📘 BUSINESS STATISTICS – SUMMARY DOCUMENT (WEEK 7)

Week 7-Course material-1

7.1 Introduction

This section introduces **hypothesis testing for proportions**.
Earlier you learned to test differences between means; now the focus is on **testing differences in proportions** when data is **categorical** (e.g., yes/no, success/failure).

Key Ideas

- Proportion tests evaluate whether **sample proportion** differs from a **population proportion** or **another sample proportion**.
- Types:
    - **One-sample test of proportions**
    - **Two independent sample test of proportions**

Both work similarly to mean tests but apply to **binary/categorical data**.

7.1.1 Key Assumptions for Proportion Tests

To ensure results are valid:

1. **Random Sampling**
2. **Independence** of observations
3. **Binary Outcomes** (success/failure)
4. **Sufficient Sample Size**
5. **Normal Approximation** (sample proportion approx. normal)

7.2 Parametric Test for Single Proportion

Used to test whether a **sample proportion (P)** differs from a **known or historical proportion (PH0)**.

Steps:

a. Formulate hypotheses
b. Extract data
c. Compute **Standard Error**

$$\sigma_p = \sqrt{\frac{P_{H0}Q_{H0}}{n}}$$

d. Compute **Z statistic**

$$Z = \frac{P - P_{H0}}{\sigma_p}$$

e. Compare with **p-value** or **critical Z**

Examples & Problems

Example 1 (p.1)

Retail company checking if new policy changed satisfaction levels (from 65% to sample 70%).

Example 2 (p.3)

Comparing 39% vs 41% historical loan proportion.

Problem 1 – Loan Proportion Test (p.3)

Conclusion:
Fail to reject H_0 → No significant change in proportion of loans to women.

Problem 2 – Customer Satisfaction Test (p.3-4)

Conclusion:
Fail to reject H_0 → Rick's claim that satisfaction remains ~86% is valid.

7.3 Test for Differences Between Two Sample Proportions

Used to compare whether **two populations differ** on a categorical attribute.

Steps:

1. Formulate hypotheses
2. Extract binary data
3. Calculate pooled **Standard Error**
4. Compute **Z-statistic**

5. Determine significance using p-value

Assumptions

- Binary outcome
- Independent samples
- True underlying proportion exists

Example (p.4)

Comparing customer satisfaction rates between Store A and Store B.

Problem 3 – Pollution Control System Effectiveness (p.4–5)

Two systems, comparing 68% vs 76%.

Conclusion:
Fail to reject H_0 → Difference not significant → **Cheaper system installed**

Problem 4 – Textbook Purchase Behavior (p.5)

Freshmen (46%) vs. Sophomores (40%).

Conclusion:
Fail to reject H_0 → No significant difference.

7.4 Non-Parametric Tests – Introduction

Non-parametric tests are used when **parametric assumptions (normality, numeric scale)** are not met.

Parametric vs Non-Parametric (p.6–7)

A comparison table is included on **Page 7** showing alternatives such as:

Parametric	Non-parametric
t-test	Mann–Whitney U, Wilcoxon
ANOVA	Kruskal–Wallis

Key Features

- Do **not require normality**

- Suitable for:
    - Ordinal data
    - Categorical data
    - Skewed distributions
- Less sensitive to outliers

Examples

- Mann–Whitney for comparing satisfaction ratings
- Wilcoxon for before–after performance improvement

7.4.1 Chi-square – One Way Classification

Used to analyze categorical data when only **counting** is possible.

Key Points (p.7–8)

- Tests whether observed categorical data fits expected distribution.
- Useful for gender counts, voter qualification, etc.

How it works

Compares **Observed (O)** vs. **Expected (E)** frequencies:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Degrees of freedom:

$$df = k - 1$$

Problem 5 – Success Definition Among Entrepreneurs (p.9–10)

Table of O, E, $(O-E)^2/E$ retained.

Conclusion:
$\chi^2 = 198.48$, df=3 → Reject H_0
→ Distribution **not the same** as for women.

Problem 6 – Age Distribution of Music Shoppers (p.10–11)

Conclusion:
Accept H_0 → Local distribution **matches** national survey.

7.4.2 Chi-square – Test of Association (Two-way Tables)

Used to determine if **two categorical variables** are dependent.

Key Aspects (p.11–12)

- Requires contingency table
- Tests whether variables are independent

$$df = (r-1)(c-1)$$

Example: Colour Preference & Gender (p.11–12)

Table provided for Green/Red/Blue vs Male/Female.

