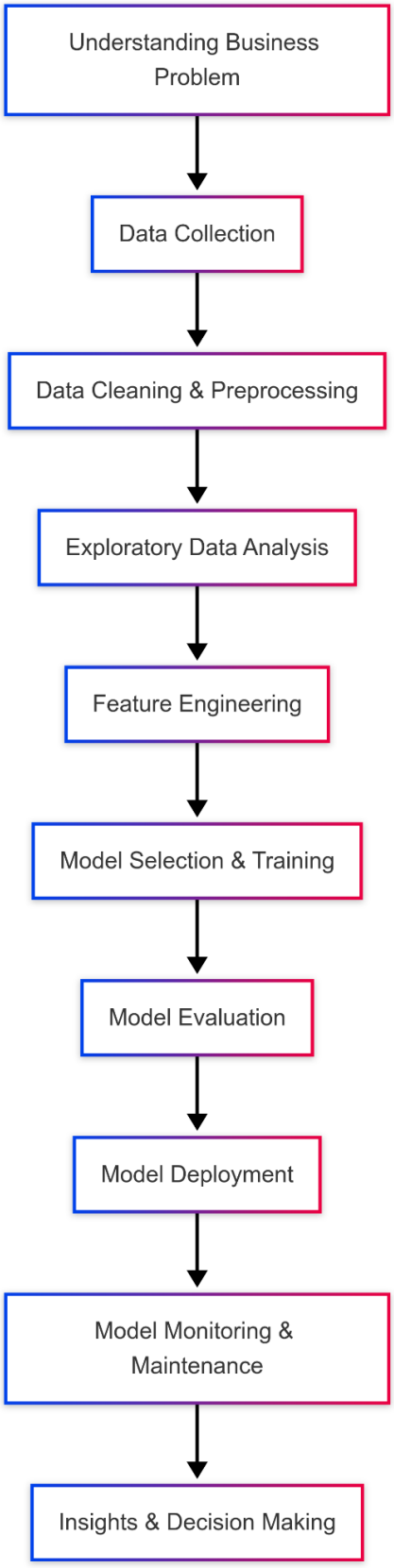
Statistics notes-Day1

Contents

Data analytics life cycle



Statistics:-

Statistics is the branch of mathematics that deals with collecting, organizing, analyzing, interpreting, and presenting data. It helps in making informed decisions based on numerical evidence.

Types of Statistics

1. **Descriptive Statistics – Summarizes and describes data.**
    - **Measures of Central Tendency: Mean, Median, Mode**
    - **Measures of Dispersion: Variance, Standard Deviation, Range**
    - **Data Visualization: Histograms, Pie Charts, Box Plots**

2. **Inferential Statistics – Makes predictions or inferences about a population based on a sample.**
    - **Hypothesis Testing (t-tests, Chi-square test, ANOVA)**
    - **Confidence Intervals**
    - **Regression Analysis**
    - **Probability Distributions (Normal, Binomial, Poisson)**

Data

Data refers to raw facts, figures, or information collected for analysis. It can be numerical (quantitative) or categorical (qualitative).

1. Population Data

- **Definition: Population data includes information collected from every individual or item in the entire group being studied.**
- **Example: If you want to study the average salary of all data analysts in India, the population data would include the salaries of every data analyst in India.**

🔹 **Key Point: Population data provides accurate results, but collecting it can be expensive and time-consuming.**

---

2. Sample Data

- **Definition: Sample data is a subset of the population, chosen to represent the whole group.**
- **Example: Instead of surveying all data analysts in India, you collect salary data from 1,000 randomly selected data analysts.**

🔹 **Key Point: Sample data is used when population data is too large to collect. Proper sampling ensures reliable estimates of the whole population.**

Real-World Example in Data Analytics

📊 **Scenario: A pharmaceutical company wants to test the effectiveness of a new drug.**

- **Population Data: Testing the drug on every patient with the condition worldwide.**
- **Sample Data: Testing the drug on 1,000 randomly selected patients from different hospitals**

Sampling Techniques

Sampling techniques are methods used to select a subset (sample) from a larger population for study. These techniques fall into two main categories:

A. Probability Sampling (Random Selection)

Definition: Every individual in the population has an equal chance of being selected. This method reduces bias and ensures representativeness.

