2 - 1)]$

Where:

- r_s = Spearman rank correlation coefficient
- d = absolute value of the difference between the ranks of each pair
- n = number of observations

Step-by-step:

Then rank order the scores for each variable (lowest = 1) individually.

Calculate the rank difference d.

Square each d to get d^2.

Sum up all d^2 values.

Plug into the formula.

Result range:

- +1: Perfect positive rank correlation
- −1: Perfect negative rank correlation
- 0: No correlation in ranks

"Activity: Spearman's Rank Correlation in Customer Preferences"

Instruction to the Student:

Imagine you surveyed 8 customers to rank five product features (Price, Quality, Durability, Packaging, Brand) in order of preference.

- Collect rankings from two customers.
- Assign ranks for each feature and compute d and d^2.
- Apply the Spearman's Rank Correlation formula:

$r_s = 1 - [6\Sigma d^2 \div n(n^2 - 1)]$

- Interpret the result: Does one customer's preferences align with the others?

Prepare a brief note summarizing your correlation value and explaining what it reveals about customer behavior similarity.

8.6.2 Tied Ranks and Adjustments

The rank is tied if there are two or more identical values in the data set. So, instead of each having its own rank, they are all ranked at their average position.

Example:

Now if three students are at 2nd with same score, we need to average rank 2,3 and 4:

Rank assigned = (2 + 3 + 4)/3 = 3

Adjustment in formula:

The former formula remains the same, but since we have to correct the ranks for ties before computing the d's and d^2's.

Did You Know?

"In Spearman's rank correlation, tied ranks are not an error—they are a natural part of ordinal data. The formula remains valid if you assign the average of the ranks for tied observations. Ignoring ties or assigning them arbitrarily can significantly distort the correlation result."

8.6.3 Advantages of Rank Correlation

No Assumption of Normality:

Even when the underlying data is not normally-distributed, the observation distribution under focus may be less sensitive as well.

Handles Non-linear Relationships:

Appropriate when the relationship is not linear, but monotonic.

Ordinal Data Friendly:

Applicable on rank/preference based, rather than numerical data.

Resistant to Outliers:

It is insensitive to the distribution of variable and does not normality assumption." Since it uses ranks, the outliers have smaller effect as with Pearson's r.

Computational Feasibility for Small Data Sets:

Particularly appropriate for contests, polls, or preference ratings.

8.6.4 Limitations of Rank Correlation

Less Precise with Numerical Data:

When real numbers are present and conditions are satisfied, the Pearson's correlation gives an accurate measurement.

Ineffective for High Tied Data Sets:

Too many ranks tied can lead to distorted correlation strength or complicated adjustments.

Cannot Detect Linear Strength:

It looks for monotonic trends, but does not quantify how strong the linear relationship is.

Not suitable for analysis of interval/ratio scale data:

Rank correlation is less satisfactory than numerical correlation for statistical modeling or prediction.

Summary Table: Spearman's vs. Pearson's

Feature	**Pearson's r**	**Spearman's r_s**
Data Type	Interval/Ratio	Ordinal or Ranked
Assumes Normality?	Yes	No
Handles Non-linear Trends?	No	Yes (monotonic)
Impact of Outliers	High	Low
Measures	Linear correlation	Rank correlation

Knowledge Check 1

Choose the correct option:

1. If the value of Karl Pearson's correlation coefficient (r) is −0.95, it indicates:

A) No correlation

B) Weak positive correlation

C) Strong negative correlation

D) Perfect correlation

2. A scatter diagram shows points sloping downward from left to right. What type of correlation is indicated?

A) No correlation

B) Positive correlation

C) Perfect correlation

D) Negative correlation

3. Which of the following is true about Spearman's Rank Correlation?

A) It assumes a linear relationship

B) It cannot handle tied ranks

C) It is based on data values, not ranks

D) It is useful for ordinal or preference data

4. The correlation coefficient is always:

A) Greater than or equal to 0

B) Between −1 and +1

C) Equal to or more than 1

D) Less than −1

5. Which of the following statements is incorrect?

A) Correlation measures association, not causation

B) A high correlation always implies a strong cause-effect relationship

C) A zero correlation means no linear relationship

D) Karl Pearson's correlation is sensitive to outliers

8.7 Summary

- In this chapter, we discussed correlation which shows the strength and direction of association of two variables.
- We started with the elementary meaning of correlation, and how it differs from "causation."

- Various forms of correlation were mentioned: positive/negative, linear/non-linear and simple, partial and multiple correlation.
- We learnt about how we can visually perceive the presence and magnitude of relationships through scatter plots and basic graphs.
- Karl Pearson's Coefficient of Correlation (r) was presented as a measure of the magnitude of r, and its formula, assumptions, and computation were shown.
- We examined Pearson r, including its characteristics and weaknesses.
- Lastly, we looked at Spearman's Rank Correlation that is a non-parametric measure appropriate for ordinal or ranked data, and how to deal with tied ranks.
- These four tools, together, provide a powerful collection of techniques for exploring relationships in statistical and business data that are indispensable for estimation, choice making, prediction and pattern identification.

8.8 Key Terms

1. Correlation - A statistic that demonstrates how two variables move in relation to each other
2. Positive Correlation- Both increase or decrease as together .
3. Negative Correlation- one variable increases and the other decreases.
4. Linear Correlation – A link between two data in which one varies proportionately to the other
5. Karl Pearson's Coefficient (r) - It is a numerical value representing the degree of linear correlation, ranging from −1 to +1.
6. Spearman's Rank Correlation - Monotonic relationship between ranked datasets
7. Scatter Diagram - A graph used to display the relationship between two variables
8. Ranks do not differ, how to handle tied ranks in Mann Whitney U What is the appropriate method when there are ties for detecting differences.
9. Partial-cornet°0 Relationship between two variables after elimination of the third variable.
10. Multiple Correlation: One dependent variable with two or more independent variables.