Problem 7 – Newspaper Readership & Education Level (p.12–13)

Large contingency table included.

Conclusion:
$\chi^2 = 32.855$ > critical value → Reject H_0
→ Readership frequency **depends on education level**.

Problem 8 – Grades vs Music Listening Hours (p.14–15)

Data for 4 categories of hours vs 5 grade categories.

Conclusion:
$\chi^2 = 63.829$, $p<0.001$ → Reject H_0
→ Grades **depend on** hours spent listening to music.

📘 WEEK 8 – SUMMARY DOCUMENT (NON-PARAMETRIC TESTS)

(Based on W8-summary-1.pdf)

W8-summary-1

---

1. Introduction to Non-Parametric Tests (Pages 2–4)

What Are Non-Parametric Tests?

Statistical tests that **do not require assumptions about population distribution** (e.g., normality). Used when data are:

- Not normally distributed
- Ordinal (ranked)
- Categorical
- Contain outliers
- Small sample size

Why Use Them?

1. **No distribution assumptions**
2. **Robust against outliers**
3. **Work with ordinal/ rank data**
4. **Useful for small samples**
5. **Flexible for skewed / non-metric data**

Common Non-Parametric Tests

Purpose	Test
Compare 2 independent groups	Mann-Whitney U / Wilcoxon Rank-Sum
Compare 2 related groups	Wilcoxon Signed-Rank
Compare 3+ independent groups	Kruskal-Wallis H
Categorical associations	Chi-Square

A visual chart of these test types appears on **Page 3** of the file.

W8-summary-1

---

2. Session 8.1 – Mann-Whitney U Test (Pages 5–8)

Purpose

Non-parametric alternative to **independent samples t-test**.
Used when:

- Data are **ordinal** or **non-normal**

- Comparing **two independent groups**

Key Concepts

- Ranks are assigned to combined data
- Tests whether groups have the **same distribution**

Steps & Formulas

1. **Hypotheses**
    - H_0: No difference between groups
    - H_1: Groups differ
2. **Compute U statistic**

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

3. **Mean of U**

$$\mu_U = \frac{n_1 n_2}{2}$$

4. **Standard deviation**

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

5. **Z-value**

$$Z = \frac{U - \mu_U}{\sigma_U}$$

Examples

- Customer satisfaction across two store layouts
- Employee performance across two departments
- Business applications include satisfaction scores, productivity, campaign performance, etc.

Global & Indian Corporate Use Cases

- Amazon, Google, Toyota, Boeing (Global)
- Infosys, HUL, ITC, TCS, Maruti Suzuki (India)

3. Session 8.2 – Kruskal-Wallis H Test (Pages 8–13)

Purpose

Non-parametric alternative to **One-Way ANOVA**.
Used when:

- Comparing **three or more independent groups**
- Data are ordinal or non-normally distributed

Key Concepts

- All observations are **ranked together**
- Compares **median differences** across groups

H Statistic Formula

$$K = \frac{12}{n(n+1)} \sum \frac{R_j^2}{n_j} - 3(n+1)$$

Where:

- R_j= sum of ranks of group j
- n_j= group size
- n= total observations

Degrees of Freedom

$$df = k - 1$$

Decision Rule

- Compare **K** to chi-square table critical value.

Examples

- Customer satisfaction across 3 branches
- Training program effectiveness across 3 programs
- Quality comparison across manufacturing plants

Global & Indian Corporate Use Cases

- Nestlé, GE, IBM, HSBC (Global)
- Wipro, M&M, HAL, BHEL (India)

---

4. Session 8.3 – Wilcoxon Signed-Rank Test (Pages 15–17)

Purpose

Used for **paired or repeated measurements**, non-parametric alternative to **paired t-test**.

Key Concepts

- Works on **differences** between paired observations
- Ranks the absolute differences
- Signs (+/–) retained
- Test statistic = smaller of **T+** or **T–**

Formulas (Large Sample Approximation)

$$\mu_T = \frac{n(n+1)}{4}$$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

$$Z = \frac{T - \mu_T}{\sigma_T}$$

Examples

- Checkout productivity using two billing systems
- Attitude survey comparing 1990 vs. 2011 views

---

5. Session 8.4 – Friedman's Two-Way ANOVA (Pages 17–18)

Purpose

Used for **three or more related groups**, non-parametric alternative to **Two-Way ANOVA (Repeated Measures)**.

