SLDS Honours Theory

Explain Categorical data and quantitative data.

📊 1. Categorical Data

✅ Definition:

Categorical data represents **qualities or characteristics**. It places data into **distinct categories** or groups.

✅ Key Characteristics:

- Values are **labels or names**
- **No numerical meaning** (usually)
- Can't be used in arithmetic operations (e.g., you can't "average" colors)

✅ Types of Categorical Data:

- **Nominal**: No natural order
  Example: Gender (Male, Female), Colors (Red, Blue, Green)
- **Ordinal**: Has a natural order
  Example: Education Level (High School, Bachelor's, Master's, PhD), Rankings (1st, 2nd, 3rd)

✅ Examples:

Variable	Type
Marital Status	Nominal
Customer Rating (Good, Better, Best)	Ordinal
Blood Type	Nominal

📈 2. Quantitative Data

✅ Definition:

Quantitative data represents **numerical values** that can be **measured or counted**.

✅ Key Characteristics:

- Values are **numeric**
- Can be used in arithmetic operations
- Tells **how much**, **how many**, or **how often**

✅ Types of Quantitative Data:

- **Discrete**: Countable numbers
  Example: Number of children, cars, students
- **Continuous**: Measurable values
  Example: Height, weight, time, temperature

✅ Examples:

Variable	Type
Age (in years)	Continuous
Number of books	Discrete
Income (in ₹)	Continuous

🔄 Comparison Summary

Feature	Categorical Data	Quantitative Data
Nature	Descriptive	Numerical
Can do math?	No	Yes
Subtypes	Nominal, Ordinal	Discrete, Continuous
Examples	Gender, Color, Rank	Age, Income, Distance
Visualized by	Bar chart, Pie chart	Histogram, Box plot

Define Binomial distribution and Poisson distribution.

🎯 1. Binomial Distribution

✅ Definition:

A **Binomial Distribution** models the number of **successes** in a **fixed number of independent trials**, where each trial has only **two possible outcomes**: success or failure.

✅ Key Conditions:

- Fixed number of trials: `n`
- Two possible outcomes: **Success** or **Failure**
- Constant probability of success: `p`
- Trials are **independent**

✅ Probability Formula:

$$P(X = r) = nCr \cdot p^r \cdot (1-p)^{n-r}$$

Where:

- $P(X = r)$: Probability of getting exactly `r` successes
- nCr: Number of combinations (ways to choose `r` successes from `n` trials)
- p: Probability of success
- $1 - p$: Probability of failure

✅ Mean and Variance:

- Mean $m = n \cdot p$
- Variance $\sigma^2 = n \cdot p \cdot (1-p)$

Variance is also given as "n.p.q"

📌 2. Poisson Distribution

✅ Definition:

The **Poisson Distribution** models the number of **events** that occur in a **fixed interval of time or space**, given the events occur **independently** and at a **constant average rate**.

✅ Key Conditions:

- Events occur **randomly and independently**
- Average rate (mean) of events in an interval is **known and constant**
- Used when number of trials `n` is large and probability `p` is small

---

✅ Probability Formula:

$$P(X = r) = \frac{e^{-m} \cdot m^r}{r!}$$

Where:

- $P(X = r)$: Probability of exactly `r` events occurring
- m: Mean number of events (previously denoted as λ)
- e: Euler's number ≈ 2.718
- r: Number of occurrences
- $r!$: Factorial of `r`

✅ Mean and Variance:

- **Mean** = m
- **Variance** = m

🔄 Comparison Summary

Feature	Binomial Distribution	Poisson Distribution
Application	Fixed number of **trials**	Count of events in **continuous interval**
Parameters	n, p, Mean $m = n \cdot p$	Mean m (average events per interval)
Formula	$P(X = r) = nCr \cdot p^r \cdot (1-p)^{n-r}$	$P(X = r) = \frac{e^{-m} \cdot m^r}{r!}$
Mean and Variance	Mean = m, Variance = $m \cdot (1-p)$	Mean = m, Variance = m

Explain Type1 and Type 2 error in detail.

✅ Hypothesis Testing Background

In hypothesis testing, we start with two hypotheses:

- **Null Hypothesis (H_0):** The default or initial assumption (e.g., "the product is not defective").
- **Alternative Hypothesis (H_1 or Ha):** The claim we want to test (e.g., "the product is defective").