8.9 Descriptive Questions

1. What is correlation and why is it important in business and economics?
2. Distinguish between positive and negative correlation with appropriate examples.
3. What are differences between linear and non linear co-relation?
4. Describe the process for constructing a scatter plot. What can it tell us?
5. Derive the formula for Karl Pearson's coefficient of correlation and explain the terms.
6. What are the assumptions that should be met to use Pearson correlation method?

7. Describe the advantages and disadvantages of Karl Pearson's coefficient of correlation.
8. Define the term rank correlation. How does it help me when there are no numbers to compare?
9. How do you handle “tied” ranks in Spearman rank correlation?
10. Compare Pearson correlation with Spearman rank-based correlation. When is each method appropriate?

8.10 References

1. Gupta, S. P. (2014). Statistical Methods. Sultan Chand & Sons.
2. Sharma, J. K. (2018). Business Statistics. Vikas Publishing House.
3. Levin, R. I., & Rubin, D. S. (2013). Statistics for Management. Pearson Education.
4. Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2016). Statistics for Business and Economics. Cengage Learning.
5. UGC e-Pathshala Modules – Statistical Analysis
6. MoSPI (Government of India) – National Statistics Handbook

Answers to Knowledge Check

Knowledge Check 1

1. C) Strong negative correlation
2. D) Negative correlation
3. D) It is useful for ordinal or preference data
4. B) Between −1 and +1
5. B) A high correlation always implies a strong cause-effect relationship

8.11 Case Study

“Relation Study between Sales and Advertising Spend”

Introduction

In this competitive market, businesses are in relentless pursuit for knowledge of determining what drives their sales. A relationship that is often presumed is an advertising and sales revenue relationship. And marketing departments regularly demand more money, since the one leads to the other. Yet for business leaders to have the ability to act, they need the data rather than assumptions.

This caselet can be used to understand the application of correlation analysis (that is, Karl Pearson's and Spearman's), in assessing industry demand relationship between advertising spend and sales. It also covers graphical aids such as scatter plots and clarifies the effect of a misreading of correlation on business strategy directed by factors that are secondary in causation.

Background

Take an example of a mid-sized retail company having presence in five metro cities. The marketing team has increased that ad spend over the last six months, which we expected to lead to a sales lift. But in certain cities, sales did get better, and some others just didn't change all that much despite increased advertising expenses.

The regional manager, Priya, decided to investigate further. (For the following, she pulled monthly figures from each of the five cities for:)

- How much is the spend on digital ads (₹)
- Monthly sales revenue (₹)

She then made a scatter graph of the data, which suggested a rising trend in three cities and either messiness or decline in the others. This led her to compute the coefficient of correlation, Karl Pearson's coefficient (r), to determine the extent of linear association between the two variables.

The results:

- City A: r = +0.87 (strong positive)
- City B: r = +0.65 (fair positive)
- City C: r = −0.10 (very weak negative)
- City D: r = +0.02 (no linear relationship)
- City E: r = +0.90 (positively strong)

She also used Spearman's Rank Correlation for customer satisfaction rankings and repeat purchases to determine patterns between ranking variables.

1: Misunderstanding Zero Correlation The first problem is the one involving confusion about zero correlation.

Many departments operated under the assumption that a correlation of almost zero meant no relationship at all. In city D, the ad spend wasn't linear with sales, however the customer footfall went up.

Solution:

The team was trained on the distinction non-linearity vs. no relationship between the variables. A scatter plot was useful in indicating that a non-linear relationship could still be present. They had explored other models like polynomial regression and concentrated on engagement metrics.

MCQ:

What does it mean when r = 0?

A no relationship of any type

B) There is no linear relationship

C) The variables are independent

D) A causal relationship exists

Answer: B) No linear relation Explanation:- From the given data 'x' and 'y' are independent.

Problem 2 : Treatment of Ties in the Spearman ' s Method

When we collected customer feedback with star ratings, several customers assigned the same rating, and this resulted in tied ranks of sameness in our data.

Solution:

The data were ranked in the original order of occurrence with average ranks for ties prior to calculation of Spearman's coefficient. This ensured that the analysis captured monotonic relationships between satisfaction and repeat purchasing.

MCQ:

What do the tied ranks mean in Spearman method?

A) Ties are ignored

B) Assign the lowest rank

C) Increment the average of tied ranks .

D) Assign the highest rank

Response: C) Assign average of tied ranks

Problem 3: Correlation = Causation! Here's a challenge for you today.

Some teams pointed to a high correlation to vindicate marketing campaigns that they believed had caused sales to rise.

Solution:

Priya's team trained teams on the difference between correlation and causation, displaying examples in which a third factor (seasonal demand or competitor discounting) influenced both ad spend and sales.

MCQ:

Which one of the following statements is correct about correlation?

A) Correlation always implies causation

B) Science correlation strength and direction of relationship

C) There is strong causal relationship between the two variables if r value is high

D) Wiggling even when unrelated – Zero correlation indicates that the variables have nothing to do with each other.

Answer: B) Correlation reflects strength between two variables as well as direction of relationship.

Conclusion

Using statistical measures like Pearson's and Spearman's correlation coefficients, the company got a clearer sense of where ad spending mattered and in what places other factors were at play. It also underscored the value of visual tools such as scatter diagrams, and of statistical literacy in order to understand data properly. Understanding the reality behind 'correlation =/= causation' can help companies make smarter strategic choices - and not to draw bogus conclusions.