Types of Probability Sampling:

1. **Simple Random Sampling**
   - **Each individual is chosen randomly, like drawing names from a hat.**
   - ✅ **Example: A lottery system selects 100 employees from a company for a survey.**
2. **Stratified Sampling**
   - **The population is divided into subgroups (strata) based on characteristics (e.g., age, gender), and samples are taken from each.**
   - ✅ **Example: A school selects students from each grade (stratum) to study academic performance.**
3. **Systematic Sampling**
   - **Every nth individual is selected from a list.**
   - ✅ **Example: A researcher surveys every 10th customer entering a mall.**
4. **Cluster Sampling**
   - **The population is divided into clusters (groups), and some clusters are randomly chosen.**
   - ✅ **Example: Instead of surveying all schools in a city, researchers randomly select 10 schools and survey all students in those schools.**

B. Non-Probability Sampling (Non-Random Selection)

Definition: Not all individuals have an equal chance of being selected. This method is easier but can introduce bias.

Types of Non-Probability Sampling:

1. **Convenience Sampling**
   - **Selecting individuals who are easily accessible.**
   - **❌ Can be biased as it doesn't represent the entire population.**
   - **✅ Example: A researcher surveys shoppers at a single mall about their shopping habits.**
2. **Judgmental (Purposive) Sampling**
   - **The researcher selects participants based on their expertise or specific characteristics.**
   - **✅ Example: Interviewing doctors about a new medical treatment.**
3. **Quota Sampling**
   - **Ensures certain quotas (e.g., 50% males, 50% females) are met.**
   - **✅ Example: A survey requires 30% responses from urban and 70% from rural areas.**
4. **Snowball Sampling**
   - **Used for hard-to-reach populations, where participants refer others.**
   - **✅ Example: Studying drug users by asking existing participants to refer others.**

4. Sampling Techniques

Sampling techniques are methods used to select a subset (sample) from a larger population for study. These techniques fall into two main categories:

A. Probability Sampling (Random Selection)

- **Definition: Every individual in the population has an equal chance of being selected. This method reduces bias and ensures representativeness.**

Types of Probability Sampling:

1. **Simple Random Sampling** 🎲
    - **Each individual is chosen randomly, like drawing names from a hat.**
    - **✅ Example: A lottery system selects 100 employees from a company for a survey.**
2. **Stratified Sampling** 📊
    - **The population is divided into subgroups (strata) based on characteristics (e.g., age, gender), and samples are taken from each.**
    - **✅ Example: A school selects students from each grade (stratum) to study academic performance.**
3. **Systematic Sampling** 🔢
    - **Every nth individual is selected from a list.**
    - **✅ Example: A researcher surveys every 10th customer entering a mall.**
4. **Cluster Sampling** 🏘️
    - **The population is divided into clusters (groups), and some clusters are randomly chosen.**
    - **✅ Example: Instead of surveying all schools in a city, researchers randomly select 10 schools and survey all students in those schools.**

B. Non-Probability Sampling (Non-Random Selection)

🔹 **Definition: Not all individuals have an equal chance of being selected. This method is easier but can introduce bias.**

Types of Non-Probability Sampling:

1. **Convenience Sampling** 🏪
    - **Selecting individuals who are easily accessible.**

- ❌ **Can be biased as it doesn't represent the entire population.**
- ✅ **Example: A researcher surveys shoppers at a single mall about their shopping habits.**

2. **Judgmental (Purposive) Sampling 🧐**
   - **The researcher selects participants based on their expertise or specific characteristics.**
   - ✅ **Example: Interviewing doctors about a new medical treatment.**
3. **Quota Sampling 📏**
   - **Ensures certain quotas (e.g., 50% males, 50% females) are met.**
   - ✅ **Example: A survey requires 30% responses from urban and 70% from rural areas.**
4. **Snowball Sampling ❄️**
   - **Used for hard-to-reach populations, where participants refer others.**
   - ✅ **Example: Studying drug users by asking existing participants to refer others.**

Probability vs. Non-Probability Sampling

Feature	Probability Sampling	Non-Probability Sampling
Selection Method	Random	Non-Random
Bias	Low	High
Accuracy	High	Lower (depends on method)
Example	Lottery draw	Asking friends for opinions

Variable

A variable is any characteristic, number, or quantity that can change or vary in a dataset. Variables are classified into two main types:

A. Quantitative Variable (Numerical Data) 📊

Definition: A variable that represents numeric values that can be measured or counted.