Based on sample data, we either:

- **Reject H_0** (if evidence supports H_1)
- **Fail to reject H_0** (if evidence is not strong enough)

❗ 1. Type I Error

🔸 Definition:

A **Type I error** occurs when we **reject the null hypothesis (H_0) even though it is actually true**.

🔸 In Simple Terms:

We **think there is an effect or change** when **in reality, there isn't**.

🔸 Example:

A medical test indicates a patient has a disease (positive result), but in reality, the patient is healthy.

🔸 Probability of Type I Error:

- Denoted by **α (alpha)**
- Called **level of significance**
- Common values: **0.05 (5%)**, **0.01 (1%)**

α = P(Type I Error) = P(rejecting H_0 | H_0 is true)

⚠️ 2. Type II Error

🔸 Definition:

A **Type II error** occurs when we **fail to reject the null hypothesis (H_0) even though it is actually false**.

🔸 In Simple Terms:

We **miss a real effect or change** that **does exist**.

🔸 Example:

A medical test fails to detect a disease (negative result), but in reality, the patient does have the disease.

🔸 Probability of Type II Error:

- Denoted by **β (beta)**

β = P(Type II Error) = P(not rejecting H_0 | H_1 is true)

🔄 Summary Table

Error Type	Description	Consequence	Probability Symbol
Type I	Rejecting a true H_0	False alarm	**α**
Type II	Not rejecting a false H_0	Missed detection	**β**

🎯 Minimizing Errors

- **Lower α (Type I Error)** → Reduce chances of false alarm, but might increase β
- **Increase sample size** → Reduces both errors and increases test **power**
- **Test power** = 1 – β → Higher power means better ability to detect true effect

---

📌 Example Scenario: Drug Effectiveness

Situation	Conclusion from Test	Actual Truth	Error Type
Drug is effective (H_0 false)	Do not reject H_0	Drug really works	**Type II Error**
Drug is not effective (H_0 true)	Reject H_0	Drug doesn't work	**Type I Error**

✅ **Type I Error (False Positive):**

1. **Definition**: It occurs when we **reject a true null hypothesis**.
2. **Meaning**: We conclude there is an effect or difference, when in reality, **there isn't**.
3. **Symbol**: Represented by **α (alpha)** – the significance level.
4. **Example**: A doctor says a healthy person is sick.
5. **Impact**: Leads to unnecessary action or treatment.

✅ **Type II Error (False Negative):**

1. **Definition**: It occurs when we **fail to reject a false null hypothesis**.
2. **Meaning**: We conclude there is **no effect**, when actually **there is** one.
3. **Symbol**: Represented by **β (beta)**.
4. **Example**: A doctor says a sick person is healthy.
5. **Impact**: Leads to missing an important finding or issue.

🔄 **Summary:**

Error Type	Action Taken	Reality	Mistake Made
Type I	Reject H_0	H_0 is true	False Alarm (False Positive)
Type II	Do not reject H_0	H_0 is false	Missed Detection (False Negative)

Define the following key terms for simple linear regression. i)Response ii) Record iii) Independent variable iv) Regression co-efficient v) Residuals

i) Response (Dependent Variable)

- It is the **output variable** we are trying to **predict or explain**.
- It **depends** on the independent variable.
- Example: If predicting marks based on study hours, **marks** are the response.

---

ii) Record

- A record is a **single observation or data point** in the dataset.
- It includes values for both independent and dependent variables.
- Example: One student's study hours and their corresponding marks.

---

iii) Independent Variable (Predictor)

- It is the **input variable** used to **predict** the response.
- It is assumed to be **not affected** by the response variable.
- Example: In predicting marks, **study hours** is the independent variable.

---

iv) Regression Coefficient (Slope)

- It represents the **rate of change** of the response with respect to the independent variable.
- It tells us **how much the response variable changes** for a one-unit change in the predictor.
- Found using the regression line: **Y = a + bX**, where **b** is the regression coefficient.

---

v) Residuals

- Residuals are the **differences between actual and predicted values** of the response variable.
- Formula: **Residual = Actual value – Predicted value**
- Smaller residuals mean better model fit.

Brief the steps in multinomial distribution goodness of fit. Elaborate the steps with an example

✅ Multinomial Distribution – Goodness of Fit Test (Chi-Square Test)

This test checks whether observed categorical data matches an expected distribution across **more than two categories**.

🔹 Steps of the Multinomial Goodness of Fit Test:

1. **State the Null and Alternative Hypothesis**
   - H0: Observed data fits the expected distribution.
   - H1: Observed data does **not** fit the expected distribution.