Statistics for Business Unit 9 V3.docx

Statistics for Business_BBA_2

Statistics for Business_BBA_2

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:127504309

Submission Date

Feb 4, 2026, 12:00 PM GMT+5:30

Download Date

Feb 4, 2026, 12:02 PM GMT+5:30

File Name

Statistics for Business Unit 9 V3.docx

File Size

148.4 KB

26 Pages

5,384 Words

29,444 Characters

4% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 15 words)

Match Groups

11 Not Cited or Quoted 4%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

2% Internet sources
0% Publications
3% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **11** Not Cited or Quoted 4%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

2% Internet sources
0% Publications
3% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Submitted works	University of Glamorgan on 2023-04-09 <1%
2	Internet	youthforpakistan.org <1%
3	Internet	www.coursehero.com <1%
4	Submitted works	Imperial College of Science, Technology and Medicine on 2024-12-12 <1%
5	Submitted works	Kaplan International Colleges on 2025-06-29 <1%
6	Internet	brightideas.houstontx.gov <1%
7	Submitted works	Myanmar Imperial University on 2025-08-28 <1%
8	Internet	www.teachmint.com <1%
9	Submitted works	Institute Of Business Management & Research, IPS on 2025-11-21 <1%
10	Submitted works	Myanmar Imperial University on 2025-10-30 <1%

Unit 9: Regression

Learning Objectives

1. Define regression and understand its importance in predicting the value of one variable based on another.

2. Differentiate between correlation and regression, especially in terms of direction and application.

3. Identify various types of regression such as simple, multiple, linear, and non-linear regression, and recognize their appropriate use cases.

4. Understand and apply the algebraic methods used to study regression, including the least squares method.

5. Construct regression lines (Y on X and X on Y) and interpret their significance in a real-world dataset.

6. Calculate regression coefficients and understand their meaning in terms of the rate of change and the direction of relationship.

7. Explain the mathematical properties of regression lines, such as how they intersect and how they relate to correlation coefficients.

Content

9.0 Introductory Caselet

"Mira's Marketing Spend: Predicting ROI with Regression"

Background:

Mira is the marketing head of a rapidly growing online grocery startup in three metro cities. Her team is frequently spending money on various channels — Google Ads, Instagram promotions and local influencer campaigns. Having a tight marketing budget, Mira found that the ROI fluctuated from month to month with zero predictability. There were months with huge increases in sales, and others that barely broke even.

In order to reconcile this inconsistency, Mira turned to a data analyst who familiarized her with regression analysis. Correlation would only tell her whether there's a strong relationship, but via regression, she could actually see how much each service drives total monthly sales and forecast them into the future.

In the last 10 months, Mira's team collected the following data:

• Cost of Google Ads (₹),

• Number of influencer posts,

• Engagement rate on Instagram,

• Total monthly sales (₹).

The analyst fitted a multiple linear regression model to this dataset with sales as the dependent variable and all the other factors were independent predictors. As a result, this equation allowed Mira to calculate what change in ad strategy might mean for sales results.

When Porter estimated from the regression model, she discovered:

• Google Ads had the best positive change on monthly sales,

• Influencer posts worked — but only up to a point,

• Engagement on Instagram had a somewhat less strong, but still positive, correlation with sales.

Now, Mira can:

• Predict sales for planned ad spend,

• Better spend budgets between channels,

• Justify marketing decisions with data.

A marketing strategy of trial and error became one of prediction through data, purely because Mira had used regression analysis on her business data.

Critical Thinking Question:

FocusingDo you think that if you were in Mira's shoes, and needed to eliminate a variable from the model (to minimize campaign costs), that (given these regression coefficients and R^2 's) it would be pretty clear which one to nuke? What might you look for when dropping a variable from a multiple regression model?

9.1 Introduction

One of the most powerful statistical tools for analyzing relationships between variables is regression analysis. Correlation, which is a degree of relationship between two variables Regression, takes it one step further and attempts to estimate or predict the value of one variable based on the other. This is why computing the regression model is particularly useful for predictions, decision making processes and business modelling.

9.1.1 Meaning and Importance of Regression

Meaning:

Regression is a statistical method to analyze the relationship between multiple variables. In its simplest form,

regression analysis investigates the impact of an independent (X) on a dependent variable (Y). It helps answer questions like:

- How much will sales rise if advertising spend is increased?
- What is the impact of education level on income?
- What is the anticipated production cost upon production volume?

Importance:

Prediction: A regression is used to predict the future results from prior data.

Quantitative Relationship: It is an equation which allows us to estimate one variable via another.

Business Decision-Support: Pricing model, budgeting and forecasting, risk analysis.

Learning Dependencies: Facilitates in spotting cause-effect-like dependencies under controlled conditions.

9.1.2 Distinction between Correlation and Regression

Although correlation and regression treat the relationships between variables, they have far different objectives and inferences.

Feature	Correlation	Regression
Purpose	Measures **degree of association**	**Estimates or predicts** one variable based on ano
Symmetry	r(x, y) = r(y, x)	Regression of Y on X ≠ Regression of X on Y
Direction	No direction implied	One variable is **dependent**, the other is **independ**
Output	One single value (correlation coefficient)	Regression **equation or line**
Use Case	Understanding strength of association	Forecasting and cause-effect modeling

Example:

- Correlation indicates that advertising and sales are related.
- Regression provides us with identity: X = a + b × Y Spend.

Did You Know?

"Correlation tells you how strong the relationship is between two variables, but regression tells you how much one variable changes with another. In other words, correlation is about association, while regression is about prediction."