Key Concepts

- Data arranged into **blocks (rows)** and **treatments (columns)**
- Ranks assigned **within each block**

Test Statistic (Friedman χ^2r)

$$\chi_r^2 = \frac{12}{nk(k+1)} \sum R_j^2 - 3n(k+1)$$

Where:

- n = number of blocks
- k = number of treatments
- R_j = sum of ranks in treatment j

Examples

- Preference for multiple calculator brands
- Ranking cities for employee relocation

---

6. Session 8.5 – Summary of Hypothesis Testing (Pages 18–19)

Measurement Scale → Appropriate Tests

Categorical Data

- One sample: Chi-Square Goodness of Fit
- Two samples: McNemar, Fisher's Exact
- K samples: Cochran's Q, Chi-Square for independence

Ordinal Data

- One sample: Kolmogorov-Smirnov
- Two sample: Sign test, Wilcoxon matched pairs
- K sample: Median test, Friedman test, Kruskal-Wallis

Numeric Data

- One sample: Z test / t-test
- Two sample: Independent t-test / Paired t-test
- K sample: One-way ANOVA / Repeated Measures ANOVA

This section summarises how to choose the correct hypothesis test based on **sample structure** and **data type**.

---

7. Conclusion & References (Pages 19–20)

Primary Book Reference

- *Statistics for Management* – Levin, Rubin, Siddiqui, Rastogi

Online Reference

- Article comparing parametric vs non-parametric tests

Case Study References

A list of 27 academic and corporate case studies demonstrating real-world usage of non-parametric tests (e.g., Amazon, IBM, Infosys, Nestlé, TCS, GE, HAL, BHEL, etc.)

📘 WEEK 9 – SUMMARY DOCUMENT (PREDICTIVE ANALYTICS)

(Based on W9-Summary-1.pdf)

W9-Summary-1

---

1. Introduction to Predictive Analysis (Pages 1–6)

Definition

Predictive analysis uses **historical data**, **statistical algorithms**, and **machine learning** to **forecast future outcomes**, behaviours, and trends.

Purpose & Importance

Predictive analytics helps organizations:

- Improve decision-making
- Anticipate risks and opportunities
- Optimize operations and supply chains
- Personalize marketing
- Improve customer satisfaction
- Reduce costs
- Enhance strategic planning

Major Business Applications

- **Retail:** Demand forecasting, inventory optimization
- **Finance:** Credit scoring, fraud detection
- **Marketing:** Customer segmentation, targeted campaigns
- **Supply Chain:** Predicting raw material needs, logistics optimization
- **Human Resources:** Predicting turnover, performance trends
- **Insurance:** Claim likelihood prediction

Key Components of Predictive Analysis

1. **Data Collection**
2. **Data Cleaning**
3. **Model Building**
4. **Validation**
5. **Prediction**
6. **Decision Making**

Example: Retail Sales Forecasting

A retail chain uses:

- Time series models (ARIMA)
- Regression analysis
- Machine learning models
  → to forecast demand, plan promotions, and optimize inventory.

Global Examples

- **Amazon:** Product recommendations
- **Netflix:** Content recommendations
- **UPS:** Route optimization
- **Coca-Cola:** Customer engagement analytics
- **Walmart:** Predictive inventory systems

Indian Examples

- **Flipkart:** Product personalization
- **ICICI Bank:** Fraud analytics
- **Ola:** Demand forecasting
- **TCS:** Workforce planning
- **Mahindra:** Predictive maintenance

---

2. Session 9.1 – Correlation & Regression (Pages 15–16)

Correlation

- Measures **strength & direction** of relationship between variables

- Does *not* imply causation

Regression

- Predicts one variable (dependent) using another(s) (independent)
- Determines magnitude & nature of relationship

Predictive Analytics Framework

- Uses regression, ML algorithms, and historical data
- Helps forecast trends such as:
    - Sales
    - Customer behaviour
    - Market movements

---

3. Session 9.2 – Karl Pearson's Correlation (Pages 16–17)

Definition

Pearson's r measures the **linear relationship** between two continuous variables.
Value range: **-1 to +1**

Formula

$$r = \frac{\sum X_i Y_i - n\bar{x}\bar{y}}{\sqrt{\sum X_i^2 - n\bar{x}^2}\sqrt{\sum Y_i^2 - n\bar{y}^2}}$$

Examples

1. **Banking Example**
   Correlation between number of bankers and waiting time:

$$r = -0.813$$

→ Strong negative correlation (more bankers → less waiting time)