2. **Calculate Expected Frequencies**
   - Use:

$$E_i = n \times p_i$$

   where:
   - n = total number of observations
   - p_i = expected probability for category i

3. **Compute the Chi-Square Test Statistic**
   - Use:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

   where O_i = observed frequency and E_i = expected frequency for category i

4. **Determine the Degrees of Freedom**
   - $df = k - 1$, where k = number of categories.

5. **Compare With Critical Value**
   - Find critical value from chi-square table for chosen significance level (usually 0.05).
   - If $\chi^2\text{calculated} > \chi^2\text{table}$, reject H_0.

6. **State the Conclusion**
   - Based on comparison, accept or reject the null hypothesis.

🔍 Example:

A company wants to test if customers prefer 4 product types equally.

Observed frequencies:

- A: 30
- B: 25
- C: 20
- D: 25

  Total = 100

Assuming equal preference:

Expected frequency for each = $100 \div 4 = 25$

Step-by-step Calculation:

Product	O_i	E_i	$(O_i - E_i)^2 / E_i$
A	30	25	$\frac{(30-25)^2}{25} = 1$
B	25	25	0
C	20	25	$\frac{(20-25)^2}{25} = 1$
D	25	25	0

$$\chi^2 = 1 + 0 + 1 + 0 = 2$$

- Degrees of freedom = $4 - 1 = 3$
- At $\alpha = 0.05$, critical χ^2 = 7.815

Since $2 < 7.815$, we **fail to reject** H_0.

✅ Conclusion: There's **no significant difference**; product preferences are likely equal.

📌 Final Summary:

The **multinomial goodness of fit test** helps check whether observed categorical outcomes match expected proportions. It's especially useful when dealing with **more than two categories**, like survey responses or product preferences.

Brief the steps in test of independence. Elaborate the steps with an example

✅ Test of Independence (Chi-Square Test)

The **Chi-Square Test of Independence** is used to determine whether **two categorical variables** are independent or associated.

🔹 Steps in the Test of Independence:

1. **State the Hypotheses:**
   - Null Hypothesis (H0: The two variables are **independent**.
   - Alternative Hypothesis (H1): The two variables are **dependent (associated)**.
2. **Create a Contingency Table:**
   - This table shows frequencies of occurrences for combinations of categories from the two variables.

3. **Calculate Expected Frequencies:**
   - For each cell:

$$E_{ij} = \frac{(\text{Row total}) \times (\text{Column total})}{\text{Grand total}}$$

4. **Compute the Chi-Square Test Statistic:**
   - Use:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

   where O_{ij} is the observed frequency, and E_{ij} is the expected frequency.

5. **Determine Degrees of Freedom:**
   - $$df = (r-1)(c-1)$$

   where r = number of rows, c = number of columns.

6. **Compare With Critical Value or p-value:**
   - Use the chi-square table to find the critical value at a chosen significance level (usually 0.05).
   - If $\chi^2_{calculated} > \chi^2_{table}$, reject H_0.

7. **Conclusion:**
   - Based on the result, conclude whether variables are independent or associated.

🔍 Example: Survey on Gender and Preference for a Product

	Like	Dislike	Total
Male	20	10	30
Female	30	40	70
Total	50	50	100

Step-by-Step:

Step 1: Hypotheses

- H_0: Gender and preference are independent.
- H_1: Gender and preference are dependent.

Step 2: Expected frequencies:

$$E_{Male,Like} = \frac{30 \times 50}{100} = 15 \quad E_{Male,Dislike} = \frac{30 \times 50}{100} = 15$$

$$E_{Female,Like} = \frac{70 \times 50}{100} = 35 \quad E_{Female,Dislike} = \frac{70 \times 50}{100} = 35$$

Step 3: Chi-Square statistic:

$$\chi^2 = \frac{(20-15)^2}{15} + \frac{(10-15)^2}{15} + \frac{(30-35)^2}{35} + \frac{(40-35)^2}{35}$$

$$\chi^2 = \frac{25}{15} + \frac{25}{15} + \frac{25}{35} + \frac{25}{35} = 1.67 + 1.67 + 0.71 + 0.71 = 4.76$$

Step 4: Degrees of freedom

$$df = (2-1)(2-1) = 1$$

Step 5: Critical value for df = 1 at 0.05 = 3.841

Since $4.76 > 3.841$, reject H_0

✅ **Conclusion:** There **is a significant association** between gender and product preference.