9.1.3 Applications of Regression in Business and Economics

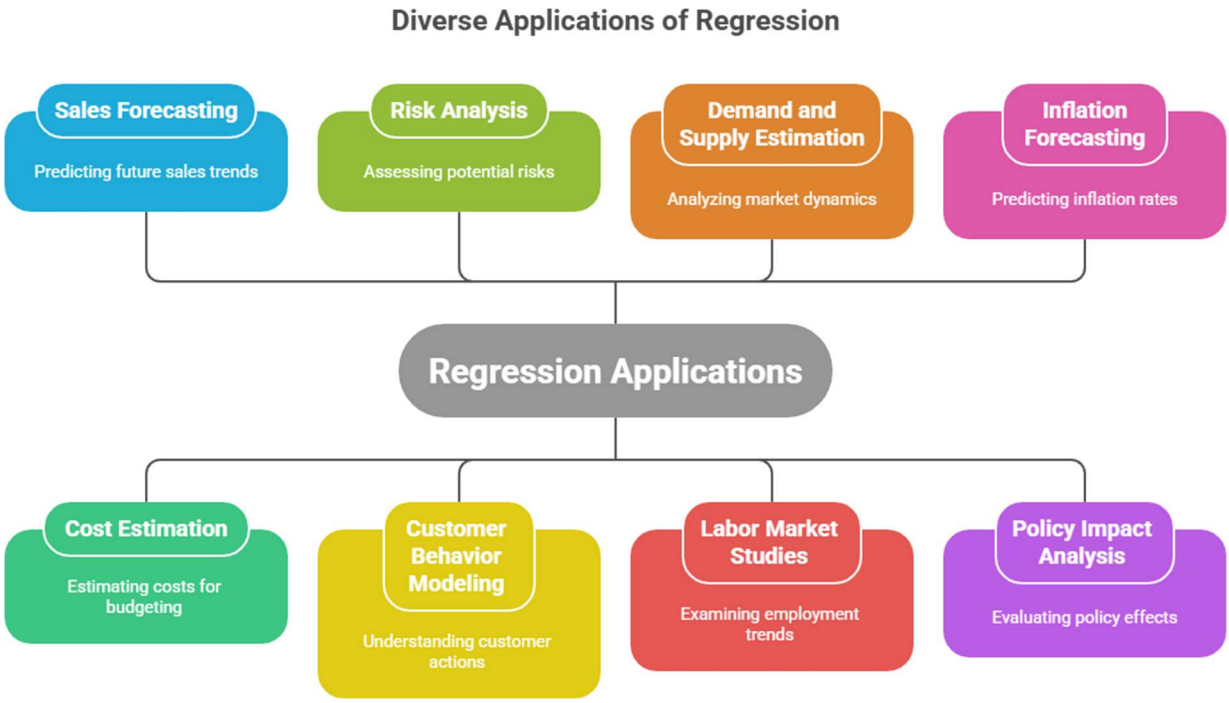


Fig.9.1. Applications of Regression in Business and Economics

Regression is commonly applied in business analytics, economics, and management sciences.

Applications in Business:

• Sales Forecast – How do we predict our future sales using what we sold in the past and how much money we spent on marketing.

• Cost estimation: Production cost estimates at different levels of output.

• Risk Analysis: The relation between risk exposure of different factors and financial returns.

• Customer Behaviour Modelling - Predict product purchase with the base of demographic and income.

Applications in Economics:

• Demand and Supply Analysis: How a product's price influences the quantity of it desired.

• Studies on the Labor Market: Researching the role of education and skills in determining wage.

• Inflation Expectations: Estimation of future inflation using interest rates and the value of the currency.

• Policy Impact Analysis: Evaluation of economic impact taxation or subsidies

9.2 Types of Regression

Depending on the number of variables and their relationships to each other, different forms of regression analysis are applied. The principal are the simple, multiple, linear and non-linear regression. Knowing these kinds can aid in choosing the right model for a particular data set and problem.

9.2.1 Simple Regression

Definition:

Simple Regression Analysis In simple regression (or bivariate regression) there are two variables, one dependent variable (Y) and one independent variable (X).

dependent variable (Y). This is a Y using X problem.

Equation:

8

Y = a + bX

Where:

• Y = Dependent variable

• X = Independent variable

• a = Intercept

• b = Regression coefficient (slope)

Example:

Predicting Y based on the value of X.

Use Case:

When you have a single predictor for the predicted variable, e.g., predicting profit based on sales.

“Activity: Forecasting Sales Using Simple Regression”

Title: "Predicting Next Month’s Sales from Advertising Spend"

Instruction to the student:

You are provided with monthly data for the past 8 months, including total advertising spend (₹ in lakhs)

and corresponding sales revenue (₹ in lakhs).

- Use the simple linear regression formula to compute the regression line:

Y = a + bX, where Y = Sales and X = Ad Spend.

- Calculate the values of a and b using the least squares method.
- Based on your model, predict the sales revenue if next month's ad spend is ₹5.5 lakhs.
- Submit your regression equation and prediction in a short report.

9.2.2 Multiple Regression

Definition:

Multiple regression, the precursor to path analysis is where you have more than one predictor of a single criterion. It is applicable when a series of factors combine to affect an outcome.

Equation:

$Y = a + b_1X_1 + b_2X_2 + \ldots b_nX_n$ Where:

Where:

• $X_1, X_2, \ldots X_n$: Numerous input variables

• $b_1, b_2, \ldots b_n$ = Regression coefficients Here is a linear Generally the slope factor of simple linear The formula for calculating SS regression involving two variables x y and 'n' data points is ∑_ (i=1)^n ((X_ i-¯X)((Y_i-¯y))))/ ∑_(i= 1)^n ×² Where n = number of pa… How to perform nonlinear regression in R?

Example:

You are interested in predicting the price of a house (Y) with size (X_1), location rating (X_2), and number of bedrooms (X_3) as independent variables.

Use Case:

Applied in marketing (predicting sales with price, promotion or advertising level), HR (predicting employee performance), finance (predicting risk).

9.2.3 Linear Regression

Definition:

In linear regression, it is assumed that the dependent and independent variables follow a linear relationship. This works for both the simple and the multiple regression so long as the relationship is linear.

Equation (Simple Linear):

$Y = a + bX$

Equation (Multiple Linear):

$Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$

Graph: A straight line in 2-d (for the simple linear) or a plane/ hyperplane in higher dimensions.

Use Case:

Handy if data demonstrates constant rate of change (such as predicting monthly utility bill based on usage).