Types of Quantitative Variables:

1. **Discrete Variable (Countable)**
   - **Takes only specific values (whole numbers).**
   - **✅ Example: Number of students in a class (10, 25, 50).**
2. **Continuous Variable (Measurable)**
   - **Can take any value within a range (including decimals).**
   - **✅ Example: Height (5.7 ft), Weight (62.5 kg), Temperature (36.8°C).**

B. Qualitative Variable (Categorical Data)

Definition: A variable that represents categories or labels instead of numbers.

Types of Qualitative Variables:

1. **Nominal Variable (No Order)**
   - **Categories without a specific order.**
   - **✅ Example: Gender (Male, Female), Eye Color (Blue, Brown, Green).**
2. **Ordinal Variable (Ordered Categories)**
   - **Categories with a meaningful order but no fixed difference.**
   - **✅ Example: Education Level (High School, Bachelor's, Master's, PhD), Customer Satisfaction (Poor, Average, Good, Excellent).**

Scales of measurement define how variables are categorized, measured, and interpreted in statistical analysis. There are four main scales of measurement:

---

1. Nominal Scale (Categorical, No Order)

🔹 **Definition: Data is categorized without any ranking or order. These are labels or names with no mathematical meaning.**

✅ **Examples:**

- **Gender (Male, Female, Other)**
- **Blood Type (A, B, AB, O)**
- **Eye Color (Blue, Green, Brown)**

📌 **Key Features:**

- **No numerical value or order**
- **Only mode can be used for analysis**

2.Ordinal Scale (Categorical, Ordered)

🔹 **Definition: Data is categorized with a meaningful order, but the difference between values is not measurable.**

✅ **Examples:**

- **Education Level (High School, Bachelor's, Master's, PhD)**
- **Customer Satisfaction (Poor, Average, Good, Excellent)**
- **Military Ranks (Private, Sergeant, Captain)**

📌 **Key Features:**

- **Ranking is meaningful but differences are not uniform**
- **Mode and median can be used for analysis**

---

3. Interval Scale (Numerical, No True Zero)

🔹 **Definition: Data is numeric, differences are meaningful, but there is no true zero (zero does not mean "absence" of the quantity).**

✅ **Examples:**

- **Temperature in Celsius or Fahrenheit (0°C does not mean "no temperature")**
- **IQ Scores (A person with an IQ of 120 is not "twice as smart" as someone with 60)**
- **Year (e.g., 2000, 2020 – no "true zero" year)**

📌 **Key Features:**

- Addition & subtraction are meaningful
- No true zero, so ratios are meaningless
- Mode, median, and mean can be used for analysis

---

4. Ratio Scale (Numerical, True Zero)

🔹 Definition: Data is numeric, has equal intervals, and has a true zero (zero means "absence" of the quantity).

✅ Examples:

- Height (0 cm means no height)
- Weight (0 kg means no weight)
- Income (0 dollars means no income)
- Age (0 years means newborn)

📌 Key Features:

- All mathematical operations (addition, subtraction, multiplication, division) are meaningful
- Mode, median, mean, ratios, and percentages can be used for analysis

Scale	Type	Order	Equal Intervals	True Zero	Example
Nominal	Categorical	❌ No	❌ No	❌ No	Eye Color, Nationality
Ordinal	Categorical	✅ Yes	❌ No	❌ No	Satisfaction Level, Education
Interval	Numerical	✅ Yes	✅ Yes	❌ No	Temperature (°C, °F), IQ Score
Ratio	Numerical	✅ Yes	✅ Yes	✅ Yes	Height, Weight, Income

ASSIGNMENT STATISTICS

Section 1: Data & Sampling (10 Marks)