In the context of multiple linear regression. Explain what is over fitting and multi collinearity.

Overfitting in Multiple Linear Regression:

Definition:
Overfitting occurs when a regression model learns not only the underlying relationship between the independent variables (predictors) and the dependent variable, but also the **random noise** or fluctuations in the training data. As a result, the model performs well on training data but poorly on unseen or test data.

Explanation:

- In multiple linear regression, adding more predictors might seem to improve model accuracy. However, too many variables can lead the model to memorize specific data points.
- This reduces the model's **generalization ability**.
- Overfitting is especially common when:
    - The number of predictors is close to or greater than the number of observations.
    - The model includes unnecessary or irrelevant variables.

Consequences:

- High R^2 on training data, but poor prediction on test data.
- Model becomes too sensitive to small changes in data.

Solution:

- Use **Cross-Validation**, **Regularization techniques (Ridge/Lasso)**, or **Feature Selection** to prevent overfitting.

Multicollinearity in Multiple Linear Regression:

Definition:
Multicollinearity refers to a situation where **two or more independent variables** in a multiple regression model are **highly linearly correlated**.

Explanation:

- When predictors are highly correlated, they provide redundant information.
- The model struggles to determine the individual effect of each predictor on the dependent variable.
- Coefficients become **unstable** and **highly sensitive** to changes in the data.
- Standard errors of coefficients increase, making **t-tests** unreliable.

Consequences:

- Difficulty in interpreting which predictor is truly influencing the outcome.
- Coefficient signs or magnitudes might not make practical sense.

Detection:

- Check **correlation matrix** among predictors.
- Use **Variance Inflation Factor (VIF)** — if VIF > 10, multicollinearity may be a problem.

Solution:

- Remove or combine correlated variables.
- Use **Principal Component Analysis (PCA)** or **Ridge Regression** which can handle multicollinearity.

Summary Table:

Concept	Overfitting	Multicollinearity
Definition	Model learns noise in training data	Predictors are highly correlated
Problem Type	Poor generalization to new data	Unstable coefficients
Detection	Cross-validation, test error	Correlation matrix, VIF
Fix	Regularization, feature selection	Remove/reduce correlated variables

Explain TIME SERIES PATTERNS i)Horizontal Pattern ii) Trend Pattern iii)Seasonal Pattern iv)Trend and Seasonal Pattern v)Cyclical Pattern

Time Series Patterns

A **time series** is a sequence of data points measured at successive points in time, often at uniform intervals. Understanding its patterns helps in forecasting and analysis. The main types of patterns are:

i) Horizontal Pattern (Stationary Pattern)

- This pattern shows **no significant upward or downward movement** over time.
- The data fluctuates around a **constant mean** or level.
- It suggests **stability** in the variable being measured.

Example: Daily temperature in a city with a constant climate, or noise level in a quiet room.

---

ii) Trend Pattern

- A trend shows a **long-term increase or decrease** in the data.
- It reflects **persistent upward or downward movement** over time.
- Causes may include economic growth, technology adoption, or population increase.

Example: Increase in sales revenue over years due to company growth.

---

iii) Seasonal Pattern

- This pattern repeats at **regular intervals (within a year)**, such as **daily, monthly, or quarterly**.
- Caused by **seasonal factors** like weather, holidays, or school schedules.
- The variation is **predictable** and consistent.

Example: Increase in ice cream sales during summer or higher shopping during Diwali/Christmas.

---

iv) Trend and Seasonal Pattern

- This is a **combination** of both trend and seasonal components.
- The data shows **long-term growth or decline (trend)** along with **short-term repeated fluctuations (seasonality)**.

Example: E-commerce sales increasing yearly (trend), but spiking every November during festive sales (seasonality).

---

v) Cyclical Pattern

- Cyclical patterns represent **long-term fluctuations** around the trend line, **not of fixed length**.
- Often linked to **economic or business cycles** such as recession and boom.
- Unlike seasonality, cycles are **irregular in timing and duration**.

Example: Stock market performance or real estate prices over decades.

Conclusion:

Understanding these time series patterns helps analysts and businesses to make **informed decisions**, **forecast future trends**, and **identify anomalies**.

Compare descriptive and inferential statistics.