9.2.4 Non-linear Regression

Definition:

Non-linear regression represents those cases when there is something that prevents the relationship between Y and X via a straight line. This curve of regression can be exponential, logarithmic or polynomial etc.

Example Equations:

- $Y = a \times e^{\wedge}(bX)$ (Exponential)
- $Y = a + bX + cX^2$ (Quadratic)

Example:

A learning curve for a new hire, in which the employee does better at a faster rate in the beginning and then improves more slowly.

Use Case:

Relevant to biology (growth models), economics (diminishing returns) and marketing (response to further advertising).

Summary Table: Comparison of Regression Types

Type of Regression	No. of Independent Variables	Relationship Form	Common Use Case
Simple Regression	One	Linear	Sales prediction from marketing spend
Multiple Regression	Two or more	Linear	House price prediction using multiple factors
Linear Regression	One or more	Linear	Predictive analytics in most domains
Non-linear Regression	One or more	Curved/Complex	Customer behavior modeling, biological growth

9.3 Methods of Studying Regression

Regression analysis is the process of finding such best fit relation between two or more variables. Techniques for the examination of regression are set into either the visual or computational domain. This section presents two of the most widely used methods, the scatter diagram approach and least squares method as well as a brief discussion on their assumptions.

9.3.1 Scatter Diagram Method

Definition:

A scatter plot is a visual tool designed to show the relationship between two data sets. It is the most elementary method of determining a linear and non-linear trend.

Steps to Draw:

(x,y).

Look at the big picture these points provide.

On the horizontal axis, plot X and on the vertical axis, plot Y.

For each value pair, plot a point

Interpretation:

• Upward trend → Positive correlation

• Downward trend → Negative correlation

• Nothing in particular → Nothing to see

• Points lie in straight line → Strong linear relationship

• Clubbing or curving of the points → Non-linear correlation may be present

Use Case:

As an exploratory tool to determine if regression analysis should be employed and (if so) what kind of model.

9.3.2 Method of Least Squares

Definition:

The least squares method is a mathematical procedure for finding the best-fitting curve to a given set of points by minimizing the sum of the squares of the vertical deviations (errors) from the points to the curve.

Objective:

To find a line of the form:

$Y = a + bX$

Where:

• a = Y-intercept

• b = Slope of the regression line.

• The line reduces the value of: $\Sigma(Y - \hat{Y})^2$, where $\hat{Y}$ = estimated value

Formulas:

• $b = [n\Sigma XY - (\Sigma X)(\Sigma Y)] \div [n\Sigma X^2 - (\Sigma X)^2]$

• $a = \bar{Y} - b\bar{X}$ Steps:

Compute ΣX, ΣY, ΣXY, and ΣX^2.

Calculate b (slope).

Calculate a (intercept).

Construct the regression equation.

Use Case:

This one is really popular in economics, business forecast and perform scheduling-of-production whenever a predictive model of a categorical response based on the previous data is used.

9.3.3 Assumptions Underlying Regression Analysis

The validity and reliability of the final results are dependent on several underlying assumptions of the regression analysis.

Assumption	Explanation
Linearity	The relationship between the dependent and independent variable is linear.
Independence	Observations are independent of one another.
Homoscedasticity	The variance of errors (residuals) is constant across all levels of X.
Normality of Errors	The residuals (errors) are normally distributed.
No Multicollinearity	In multiple regression, independent variables are not highly correlated.

Importance of Assumptions:

• Violations may cause the estimates to be biased or misleading such as ignoring error that is not negligible because it underestimates error, ignoring significance that is unjustified since it overestimates significance.

• Assumptions should be verified using diagnostics such as residual plots and statistical tests

9.4 Lines of Regression

If it's not clear, the regression line is the line that "fits" the scatter plot of points we obtained. There are two lines:

• One that predicts Y from any X value in this model: the regression of Y on X

• The one that predicts X on a given Y, known as the regression line of X on Y

Each of these lines minimizes the sum of paired deviations in dependent variable that is being predicted.

9.4.1 Regression Line of X on Y

Objective:

This line helps to predict X (predictor) from given values of Y (outcome).

Equation format:

$X = a + bY$

Where:

- X = Estimated value of the predictor
- Y = Value of the dependent variable that is known
- a = Intercept
- b = Regression coefficient of X with respect to Y

Formula for slope (b):

$b_{xy} = r \times (\sigma_x \div \sigma_y)$

Here:

- r = Pearson correlation coefficient
- σ_x = Standard deviation of X
- σ_y = Standard deviation of Y

Interpretation:

This very rare line is only useful if you happen to have a specific interest in predict X from Y (e.g. predicting ad budget limiting sales target).

9.4.2 Regression Line of Y on X

Purpose:

This is the most frequently employed line, it helps determine Y (dependent variable) for a given X.

Equation format:

5

$Y = a + bX$

Where:

- Y = Predicted value of the dependent variable
- X= Value of the independent variable known or given
- a = Intercept

• b = Regression coefficient of Y on X If Subscribe to view the full document.

Formula for slope (b):

$b_{yx} = r \times (\sigma_y \div \sigma_x)$

Interpretation:

We're dealing with this line when we are interested in prediction, for example predicting sales from advertising or forecasting expenses from output.

9.4.3 Properties of Regression Lines

Two Lines Exist:

There are two distinct regression lines if the correlation is imperfect ($r \neq \pm 1$), while in case of perfect correlation ($r = \pm 1$) both regression lines

coincide.

Intersection at Mean Point:

The two regression lines will intersect at all times at $(\bar{X}, \bar{Y})$ which is the means of the X and Y.

Relation with Correlation Coefficient:

The project of the two with regression coefficients equals the square of correlation coefficient:

$b_{yx} \times b_{xy} = r^2$

Directionality Matters:

The Y on X regression line is not the same as the X on Y one; they do different things!

Least Squares Criterion:

Every line minimizes the sum of square deviation in its own predicted variable (X or Y).

Did You Know?