2. **Zippy Cola Advertising**

$$r = 0.7867$$

→ Strong positive relationship between ads seen & cans purchased

Characteristics

- Symmetric
- Unit-free
- Sensitive to outliers

Business Applications

- Market research
- Financial analysis
- CRM analytics
- Supply chain optimization
- HR analytics
- Strategic planning

---

4. Session 9.3 – Spearman's Rank Correlation (Pages 17–18)

Definition

A **non-parametric**, rank-based measure of monotonic relationship.
Used when:

- Data are ordinal
- Not normally distributed
- Relationship is monotonic but not linear

Formula

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Example Calculation

A real estate study shows:

$$\rho = 0.49$$

→ Moderate positive correlation between home price rank & days to sell

Applications

- Education (student ranking)
- Finance (portfolio risk ranking)

- Healthcare (symptoms vs treatment outcomes)
- Market research
- HR performance ranking
- Supply chain ranking

---

5. Session 9.4 – Simple Regression 1 (Pages 18–19)

Definition

Simple regression predicts **one dependent variable (Y)** using **one independent variable (X)**.

Assumptions

1. Numeric variables
2. Linearity
3. Normal distribution
4. Independence
5. Homoscedasticity

Equations

Slope (b):

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

Intercept (a):

$$a = \bar{y} - b\bar{x}$$

Regression line:

$$Y = a + bX$$

Interpretations

- **r^2 (Coefficient of Determination):** % variance explained
- **β (Standardized Coefficient):** strength of effect
- **B (Unstandardised Coefficient):** actual change in Y per unit X

Examples

- Advertising → Sales
- Square footage → House prices

---

6. Session 9.5 – Simple Regression 2 (Pages 19–20)

Real-World Examples

1. **Appliance Sales vs Housing Starts**
   Regression model used to forecast appliance sales.
2. **Supervisor Interruptions vs Worker Hostility**
   Predict hostility score based on number of interruptions.

Key Steps

- Fit regression line
- Predict future scores
- Interpret coefficients

---

7. Conclusion & References (Pages 20–21)

Summary

Predictive analytics:

- Forecasts future trends
- Identifies risks & opportunities
- Enhances business decision-making

Future Trends

- Increased use of **AI & Machine Learning**
- Improved predictive accuracy
- Automated decision systems

References

Includes reports from:

- Deloitte
- Harvard Business Review
- Forbes

- Amazon, Netflix, UPS technology blogs
- Academic journals (JSTOR, Springer, Wiley, etc.)

📘 WEEK 10 – SUMMARY DOCUMENT (MULTIPLE REGRESSION & p-VALUE APPROACH)

Based on Week 10 – Course material_DA-1.pdf

Week 10-Course material_DA-1

10.1 Introduction (Page 1)

Predictive Analytics Overview

Predictive analytics uses:

- **Historical data**
- **Statistical techniques**
- **Machine learning models**

To forecast:

- Customer behaviour
- Risks
- Sales trends
- Operational outcomes

Helps organizations in **marketing, HR, finance, operations** to make **data-driven decisions**.

Simple Regression Refresher

Models the relationship between **one independent variable (X)** and **one dependent variable (Y)**:

$$Y = a + bX$$

Example: Predicting sales from advertising expenditure.

Multiple Regression Introduction

Expands simple regression to include **multiple predictors**:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3$$

Provides a more comprehensive view of relationships between variables.

10.2 Multiple Regression & Assumptions (Pages 1–2)

Definition

Multiple regression involves:

- **One dependent variable (Y)**
- **Two or more independent variables (X_1, X_2, ...)**

Assumptions

1. **One dependent + 2 or more independent variables**
2. **Numeric data**
3. **Normal distribution of data**
4. **No multicollinearity** among predictors
    - Multicollinearity distorts β-coefficients
    - Must be checked and resolved

10.3 Key Aspects of Multiple Regression (Page 2)

Components

- **Y** = Dependent variable
- **X_1, X_2, ... X_n** = Predictors
- **Regression coefficients (β_0, β_1, ...)** indicate how much Y changes per unit change in each X.
- **β_0 (Intercept)** is the value of Y when all Xs = 0.
- Error term (**ϵ**) accounts for unexplained variance.

Model Equation

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n + \epsilon$$

10.4 Applications of Multiple Regression (Page 3)

Business Use Cases

1. **Sales Forecasting**
   Predict sales using advertising, pricing, location.