Aspect	Descriptive Statistics	Inferential Statistics
Definition	Summarizes and organizes data from a sample or population.	Makes predictions or generalizations about a population based on a sample.
Purpose	To describe the basic features of data in a study.	To draw conclusions and make decisions under uncertainty.
Techniques Used	Mean, median, mode, standard deviation, charts, graphs.	Hypothesis testing, confidence intervals, regression, etc.
Data Scope	Works only with the data available (no generalization).	Goes beyond the data to infer about a larger group.
Example	"The average height of students in a class is 165 cm."	"Based on a sample, we estimate the average height of all students in the school."

What is sampling? State and explain different sampling methods.

◆ What is Sampling? (2 marks)

Sampling is the process of selecting a portion (sample) from a larger group (population) to study and make inferences about the entire population. It helps in saving **time, cost, and effort**, especially when the population is large.

◆ Types of Sampling Methods (8 marks)

1. Sampling from a Finite Population

- The population has a **limited or fixed number of elements**.
- **Example:** Sampling 50 students from a class of 200.

2. Sampling from an Infinite Population

- The population is **too large or uncountable**.
- **Example:** Testing bulbs coming from a continuous production line.

3. Stratified Random Sampling

- Population is **divided into strata** (subgroups) based on a characteristic.
- A **random sample** is taken from each stratum.
- Ensures representation of all subgroups.

- **Example:** Dividing students by gender and sampling both groups.

4. Cluster Sampling

- The population is divided into clusters (groups), then **entire clusters are randomly selected**.
- Cost-effective and easy when population is widely spread.
- **Example:** Selecting 3 departments in a university and surveying all students in them.

5. Systematic Sampling

- Every **k^{th} item** is selected from a list after choosing a random start.
- **Formula:** k=Nnk = \frac{N}{n}k=nN (where N = population size, n = sample size).
- **Example:** Selecting every 10th visitor to a website.

6. Convenience Sampling

- Samples are chosen based on **ease of access or availability**.
- Not reliable for generalizing to population.
- **Example:** Surveying people in your friend circle.

7. Judgment Sampling (Purposive Sampling)

- The researcher **intentionally selects** units they believe are most useful.
- Based on **expert judgment**.
- **Example:** Choosing experienced teachers to evaluate a new syllabus.

8. Other Sampling Methods

- Includes **Quota Sampling, Snowball Sampling, Multistage Sampling**, etc.
- Used when traditional methods are not feasible or specific needs arise.

✅ Conclusion:

Understanding various sampling methods allows researchers to choose the most appropriate one based on the **population type, data requirement, and resource constraints**.

What are Non-Parametric Tests

Non-Parametric Tests – Definition and Explanation

Non-parametric tests are statistical tests that do **not assume a specific distribution** (such as normal distribution) for the data. These tests are used when:

- The data is **ordinal**, **ranked**, or **categorical**.
- The sample size is small.
- The data **does not meet the assumptions** required for parametric tests (like equal variance or normality).
- We need to compare **medians** instead of means.

Key Features of Non-Parametric Tests:

- Do not require estimation of population parameters (mean, standard deviation).
- Based on **ranks** rather than actual values.
- More **robust** to outliers and skewed distributions.
- Used in **qualitative** or **non-numeric** data analysis.

Examples of Non-Parametric Tests:

Test	Purpose
Mann-Whitney U Test	Compares two independent samples
Wilcoxon Signed-Rank Test	Compares two related samples
Kruskal-Wallis Test	Non-parametric version of ANOVA
Spearman's Rank Correlation	Measures correlation using ranks
Chi-Square Test	Tests independence or goodness-of-fit in categorical data

---

When to Use Non-Parametric Tests:

- Data is not normally distributed.
- You are analyzing **ordinal data** (e.g., survey responses like good, average, bad).
- Sample size is **too small** for reliable parametric analysis.

Explain the various decomposition models used in time series data. Also, state which decomposition model will be appropriate for the following condition a) When the seasonal variation is relatively constant over time.

Decomposition Models in Time Series Analysis

Time series decomposition is a technique used to break down a time series into its individual components to better understand its structure. The primary components are:

1. **Trend (T)**: The long-term progression or general direction of the data (upward, downward, or flat).
2. **Seasonality (S)**: Regular patterns that repeat over a known, fixed period (e.g., daily, monthly, yearly).
3. **Irregular/Residual Component (R)**: Random noise or irregular fluctuations that cannot be explained by trend or seasonality.