"The two regression lines—Y on X and X on Y—always intersect at the mean point $(\bar{X}, \bar{Y})$ of the data. This is true regardless of the direction or strength of the correlation."

9.4.4 Geometric Representation of Regression Lines

With a scatter plot, the two regression lines form best-fit descriptions of the trend of the data.

Visual Representation:

• Y o X line minimizes the vertical deviations $(Y- \hat{Y})^2$.

• The line of X on Y is the best fit with respect to horizontally non-offsetting distortions as measured by $(X-\hat{X})2$.

• Both lines pass through the point $(\bar{X}, \bar{Y})$ which is the mean of X and Y.

• The two lines are separated by an angle that accommodates the degree of correlation:

o If r = ±1, both the lines coincide to become a single straight line.

o If r = 0, they intersect at 90°.

Summary Table: Regression Line Comparison

Feature	Y on X	X on Y
Dependent Variable	Y	X
Independent Variable	X	Y
Equation Format	$Y = a + bX$	$X = a + bY$
Minimizes Deviation In	Y (vertical distances)	X (horizontal distances)
Slope Formula	$b = r \times (\sigma_y \div \sigma_x)$	$b = r \times (\sigma_x \div \sigma_y)$
Common Use	Forecasting, prediction models	Estimating X from known Y

9.5 Regression Coefficients

Regression coefficients are key to knowing the relationship between variables in a regression analysis. These variables measure how much the dependent variable rises when the independent variable increases by one.

9.5.1 Concept and Calculation of Regression Coefficients

and â is the coefficient that measures how much does change in Y for every one unit of change of X, other things remaining constant.

In ordinary linear regression, typically, we take:

• b_{yx}: Regression coefficient for Y on X

• b_{xy} : The regression coefficient of X on Y

The same coefficients occur in the relations:

• $Y = a + b_{yx}X$

• $X = a + b_{xy}Y$

Formulas:

• $b_{yx} = r \times (\sigma_y \div \sigma_x)$

• $b_{xy} = r \times (\sigma_x \div \sigma_y)$

Where:

• r = Pearson's correlation coefficient

• σ_x, σ_y = The standard deviation of X and Y respectively or otherwise, mathematically speaking using raw data:

• $b_{yx} = \Sigma(xy) \div \Sigma(x^2)$

• $b_{xy} = \Sigma(xy) \div \Sigma(y^2)$

9.5.2 Relationship between Correlation Coefficient and Regression Coefficients

The correlation coefficient (r) and the two regression coefficients (b_{yx} and b^{y}_{x}) are mathematically related.

b_{xy}):

$r = \sqrt{(b_{yx} \times b_{xy})}$

Or:

$b_{yx} \times b_{xy} = r^2$

Interpretation:

•If the two regression coefficients are both negative, then r will be negative.

• If both coefficients of regression are negative, r is also negative.

•If one is positive and the other negative, there's an inconsistent relationship which shouldn't generally arise in well-formed data.

9.5.3 Properties of Regression Coefficients

Two Coefficients:

Between any two variables, there are 2 regression coefficients: b_{yx} and b_{xy}.

Signs Match Correlation Coefficient:

Both regression coefficients are of the same sign as r.

Geometric Mean Relationship:

The two coefficients multiplied equal to the correlation coefficient squared:

$b_{yx} \times b_{xy} = r^2$

Unit Sensitivity:

Regression coefficients are not unit-free. Their numerical values become different when the units with which the two variables are measured, say from kg to g.

No Fixed Range:

While correlation lies between −1 and +1, regression coefficients may or may not be on this interval.

(positive or negative).

When r = ±1:

The two lines of regression are superimposed, and both the regression coefficients equal the ratio of the standard deviations:

o $b_{yx} = \sigma_y \div \sigma_x$ (for r = +1)

o $b_{yx} = -(\sigma_y / \sigma_x)$ (r = −1)

Did You Know?

“The product of the two regression coefficients ($b_{yx} \times b_{xy}$) is always equal to the square of the correlation coefficient (r^2). This provides a mathematical bridge between correlation and regression.”

“Activity: Identifying Key Predictors Using Multiple Regression”

Title: "Which Factors Drive Revenue? A Data-Driven Approach"

Instruction to the student:

A startup has collected data over 10 weeks, including the following for each week:

- Number of social media posts,

- Average daily website traffic,
- Weekly ad spend,
- Weekly revenue. You are required to:

1. Run a multiple linear regression to predict revenue using the three predictors.
2. Identify the regression coefficients for each variable and interpret their meanings.
3. Determine which variable has the most significant impact on revenue based on the magnitude of the coefficient.
4. Write a short explanation (100–150 words) advising the company where to focus its marketing efforts for maximum return.

9.5.4 Interpretation of Regression Coefficients

The slope (b) of a regression line equation has an operational...

• b_{yx} = 2.5 means:

"If we increase X by one when that is the only available change, Y increases by 2.5"

• b_{yx} = −0.75 means:

“Every time X increases by 1 unit, on average Y would decrease by 0.75 units”.

This is to help assess results, impact and shape decisions.

Real-life Example:

In marketing, if b_{yx} = 5.2 in the regression equation Sales = a + 5.2 × AdSpend, then for every ₹1,000 increase in

ad-spend, sales rise by ₹5,200 on average.

Summary Table: Regression Coefficient Essentials

Feature	**Value or Explanation**
Number per pair of variables	2 (Y on X and X on Y)
Relation to correlation coefficient	$r^2 = b_{yx} \times b_{xy}$
Sign	Same as that of **r**
Unit-dependence	Yes
Range	No fixed limits

Interpretation	Rate of change in dependent variable

9.6 Properties of Lines of Regression (Linear Regression)

Linear regression mean is fitting a line to the data points that more or less represents the relationship between independent and dependent variable. This is where we will discuss some properties of the regression line, the purpose of least squares and standard error, as well as its limitations.