2. **Employee Performance Evaluation**
   Predict performance using experience, education, training hours.
3. **Customer Satisfaction Analysis**
   Factors like service speed, cleanliness, staff friendliness.
4. **Market Research**
   Understand impact of various marketing channels.
5. **Real Estate Pricing**
   Predict house price using square footage, bedrooms, location, age.

10.5 Steps to Develop a Multiple Regression Equation (Pages 3–4)

To find the regression equation:

$$Y = a + b_1X_1 + b_2X_2 + \epsilon$$

We solve the system:

1. $$\sum Y = na + b_1\sum X_1 + b_2\sum X_2$$
2. $$\sum X_1Y = a\sum X_1 + b_1\sum X_1^2 + b_2\sum X_1X_2$$
3. $$\sum X_2Y = a\sum X_2 + b_1\sum X_1X_2 + b_2\sum X_2^2$$

These three equations are solved to get **a, b_1, b_2**.

📌 **PROBLEMS & SOLUTIONS**

Problem 1 (Pages 4–7)

Objective: Predict apartment rent using:

- X_1 = Number of rooms
- X_2 = Distance from downtown

Regression Equation Derived

$$\text{Rent} = 96.4375 + 136.4845X_1 - 2.4X_2$$

Prediction

For a **2-bedroom apartment, 2 miles away**:

$$Y = 96.4375 + 136.4845(2) - 2.4(2) = \$364.61$$

Problem 2 (Pages 7–10)

Predict **car price (in thousands)** using:

- X_1 = Year
- X_2 = Miles driven

Regression Equation Derived

$$\hat{Y} = -4243.1682 + 2.1315X_1 + 0.2135X_2$$

Prediction

1991 car with 40,000 miles:

$$\hat{Y} = 9.188 \text{ (thousand dollars)}$$

Problem 3 (Pages 10–13)

Federal Reserve predicting **% change in GNP** using:

- X_1 = Federal deficit (in billions)
- X_2 = Mean Dow Jones Index

Regression Equation Derived

$$\hat{Y} = -3.6963 - 0.0136X_1 + 0.0076X_2$$

Prediction

If deficit = 120B and Dow = 1000:

$$\hat{Y} = -3.6963 - 0.0136(120) + 0.0076(1000)$$

10.6 Data Analytics Using the 'p-Value' Approach (Pages 13–16)

Two major approaches to hypothesis testing:

10.6.1 Critical Value Approach (Page 13)

Steps:

1. State H_0 & H_1
2. Choose α
3. Find **critical value**
4. Compute **test statistic**
5. Compare → **Reject or fail to reject H_0**

10.6.2 p-Value Approach (Pages 13–14)

Steps:

1. State hypotheses
2. Choose α
3. Compute test statistic
4. Calculate **p-value**
5. If **$p \leq \alpha$ → Reject H_0**

Difference:

- Critical-value method compares against threshold
- p-value method measures *probability of observing the result by chance*

Both lead to the same decision.

10.6.3 Decision Making (Page 14)

- Small p-value → Strong evidence against H_0
- Large p-value → Insufficient evidence to reject H_0

10.6.4 Application to Predictive Methods (Page 14)

p-values apply to:

- Regression (β significance)
- Correlation (significance of r)

Helps validate predictive models statistically.

10.6.5 Using p-Value in Excel (Page 15)

Steps:

1. Enter data
2. State hypotheses
3. Use functions such as:
    - **T.TEST()**
    - **Z.TEST()**
4. Excel outputs:
    - Test statistic
    - p-value

10.6.6 Advantages of p-Value Approach in Excel (Page 15–16)

- Easy to use
- No programming needed
- Fast calculations
- Widely available
- Supports multiple tests
- Great for visualization and reporting
- Integrates with add-ins (e.g., Analysis ToolPak)

📘 **WEEK 11 – SUMMARY DOCUMENT (FACTOR ANALYSIS & MODEL FITNESS)**

Based on Week 11 – Course material_DA-1.pdf

Week 11- Course material_DA-1

11.0 Overview

In data analysis, beyond describing and predicting, analysts also **classify** or **segment** data into groups that share internal homogeneity.
Why this matters (Page 1):

- Customer behaviours differ across regions/countries
- A single strategy doesn't fit all
- Segmenting helps build targeted strategies aligned to demographic, behavioural, and cultural differences

---

11.1 Classification & Segmentation Methods

11.1.1 Classification Methods (Page 1–2)

Classification assigns data points to **predefined labels** (known categories).
Used for:

- Market segmentation (demographics, behaviours)
- Credit scoring (finance)
- Disease classification (healthcare)

It requires a **labelled dataset** and predicts which class a new observation belongs to.