There are **two main types of decomposition models** used:

---

1. Additive Decomposition Model

- Assumes that the components add together to form the time series:

$$Y_t = T_t + S_t + R_t$$

- **Assumptions**:
  - Seasonal variations are **constant** over time and do not depend on the level of the trend.
  - Best used when the magnitude of the seasonal effect **does not change** as the trend increases or decreases.

---

2. Multiplicative Decomposition Model

- Assumes that the components multiply together:

$$Y_t = T_t \times S_t \times R_t$$

- **Assumptions**:
  - Seasonal variations **increase or decrease proportionally** with the trend.
  - Best used when the seasonal effect **grows or shrinks** with the level of the series.

Hybrid Model (Optional Advanced)

- In some cases, a **log transformation** can be applied to convert a multiplicative model into an additive one:

$$\log(Y_t) = \log(T_t) + \log(S_t) + \log(R_t)$$

- This allows using additive modeling techniques on a multiplicative series.

(a) Appropriate Model for Constant Seasonal Variation

If the **seasonal variation is relatively constant over time**, the **Additive Decomposition Model** is the appropriate choice.

Summary Table

Condition	Appropriate Model
Seasonal variation is constant	**Additive**
Seasonal variation changes with trend level	**Multiplicative**

Elaborate moving average and exponential smoothing techniques?

What is Smoothing?

Smoothing is a technique used in time series analysis to reduce random variation or noise and highlight underlying patterns such as trend or seasonality. It helps in making the data more interpretable and suitable for forecasting.

1. Moving Average (MA)

✅ Definition:

The **Moving Average** technique smooths a time series by calculating the average of values over a fixed number of past periods, often referred to as the "window size" or "span".

✅ Formula:

For a time series $Y = \{y_1, y_2, y_3, ..., y_t\}$, a **Simple Moving Average** of window size n at time t is:

$$MA_t = \frac{y_{t-n+1} + y_{t-n+2} + \cdots + y_t}{n}$$

✅ Types:

- **Simple Moving Average (SMA)**: Equal weights to all observations in the window.
- **Weighted Moving Average (WMA)**: Assigns different weights (often more to recent values).
- **Centered Moving Average**: Aligns the moving average centrally to capture trend better.

✅ Pros:

- Easy to compute and understand.
- Good for removing short-term fluctuations.

✅ Cons:

- Lags behind actual values.
- Doesn't handle seasonality or trend well unless specifically adjusted.

---

2. Exponential Smoothing (ES)

✅ Definition:

Exponential Smoothing is a forecasting method that gives **exponentially decreasing weights** to older observations. More recent observations are weighted more heavily.

✅ Types and Formulas:

a) **Simple Exponential Smoothing (SES)**:

Used when the data has **no trend or seasonality**.

$$S_t = \alpha y_t + (1 - \alpha) S_{t-1}$$

- S_t: Smoothed value at time t
- y_t: Actual value at time t
- $\alpha \in (0, 1)$: Smoothing constant

b) **Holt's Linear Trend Method**:

Used when the data has a **trend** but no seasonality.

- Two equations:
  - Level: $L_t = \alpha y_t + (1-\alpha)(L_{t-1} + T_{t-1})$
  - Trend: $T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$
  - Forecast: $\hat{y}_{t+h} = L_t + hT_t$

c) Holt-Winters Method:

Used when the data has both **trend and seasonality**.

- Can be **additive** or **multiplicative** depending on the nature of the seasonal component.

✅ Pros:

- Adaptable to different data patterns.
- Handles level, trend, and seasonality (with Holt-Winters).
- More responsive to changes than moving average.

✅ Cons:

- Requires careful selection of parameters (α,β,γ).
- More complex than moving averages.

🔄 Comparison

Feature	Moving Average	Exponential Smoothing
Weights	Equal (or manual for WMA)	Exponentially decreasing
Reacts to recent changes	Slowly	More quickly
Handles Trend	No	Holt's method
Handles Seasonality	No	Holt-Winters method
Complexity	Simple	Moderate to complex

Define sampling and central limit theorem ? Elaborate stratified sampling, judgment sampling , systematic sampling and cluster sampling

1. Sampling and Central Limit Theorem

✅ What is Sampling?

Sampling is the process of selecting a subset of individuals, items, or observations from a larger population to estimate characteristics of the whole population.

- Saves time, cost, and effort.
- Should represent the population accurately for valid conclusions.