9.6.1 Best Fit Line and Principle of Least Squares

Best Fit Line :

In linear regression, it finds the best fitting line which describes the relationship between the output and input by minimising (or optimising) the vertical distance between actual data points y and estimated values that lie on a line. These vertical wise distances are called residuals.

Principle of Least Squares:

The least squares approach finds the line that minimizes the sum of squared differences in these values, or residuals.

(i.e., the discrepancies between observed and predicted values of their dependent variable).

Objective: Minimize: $\Sigma(Y - \hat{Y})^2$ Where:

- Y = Actual value
- $\hat{Y}$ = Predicted value from regression equation

Equation form (for Y on X):

$\hat{Y} = a + bX$

- a = intercept
- b = slope (regression coefficient)

In this way, the total error is minimal, and on the whole, regression equation obtained are more practical to use for prediction.

9.6.2 Properties of Linear Regression Line

Passes through Mean Point:

The line of regression always passes through $(\bar{X}, \bar{Y})$, where X and Y are the means of the variables.

Two Separate Lines:

If the correlation is not a perfect one ($r = \pm 1$), the line of regression of Y on X is different from the line of regression of X on Y.

Minimizes Sum of Squares:

The least squares regression line minimizes $\Sigma(Y - \bar{Y})^2$ not $\Sigma(Y - \bar{Y})$, or some other measure of error.

Linear in Parameters:

Even after doing a change of variables the equation is linear in (a, b).

AWC Slope Corr: StDev Combination:

The slope of the regression line of Yon X is: 14.

$b = r \times (\sigma_y \div \sigma_x)$

Direction Given By the Sign of r:

If $r > 0$, the regression line in a positive slope; If $r < 0$, the regression has a negative slope.

9.6.3 Errors in Estimation and Standard Error of Regression

Errors in Estimation (Residuals):

There is no guarantee that each forecast of the regression line will be accurate. This difference is the residual:

Residual (e) = $Y - \hat{Y}$

Residuals are also key in examining how well the regression model used fits in with the data.

Standard Error of Estimate (S_e):

It's essentially a measure of the average size of residuals. A smaller standard error indicates more accurate predictions.

Formula:

$S_e = \sqrt{[\Sigma(Y - \hat{Y})^2 \div n]}$

Where:

- Y = Actual values

• $\hat{Y}$ = Predicted values from regression model

• n = Number of observations

Interpretation:

• A low S_e tells that the regression line is an excellent fit.

• A high S_e value indicates that the model may not be reliable for prediction.

9.6.4 Limitations of Linear Regression

Assumes Linearity:

The straight-line relationship is the model's hypothesis. This will produce misleading results if the real relationship is non-linear.

Sensitive to Outliers:

High outliers can skew the slope and intercept, which can affect accuracy.

Dependent on Assumptions:

The results hold only under assumptions such as normality, homo-scedasticity and independency are satisfied.

Does Not Prove Causation:

Causality should not be considered for a predictive model, no matter how good the correlation is.

Only One Dependent Variable:

Simple linear regression describes only one dependent variable. In practice various factors all have an effect on outcomes.

Limited in Multicollinearity Scenarios:

If you're working with multiple regression models (i.e., not simple linear) and your independent variables have a high correlation, then it can cause trouble for the model the regression coefficients.

Knowledge Check 1

Choose the correct option:

1. In simple linear regression, the dependent variable (Y) is predicted using:

A) Only one independent variable (X)

B) Multiple independent variables

C) The mean of the data

D) The correlation coefficient

2. What is the key difference between correlation and regression?

A) Correlation measures how one variable causes another

B) Regression predicts the value of one variable based on another

C) Regression only applies to nominal data

D) Correlation calculates the slope of a line

3. In the regression equation Y = a + bX, what does the coefficient b represent?

A) The predicted value of Y

3 B) The slope of the regression line (change in Y per unit change in X)

C) The value of X when Y = 0

D) The intercept of the regression line

4. The formula for calculating the regression coefficient of Y on X is:

A) $b_{yx} = r \times (\sigma_x \div \sigma_y)$

B) $b_{yx} = (\Sigma X \times \Sigma Y) \div (\Sigma X^2 \times \Sigma Y^2)$

C) $b_{yx} = \Sigma(Y - \hat{Y}) \div \Sigma X$

D) $b_{yx} = (\sigma_x \div \sigma_y) \times r$

5. What is the interpretation of a regression coefficient of 0.5 for X in a model predicting Y?

2 A) For every 1 unit increase in Y, X increases by 0.5 units

B) For every 1 unit increase in X, Y increases by 0.5 units

C) Y has no effect on X

D) The regression model is not valid

9.7 Summary

- The discussion in this chapter focused on the basics and purpose of regression analysis, which is a main statistical method used to describe as well as model relationships between variables.

- We started off with the definition and significance of regression, as well as it's comparison to correlation.
- We had studied the type of regressions — simple, multiple, linear and non-linear forms and when each one is suitable?
- We discovered how to investigate regression both graphically and algebraically in the scatter diagram method and method of least squares.
- You should have like then begun discussing the lines of regression- Y on X and X on Y respectively, and how these intersect in relation to how they interact with the mean point ($\bar{X}$, $\bar{Y}$).
- We figured how to use regression coefficients and interpreted them, examining their relation with the correlation coefficient.
- At last, we discussed the characteristics of the regression lines covering many areas such as its assumptions, how to estimate error and a few drawbacks of linear regression models.

❖ And as much as regression is vital to business, economics and data science for prediction, trend explanations and historical data-driven decision-making…

9.8 Key Terms

1. regression - in statistics, the process of dealing with interrelationships among causes and effects without assuming causative processes.
2. Univariate Regression – Short for regression with one independent and one dependent variable
3. Multiple Regression - A regression with two or more predictors
4. Linear Regression- A regression wherein the relationship is modeled by a straight line.
5. Non-linear Regression — regression that is curved, or non-linear
6. Least Squares Method - Procedure that reduces the sum of squared deviates
7. Regression Coefficient - A measure of the velocity of change between Y and X.
8. Residual - The value by which a predicted value differs from the actual value
9. SEE = Standard Error of Estimate, with which predictions from the regression model are correct.
10. Best Fit Line– The line that minimizes the error of prediction.