11.1.2 Segmentation Methods (Page 1–2)

Segmentation **discovers natural groupings** in data without predefined labels.
Used for:

- Customer targeting
- Market analysis
- Pattern recognition
- Personalized engagement

Segmentation is *exploratory*, not predictive.

11.1.3 What Are Predefined Labels? (Page 2)

These are categories that exist **before** analysis—for example:

- Spam / Not Spam

- Good customer / Bad customer

The model is trained to classify new observations into these categories.

11.1.4 Predictive vs Segmentation Models (Page 2–3)

Predictive Models

- Have a **dependent** & **independent** variable
- Example: Multiple regression
- Goal → Forecast Y from X_1, X_2, ...

Segmentation Models

- No dependent/independent relationship
- Examine interrelationships among variables
- Used when variables are many and may be correlated
- Aim → Identify latent structures or groups

When variables increase, multicollinearity and overlapping influences appear, requiring techniques like **factor analysis** and **cluster analysis**.

11.1.5 Introduction to Factor & Cluster Analysis (Page 3)

Factor Analysis

- Groups **variables** based on correlation
- Extracts underlying **latent factors**
- Also called **R-analysis**

Cluster Analysis

- Groups **individuals/objects** based on similarity
- Items in same cluster are similar; different clusters are dissimilar
- Also called **Q-analysis**, **numerical taxonomy**, **typology construction**

11.2 Factor Analysis (Pages 3–5)

Factor analysis aims to uncover underlying structures in data by grouping correlated variables.

Objectives of Factor Analysis (Page 3–4)

1. **Data Reduction** – reduce many variables into fewer factors
2. **Structure Identification** – discover underlying patterns
3. **Data Transformation** – convert correlated variables into uncorrelated factors
4. **Create factors for further analysis** (e.g., regression)

Example: Student Personality Assessment (Page 4–5)

Variables include subject marks + height + weight.
Outcome:

- Academic subjects → "Intelligence"
- Height & Weight → "Physical Fitness"

Example: Restaurant Food Quality (Page 5–6)

Variables such as waiting time, cleanliness, taste, temperature, freshness grouped into:

- Service Quality
- Food Quality

This demonstrates **latent variables** inferred from observed ones.

11.3 How Factor Analysis Works (Pages 6–7)

Latent Variables

- Unobserved constructs inferred from observed variables
- Example: "Food Quality" inferred from taste, temperature, freshness

Observed/Manifest Variables

- Directly measured characteristics

Dimensionality Reduction

- Groups highly correlated variables
- Produces fewer, interpretable latent factors

Naming Factors

- Based on interpreting variables with highest loadings
- Requires domain knowledge + intuitive understanding

Problem 1 – Toothpaste Benefits Study (Pages 7–9)

A study with 30 respondents rating six toothpaste attributes (e.g., cavity prevention, shiny teeth, fresh breath).

Goal → Identify underlying benefit factors such as:

- **Medicinal benefits** (cavity protection, gum strength, decay prevention)
- **Aesthetic benefits** (shiny teeth, attractive teeth, fresh breath)

Factor analysis can reveal these latent benefit groups.

11.2.2 Model Fitness – Before Running Factor Analysis (Pages 9–12)

To check if data is suitable for factor analysis, review **three model fit indicators**:

1. Correlation Matrix (Page 9–11)

- Check if correlations ≥ **0.30**
- High correlations indicate reduction potential
- Low correlations → factor analysis inappropriate

Example: Cavity prevention correlates strongly (0.873) with gum strengthening → likely to form a factor.

---

2. Kaiser-Meyer-Olkin (KMO) Measure (Page 11–12)

- Indicates sampling adequacy
- Values:
    - **≥ 0.80** → Excellent
    - **≥ 0.70** → Good
    - **≥ 0.60** → Fair
    - **≥ 0.50** → Poor
    - **< 0.50** → Unacceptable

In the example: **KMO = 0.660 (Fair but acceptable)**

KMO improves with:

- Larger samples
- Higher correlations
- Relevant variables

- Fewer factors

---

3. Bartlett's Test of Sphericity (Page 12–14)

- Tests whether correlation matrix differs from identity matrix
- If **significant ($p < 0.05$)** → suitable for factor analysis
- If not significant → variables not correlated enough

Model Fitness Summary (Page 14)

Factor analysis is appropriate if **any one** of the following is satisfied:

1. Correlation coefficients > 0.30
2. KMO > 0.50
3. Significant χ^2 in Bartlett's Test

11.4 Communality (Pages 14–15)

Communality = amount of variance in a variable explained by the factors.