✅ Central Limit Theorem (CLT)

The **Central Limit Theorem** is a fundamental statistical concept which states:

"Regardless of the shape of the population distribution, the sampling distribution of the sample mean approaches a normal distribution as the sample size increases (typically n ≥ 30), provided the samples are independent and identically distributed."

📌 Key Points:

- Mean of sampling distribution = population mean (μ)
- Standard deviation of sampling distribution (Standard Error) = $\sigma / \sqrt{n}$
- Becomes approximately **normal** as $n \to \infty$

---

2. Types of Sampling Methods

✅ a) Stratified Sampling

- **Definition**: The population is divided into **homogeneous subgroups (strata)** based on a specific characteristic (e.g., age, gender), and **random samples** are taken from each stratum.
- **Use Case**: When population has distinct groups.
- **Example**: Divide students into classes (strata) and randomly pick from each class.

🟢 **Advantages**:

- Increases representativeness.

- Improves precision of estimates.

✅ b) Judgment Sampling (Purposive Sampling)

- **Definition**: Samples are selected **based on the researcher's judgment** about which elements will be the most useful or representative.
- **Use Case**: When expertise or prior knowledge is used to select typical or critical cases.
- **Example**: Choosing only experienced doctors for a survey on advanced surgical procedures.

🔴 **Disadvantages**:

- Can introduce **bias**.
- Not statistically generalizable.

✅ c) Systematic Sampling

- **Definition:** Select every **k-th** element from a list or ordered population after a random starting point.
- Formula for step size: $k = \frac{N}{n}$, where N is population size and n is sample size.
- **Example:** From a list of 1,000 names, selecting every 10th person starting from the 5th.

🟢 **Advantages:**

- Simple and quick.
- Ensures evenly spread sample.

🔴 **Disadvantage:**

- Can be biased if there's a hidden pattern in the data list.

✅ d) Cluster Sampling

- **Definition**: The population is divided into **clusters** (often geographically), and **entire clusters are randomly selected**. All individuals within selected clusters are surveyed.
- **Example**: Randomly selecting 5 schools (clusters) out of 50 and surveying all students in each.

🟢 **Advantages**:

- Cost-effective for large populations.
- Useful when a complete list of individuals is unavailable.

🔴 **Disadvantages**:

- Higher sampling error than stratified or simple random sampling.

✅ Summary Table

Sampling Method	Key Feature	Best For
Stratified Sampling	Divide population by groups, sample from each	Heterogeneous population
Judgment Sampling	Researcher selects based on expertise	Expert or critical cases
Systematic Sampling	Select every k-th element	Ordered lists
Cluster Sampling	Randomly select whole clusters	Geographically spread large populations

Explain any 3 numerical measures for : a. Measures of variability

b. Measures of location c. Measures of distribution shape

a) Measures of Variability

1. **Range**
   Range is the simplest measure of variability. It is calculated as the difference between the maximum and minimum values in the dataset.
   Formula:

$$\text{Range} = \text{Max} - \text{Min}$$

   Description: It gives a quick sense of the spread but is highly sensitive to outliers.

2. **Variance**
   Variance measures the average of the squared differences between each data point and the mean.
   Formula:

$$\sigma^2 = \frac{1}{n}\sum(x_i - \mu)^2$$

Description: A higher variance indicates greater data dispersion. It is a foundational concept in statistics.

3. **Standard Deviation (SD)**
   Standard deviation is the square root of the variance. It expresses how much data points typically differ from the mean.
   Formula:

$$\sigma = \sqrt{\frac{1}{n}\sum(x_i - \mu)^2}$$

Description: Widely used in data analysis and probability. Lower SD means data is clustered; higher SD means more spread out.

4. **Interquartile Range (IQR)**
   IQR is the range of the middle 50% of the data.
   Formula:

$$IQR = Q3 - Q1$$

Description: It removes the influence of outliers and is useful for understanding data consistency.

5. **Coefficient of Variation (CV)**
   CV is the ratio of the standard deviation to the mean, expressed as a percentage.
   Formula:

$$\text{CV} = \frac{\sigma}{\mu} \times 100\%$$

Description: Useful to compare the variability of datasets with different units or means.

---

b) Measures of Location (Central Tendency)

1. **Mean**
   The mean is the arithmetic average of a dataset.
   Formula:

$$\bar{x} = \frac{1}{n}\sum x_i$$

Description: It uses all data values but can be heavily influenced by outliers.

2. **Median**
   The median is the middle value when data is arranged in order.