9.9 Descriptive Questions

1. Discuss what is meant by regression and why it is important for business forecasting.
2. Differentiate between Simple and Multiple Regression with examples.
3. Explain the difference between linear and non-linear regression. Give one use case for each.
4. Explain The method of least squares. How is the regression line obtained using it?

5. Explain the regression lines of Y on X and X on Y. Under what circumstances do they coincide?
6. What is difference between correlation coefficient and regression coefficients?
7. Explain any three characteristics the linear regression line.
8. What does the standard error of estimate mean in regression analysis?
9. Describe any four limitations of linear regression, when applied in a practical setting.
10. What are the assumptions that need to be satisfied for linear regression to be valid?

9.10 References

1. Gupta, S. C. (2014). Fundamentals of Statistics. Himalaya Publishing House.
2. Sharma, J. K. (2018). Business Statistics. Vikas Publishing House.
3. Levin, R. I., & Rubin, D. S. (2017). Statistics for Management. Pearson Education.
4. Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2016). Statistics for Business and Economics. Cengage Learning.
5. UGC e-Pathshala – Statistical Methods Modules
6. National Statistical Office (NSO) Reports, Govt. of India

Answers to Knowledge Check

Knowledge Check 1

1. A) Only one independent variable (X)
2. B) Regression predicts the value of one variable based on another
3. B) The slope of the regression line (change in Y per unit change in X)
4. A) $b_{yx} = r \times (\sigma_x \div \sigma_y)$
5. B) For every 1 unit increase in X, Y increases by 0.5 units

9.11 Case Study

“Forecasting Success” or “Success Never Comes So Easy”?

Introduction

In the ever changing world of retail, data driven decision making is essential to remain competitive. Companies frequently use historical information to forecast future performance and adjust their operations. One of the techniques being regression analysis, which is a statistical methodology applied to modelling and understanding the relationship between dependent and independent variables.

This case study details a mid-sized retail company called TrendMart has used regression analysis to forecast monthly sales with advertising spend, foot traffic and online leads. The company also achieved marketing strategy optimization, resource allocation improvement and profit forecasting degree by a systematic method of past data analysis.

Background

TrendMart is based out of several urban markets and also, online. Over the past 12 months, management began to observe a discrepancy between the company's advertising investment and results produced in actual sales. While some campaigns resulted in a lift to foot traffic or social media engagement, the impact on monthly revenue was uneven.

In attempt to solve the above, the analytics team suggested using multiple linear regression to capture the relationship between monthly sales (Y) and the following independent variables:

• Ad Spend (X_1) in ₹,

• Foot Traffic (X_2) as mean daily number of store visitors,

• Social Media Engagement (X_3) measured by the number of likes, shares and comments.

Regression analysis was performed and the following mathematical model based on the 12 months of historical data obtained from teams was established:

Sales = 2.5 + 3.1(Ad Spend) + 2.8(Foot Traffic) +0.6(Engagement), proportionate to last year's ad spend (theglobeandmail.com).

The model R^2 was 0.89; therefore, it can be said that 89% of the variance in sales could be explained by three predictors.

Problem Statement 1: Breakdown in Ad Spend and Sales Calculus.

Ads were placed, and the spending was steady; yet the management discovered that not all ads eventually led to an increase in sales.

Solution:

The regression model indicated that traffic also had a slight greater importance than ad spend in moving the sales needle. So instead of blindly growing ad budgets, TrendMart reallocated the spend towards local targeting and in-store promotions that were more closely linked to store visits - and sales.

Problem 2: How to engage online but underestimate its real effects?

Prior to now, marketing departments discounted social media strength as a viable revenue driver.

Solution:

Results from the regression analysis indicated that despite being the weakest predictor, online engagement remained a significant one. With this in mind, the digital team commenced schedule testing.

promotions based on social peaks leading to increased campaign matching and higher conversion.

Problem 3: Predicting Next Quarter's Sales from Few Data Inputs

TrendMart had to make a prediction about the sales of the next quarter, but did not have half of its transactions for that very same month.

Solution:

The team dropped the values that they had for Ad Spend and Foot Traffic into the regression model, and filled in an estimated value for Engagement based on historical performance. This allowed for an initial forecast that could be updated when the numbers were actually in.

MCQ (Knowledge Check)

In the regression model Sales = 2.5 + $3.1X_1$ + $2.8X_2$ +.$6X_3$, what is 3.1?

A) Intercept

B) Error term

C) Regression coefficient of Ad Spend

D) Total variance

Solution: C) Regression coefficient of Ad Spend

What is the meaning of 0.89 R^2 in a regression model?

A) The model has low accuracy

B) The variables are not related

C) 89% of variance in sales can be accounted for by predictors

D) The prediction error is 89%

Answer: C) 89% of the variation in sales can be accounted for by predictors.

What does it mean if the coefficient associated to Social Media Engagement is small and possitive?

A) It should be removed

B) It has no influence

C) It has a positive impact, though less than the other factors.

D) It is negatively correlated

Correct Answer: C) It makes a favorable contribution but not as much as the other factors

Conclusion

It is a classic example which underlines the usefulness of regression in our business environment. By effectively measuring the effect of various factors on sales, TrendMart was able to make the transition from making decisions based on gut feelings to forecast driven by data. The end result was increased campaign effectiveness, smarter money allocation and more predictable revenue.

Regression empowered the company to:

- Discover the most significant sales predictors,
- Forecast future revenue more accurately,
- Refine marketing tactics using data-driven intelligence.