1. **Define data and explain the difference between population data and sample data with real-life examples. *(2 Marks)***
    - **Data: Data refers to raw facts, figures, or information collected for analysis. It can be numerical, categorical, or descriptive.**
    - **Population Data: This includes all members of a defined group. Example: The total number of employees in a multinational company.**
    - **Sample Data: A subset of the population used for analysis. Example: Surveying 500 employees from the total workforce of 10,000.**
2. **A company wants to survey customer satisfaction.**
    - **Should they collect population data or sample data? Why? *(2 Marks)***
    - **Sample Data should be collected because surveying every customer (population data) is impractical and expensive. A well-selected sample can provide reliable insights with fewer resources.**
3. **Identify the best sampling technique for the following scenarios and justify your answer: *(6 Marks)***
    - **a) A university selects 50 students randomly from each department.**
        - **Stratified Random Sampling: The population is divided into strata (departments), and students are randomly selected from each.**
    - **b) A doctor studies rare cancer cases by interviewing one patient, who refers others.**
        - **Snowball Sampling: Used for rare populations where participants recruit others with similar conditions.**
    - **c) A supermarket manager asks customers at the entrance about their shopping experience.**
        - **Convenience Sampling: Easily accessible customers are surveyed, but it may not represent the entire customer base.**

Section 2: Sampling Techniques (10 Marks)

4. **Define probability sampling and non-probability sampling. *(2 Marks)***
    - **Probability Sampling: Every individual in the population has a known and equal chance of being selected. Example: Random Sampling.**
    - **Non-Probability Sampling: Selection is based on non-random methods like convenience or judgment. Example: Snowball Sampling.**
5. **Classify the following as probability or non-probability sampling and explain why: *(4 Marks)***
    - **a) Selecting every 5th person in a queue.**
        - **Systematic Sampling (Probability Sampling): Follows a fixed interval selection pattern.**
    - **b) Interviewing only top-performing employees in a company.**
        - **Judgmental Sampling (Non-Probability Sampling): Based on a pre-determined selection criterion (performance level).**
6. **Explain two advantages and two disadvantages of non-probability sampling. *(4 Marks)***
    - **✅ Advantages:**
        1. **Faster and cost-effective.**
        2. **Useful for difficult-to-reach populations (e.g., rare diseases).**
    - **❌ Disadvantages:**
        1. **Higher risk of bias.**
        2. **Results may not be generalizable to the entire population.**

Section 3: Variables (10 Marks)

7. **Differentiate between quantitative variables and qualitative variables with two examples for each. *(4 Marks)***
    - **Quantitative Variables (Numerical Data): Can be measured numerically.**
        - **Example 1: Age (25 years, 30 years)**
        - **Example 2: Salary (₹50,000, ₹75,000)**
    - **Qualitative Variables (Categorical Data): Represents categories.**
        - **Example 1: Eye Color (Blue, Brown, Green)**
        - **Example 2: Car Brand (Toyota, Ford, Honda)**

8. **Identify the type of variable (Discrete, Continuous, Nominal, Ordinal) for each of the following:** ***(6 Marks)***
    - **a) Monthly income of employees. → Continuous**
    - **b) Favorite movie genre. → Nominal**
    - **c) Time taken to complete a task. → Continuous**
    - **d) Shirt sizes (S, M, L, XL). → Ordinal**
    - **e) Number of followers on Instagram. → Discrete**
    - **f) Ratings of a product (1 star, 2 stars, … 5 stars). → Ordinal**

Section 4: Scales of Measurement (10 Marks)

9. **Match the following variables to the correct scale of measurement (Nominal, Ordinal, Interval, Ratio):** ***(6 Marks)***
    - **a) Distance traveled by a car (in km). → Ratio**
    - **b) Clothing brands (Nike, Adidas, Puma). → Nominal**
    - **c) Temperature in Fahrenheit. → Interval**
    - **d) Movie rankings (1st, 2nd, 3rd). → Ordinal**
    - **e) Number of students in a classroom. → Ratio**
    - **f) Exam grades (A, B, C, D). → Ordinal**
10. **Explain why temperature in Celsius is an Interval scale and height is a Ratio scale.** ***(4 Marks)***

- **Temperature in Celsius (Interval Scale): The difference between values is meaningful (e.g., 20°C to 30°C), but there is no true zero (0°C does not mean 'no temperature').**
- **Height (Ratio Scale): Has a true zero (0 cm means no height), and ratios make sense (e.g., 180 cm is twice as tall as 90 cm).**