Description: If the number of values is even, it is the average of the two middle values. It is not affected by outliers and is better for skewed distributions.

3. **Mode**
   The mode is the value that appears most frequently in the dataset.
   Description: There can be one mode, more than one (bimodal or multimodal), or none. It's useful for categorical and nominal data.
4. **Percentiles**
   A percentile indicates the value below which a given percentage of observations fall.
   Description: For example, the 75th percentile (P75) means 75% of data lies below that value. Common in exams and scores.
5. **Quartiles**
   Quartiles are special percentiles that divide data into four equal parts:

- Q1 = 25th percentile
- Q2 = 50th percentile (Median)
- Q3 = 75th percentile
  Description: They are used in box plots and IQR calculation to assess spread and central location.

---

c) Measures of Distribution Shape

1. **Skewness**
   Skewness measures the asymmetry of the data distribution.
   Formula (sample):

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Description: A skewness of 0 implies symmetry. Positive skew has a longer right tail; negative skew has a longer left tail.

2. **Kurtosis**
   Kurtosis measures the "tailedness" or peak of the distribution.
   Description: High kurtosis (leptokurtic) means heavy tails and sharp peak; low kurtosis (platykurtic) indicates light tails and flatter peak.
3. **Z-score**
   A Z-score indicates how many standard deviations a data point is from the mean.
   Formula:

$$Z = \frac{x - \mu}{\sigma}$$

Description: Useful for identifying outliers and standardizing data. Z-scores above 3 or below -3 often indicate outliers.

4. **Coefficient of Skewness**
   This is a quick estimate of skewness using mean, median, and standard deviation.
   Formula:

$$\text{CS} = \frac{3(\bar{x} - \text{Median})}{\sigma}$$

Description: Helps assess whether the data is right or left skewed. Positive = right-skewed, negative = left-skewed.

5. **Histogram Shape (visual + numeric)**
   While primarily graphical, histograms help identify the shape of the data distribution such as normal, skewed, or uniform.
   Description: This complements numerical measures and helps in quick visual interpretation of distribution shape.

Difference between a. Parametric and non-parametric test b. Discrete and Continuous probability distribution.

a. Parametric vs Non-Parametric Tests

Aspect	Parametric Test	Non-Parametric Test
1. **Definition**	Assumes underlying statistical distribution (e.g., normal)	No assumption about population distribution
2. **Assumptions**	Requires assumptions like normality, homogeneity of variance	Few or no assumptions required
3. **Data Type**	Interval or ratio scale (quantitative data)	Ordinal, nominal, or non-normal interval/ratio data
4. **Sample Size**	More suitable for large samples	Can be used with small samples
5. **Test Statistic**	Based on mean and standard deviation	Based on ranks, medians, or frequencies
6. **Power of Test**	More powerful when assumptions are met	Less powerful but more robust when assumptions are violated
7. **Examples**	t-test, z-test, ANOVA, Pearson correlation	Mann-Whitney U, Kruskal-Wallis, Chi-square, Spearman correlation
8. **Computational Complexity**	Simpler formulas and more standard tools	May involve ranking or resampling methods
9. **Outlier Sensitivity**	Sensitive to outliers and non-normality	Less sensitive to outliers
10. **Use Case**	When population parameters (mean, SD) are known or can be estimated	When distribution is unknown or data is ordinal

---

b. Discrete vs Continuous Probability Distributions

Aspect	Discrete Probability Distribution	Continuous Probability Distribution
1. **Definition**	Probability of **countable** outcomes	Probability of **uncountable/infinite** outcomes
2. **Values Taken**	Specific, separate values (e.g., 0, 1, 2…)	Any value in a continuous range (e.g., 2.15, 3.67…)
3. **Probability Function**	Probability Mass Function (PMF)	Probability Density Function (PDF)
4. **Probability of Exact Value**	Greater than zero (e.g., P(X = 2) = 0.3)	Always zero (P(X = 2) = 0); probabilities are over intervals
5. **Graph Type**	Bar graph	Smooth curve
6. **Area Interpretation**	Sum of probabilities = 1	Area under curve = 1
7. **Examples**	Binomial, Poisson, Geometric	Normal, Exponential, Uniform
8. **Use Case**	Counting events like number of students, defects	Measuring time, height, weight, etc.
9. **Mathematical Handling**	Uses summation	Uses integration
10. **Real-world Example**	Tossing a coin, number of calls received	Temperature readings, time taken to finish a task