Key Points

- Range: **0 → 1**
- High communality → variable well explained
- Low communality → variable not represented well
- Helps evaluate quality of factor solution

11.5 Eigenvalues (Pages 15–16)

Eigenvalue = amount of variance explained by each factor.

Interpretation

- High eigenvalue → important factor
- Low eigenvalue → weak factor
- **Kaiser Criterion:** retain factors with eigenvalue > **1**
- Scree plot helps identify "elbow" (number of meaningful factors)

Example:
If Factor 1 eigenvalue = 2.5, Factor 2 = 1.5 → Together explain 80% variance.

11.6 Rotation Methods (Pages 16–17)

Rotation improves interpretability without changing total variance.

Orthogonal Rotation (Factors uncorrelated)

- **Varimax** → most common
- **Quartimax**
- **Equamax**

Oblique Rotation (Factors allowed to correlate)

- **Direct Oblimin**
- **Promax**

Why Rotate?

- Simplifies structure
- Makes interpretation easier
- Groups variables clearly under each factor

11.7 Factor Extraction Methods (Pages 17–18)

Common methods include:

1. PCA – Principal Component Analysis

- Best for data reduction
- Uses total variance

2. PAF – Principal Axis Factoring

- Focus on **shared variance**
- Good for uncovering latent constructs

3. ML – Maximum Likelihood

- Allows hypothesis testing
- Requires multivariate normality

4. ULS – Unweighted Least Squares

5. GLS – Generalized Least Squares

6. Alpha Factoring (maximizes reliability)

11.8 Factor Loadings & Interpretation (Pages 18–19)

Factor Loading Thresholds

- **≥ 0.30** → minimal significance
- **≥ 0.40** → more meaningful
- **≥ 0.50** → practically significant
- **≥ 0.70** → strong loading

Interpretation Steps

1. For each variable, find highest loading
2. Underline/mark it
3. Use contributing variables to name the factor
4. Keep labels simple, meaningful
5. Consider direction (+/–)

High loadings = Variables important to the factor

📘 WEEK 12 – SUMMARY DOCUMENT (CLUSTER ANALYSIS & SEGMENTATION)

Based on Week 12_Course Material_DA-1.pdf

Week 12_Course Materrial_DA-1

12.1 Introduction to Cluster Analysis (Pages 1–2)

Cluster analysis and segmentation are **data mining techniques** used to uncover hidden patterns and group data into meaningful clusters.

Segmentation

- Divides a heterogeneous population into **homogeneous groups**.
- Common segmentation bases:
    - **Demographics** – age, gender, income

- **Geographic** – country, region, city
- **Behavioral** – usage, purchase patterns, loyalty
- **Psychographic** – lifestyle, values, personality

Used heavily in **marketing** for targeted communication.

Cluster Analysis

- An **unsupervised learning** method
- Automatically groups data points based on similarity
- No predefined labels required
- Common algorithms: **K-means**, **Hierarchical clustering**, **DBSCAN**

Combined Use

A company may:

1. Segment customers by demographics
2. Then apply clustering within each segment to find micro-groups

This improves personalization (e.g., targeted offers, product recommendations).

12.2 Cluster Analysis – Concept (Pages 2–3)

Cluster analysis groups objects so that:

- **Within a cluster** → objects are *similar*
- **Across clusters** → objects are *distinct*

Use cases (Page 3)

1. Customer segmentation
2. Product development insights
3. Churn prediction
4. Supply chain optimization
5. Credit risk assessment
6. Social media audience grouping

Cluster analysis helps businesses:

- Identify behavioural patterns
- Tailor strategies

- Improve customer satisfaction and revenue

12.2.1 Types of Cluster Analysis (Pages 2–3)

1. K-Means Clustering

- Most widely used
- User specifies K clusters
- Minimizes within-cluster variation

2. Hierarchical Clustering

- Builds a tree (dendrogram)
- Does not need predefined K
- Merges or splits clusters iteratively

3. Centroid-Based Clustering

- Uses a centroid representing each cluster
- Assigns points to nearest centroid

4. Distribution-Based Clustering

- Assumes data comes from probabilistic distributions

5. Spectral Clustering

- Uses similarity matrix + dimensionality reduction

12.2.2 Assumptions in Cluster Analysis (Pages 3–4)

1. Data Type

- Numerical → Euclidean, Manhattan distance
- Categorical → Jaccard, Hamming distance

2. Homogeneity

Clusters must be internally similar and externally different.

