UNIT 1: STATISTICS IN DATA SCIENCE

1️⃣ WHAT IS STATISTICS?

🔹 Professional Definition

Statistics is a branch of mathematics concerned with the collection, organization, analysis, interpretation, and presentation of data for the purpose of decision-making under uncertainty.

🔹 Simple Explanation

Statistics ek aisa tool hai jo raw data ko meaningful information me convert karta hai taaki hum logical decisions le saken.

Data khud kuch nahi bolta — statistics usse bolna sikhata hai.

🔹 Role of Statistics in Data Science (Advanced Perspective)

In Data Science, statistics:

- Forms the mathematical foundation of Machine Learning
- Helps in model evaluation
- Quantifies uncertainty
- Supports hypothesis testing
- Enables predictive modeling
- Helps in feature selection
- Validates assumptions of algorithms

Without statistics:

- No probability models
- No regression analysis
- No confidence intervals
- No hypothesis testing
- No inference from sample to population

🔹 Real-Life Connections

1. Netflix Recommendation System → Uses probability & inference
2. Credit Risk Assessment → Uses inferential statistics
3. Medical Research → Drug effectiveness testing
4. Placement Analysis → Predicting future placement trends

2️⃣ TYPES OF STATISTICS

Statistics is broadly classified into:

1. Descriptive Statistics
2. Inferential Statistics

A) DESCRIPTIVE STATISTICS

◆ Professional Definition

Descriptive statistics refers to methods used to summarize, organize, and present data in a meaningful way without drawing conclusions beyond the dataset.

◆ Simple Explanation

Past data ko summarize karna.

It answers:

- What happened?
- What is the average?
- How spread out is the data?

◆ Major Components (Advanced Coverage)

1. Measures of Central Tendency

- Mean
- Median
- Mode

2. Measures of Dispersion

- Range
- Variance
- Standard Deviation
- Interquartile Range

3. Shape of Distribution

- Skewness
- Kurtosis

◆ Real-Life Example

Example: 100 students ke marks

Descriptive statistics batayega:

- Average marks kya hai
- Sabse zyada marks kya hai
- Data consistent hai ya scattered

But it will NOT predict future performance.

B) INFERENTIAL STATISTICS

🔹 Professional Definition

Inferential statistics uses sample data to make generalizations, predictions, or decisions about a larger population with quantified uncertainty.

🔹 Simple Explanation

Chhote data (sample) se bade group (population) ke baare me decision lena.

🔹 Core Concepts (Masters Level)

- Probability Theory
- Sampling Distribution
- Central Limit Theorem
- Confidence Interval
- Hypothesis Testing
- Regression Analysis
- Correlation

🔹 Real-Life Example

Company surveys 500 customers out of 50,000.

Using inferential statistics:

- Estimate overall satisfaction rate
- Predict buying behavior
- Test if new marketing strategy works

3️⃣ POPULATION AND SAMPLE

🔹 Population

Professional Definition

Population refers to the complete set of all observations or elements of interest in a particular study.

Simple Explanation

Jis pure group ko hum study karna chahte hain.

Example

- All students of your university
- All diabetic patients in India

🔹 Sample

Professional Definition

A sample is a subset of the population selected for analysis to draw conclusions about the population.

Simple Explanation

Population ka chhota representative part.

Example

- 200 students selected from university
- 1000 patients selected from hospitals

🔹 Important Concept: Parameter vs Statistic

Parameter → Population value
Statistic → Sample value

Example:

- Population Mean = μ
- Sample Mean = $\bar{x}$

4️⃣ TYPES OF SAMPLING TECHNIQUES

Sampling techniques are divided into:

1. Probability Sampling
2. Non-Probability Sampling

A) PROBABILITY SAMPLING

Every element has equal chance of selection.

1. Simple Random Sampling

Random selection using random numbers.

Example:
Lottery system.

Practical:
Excel → RAND() function
Python → random.sample()

---

2. Systematic Sampling

Select every k-th element.

Example:
Every 10th student in attendance list.

---

3. Stratified Sampling

Population divided into strata (groups).

Example:
Students divided by department, then sample taken from each.

Used in:
Market research
Medical trials

---

4. Cluster Sampling

Population divided into clusters; entire clusters selected.

Example:
Selecting 5 schools randomly and surveying all students in them.

B) NON-PROBABILITY SAMPLING

Selection based on convenience0 or judgment.

1. Convenience Sampling

Easily available data.

2. Judgment Sampling

Researcher selects based on expertise.

3. Snowball Sampling

Participants refer others.

Used in:
Social research
Rare disease studies