3. Independence

Observations must be independent.

4. Scale of Measurement

- Variables must be standardized if scales differ

- Otherwise larger-scaled variables dominate clustering

5. Sample Size

- Larger datasets → more stable clusters

6. Outlier Sensitivity

Outliers distort cluster boundaries.

7. Normality (optional)

Models like Gaussian Mixture assume normal distribution.

12.2.3 Distance Measures Used in Clustering (Pages 4–5)

1. Euclidean Distance

$$d = \sqrt{\sum (x_i - y_i)^2}$$

- Most common for numerical data
- Sensitive to scale and outliers

2. Manhattan Distance

$$d = \sum | x_i - y_i |$$

- Less sensitive to outliers
- Good for grid-like data

3. Cosine Similarity

- Measures angle between vectors
- Ideal for text data, high-dimensional datasets

4. Jaccard Index

- Measures similarity for binary/categorical sets

12.3 K-Means Clustering (Pages 6–9)

K-means is the **most widely used** clustering method.

How K-Means Works (Page 6)

1. **Initialize** K random centroids

2. **Assign** each point to the nearest centroid
3. **Update** centroids by averaging points in each cluster
4. **Iterate** until stable

Goal

Minimize **within-cluster variance** and maximize **between-cluster difference**.

Key Points (Page 6–7)

- Unsupervised learning
- Sensitive to initial centroid placement
- Assumes spherical clusters
- Requires predefined number of clusters (K)
- Works best with **numerical** data

Real-World Business Examples (Pages 7–8)

- Customer segmentation for marketing
- Market segmentation by region
- Retail inventory grouping
- Churn prediction
- Banking: credit risk grouping
- Healthcare: patient clusters
- E-commerce personalization
- Real estate pricing clusters
- Supply chain optimization
- Social media audience grouping

These examples demonstrate the broad applicability of clustering.

12.3.4 Assumptions of K-Means (Page 8)

1. **Spherical clusters** (similar shape)
2. **Equal cluster variance**

3. **Euclidean distance** is appropriate
4. **Continuous numerical data**
5. Must **pre-set K**
6. Hard clustering (each point belongs to only one cluster)

12.3.5 Conceptual Problem – Telecom Company (Pages 9–12)

A telecom company wants to build personalized campaigns by clustering customers using variables like:

Numerical variables selected (Page 9–10):

- Age
- Monthly revenue
- Data usage
- Call duration
- Customer tenure
- Customer support calls

Categorical variables excluded

- Gender
- Contract type
- Plan type
  (Because distance metrics for K-means require numeric data)

Interpreting Clusters (Pages 10–12)

The analysis produced **4 clusters**, each representing a customer persona.

Cluster 1 – High-Revenue, High-Data Users, Long Tenure

- High spending
- Heavy data usage
- Loyal customers
  Strategy: Retention programs, premium offers

Cluster 2 – Low-Revenue, High Support Call Customers

- Low spend

- Many support issues
- On basic plans
  Strategy: Improve service quality, address service issues, upgrade incentives

Cluster 3 – Moderate Revenue, Family Plan Users

- Medium data usage
- Mid-level spenders
  Strategy: Promote family plans and data-sharing features

Cluster 4 – High-Revenue, Voice-Dominant Users with Long Contracts

- Prefer voice calls over data
- High loyalty
  Strategy:
- Maintain voice quality
- Encourage data usage through promotional plans

Table: Number of Clusters (Page 12)

(As provided in the document)

Cluster	Count
1	4
2	3
3	2
4	1

Final Cluster Centers (Page 12)

(Copied from image on Page 12)

Cluster Centers

Variable	C1	C2	C3	C4
Monthly Revenue	50	110	83	40
Data Usage	6	23	13	4

Variable	C1	C2	C3	C4
Call Duration	128	290	190	250
Customer Tenure	17	54	30	48
Customer Support Calls	1.255	2.33	1	3

These centers represent typical customers inside each cluster.

📌 **Summary of Week 12**

- Cluster analysis is a **powerful unsupervised method** for grouping similar observations.
- Segmentation + clustering = highly actionable business insights.
- K-means is the most common clustering algorithm.
- Choosing correct variables, scaling data, and handling outliers is essential.
- Cluster interpretation enables targeted marketing and strategic decision-making.
