Day 35 – Introduction to Statistics

In this notebook, I explored the basics of **statistics**, its role in **machine learning**, and some fundamental concepts like **population vs sample** and **types of statistics**.

What is Statistics?

Statistics is the **foundation of Machine Learning (ML)**. It deals with the collection, organization, analysis, and interpretation of data. In ML, almost every model is built on statistical principles – from probability distributions to hypothesis testing and regression models.

Without statistics, building reliable ML models is nearly impossible, as statistics helps us:

- Understand patterns and variability in data
- Make informed decisions with limited information
- Validate assumptions and test hypotheses
- Build predictive models

Role of Statistics in Machine Learning

- **Data Understanding** → summarizing large datasets into meaningful insights.
- **Modeling Uncertainty** → probability and distributions help capture randomness.
- **Hypothesis Testing** → deciding whether patterns are real or random.
- **Evaluation Metrics** → errors like MSE, RMSE, R^2 are statistical concepts.
- **Feature Importance** → identifying relationships between variables.

In short: Statistics forms the backbone of every ML algorithm, from linear regression to advanced neural networks.

Population vs Sample

Whenever we hear the term "population," we often think of a large group of people. In statistics:

- **Population** → The complete set of all elements under study (people, objects, events, etc.) that share at least one common characteristic.
- **Sample** → A smaller subset of the population, selected to represent it.

For example:

- All citizens of India → Population
- 10,000 people surveyed across India → Sample

Population

- Represents the entire group (people, units, objects, events, etc.)
- Characteristic is called a **Parameter**
- Data collection → **Census** or complete enumeration
- Focus → Identifying characteristics of all elements

Types of Population:

1. **Finite Population** – countable, e.g., workers in a factory
2. **Infinite Population** – uncountable, e.g., stars in the universe
3. **Existent Population** – real and observable, e.g., cars in a city
4. **Hypothetical Population** – imagined, e.g., outcomes of rolling dice infinitely

---

Sample

- A subset of the population chosen for study
- Characteristic is called a **Statistic**
- Data collection → **Sampling** or survey
- Focus → Making inferences about the population

Properties of a good sample:

- Random selection
- Free from bias
- Representative of the population

Population vs Sample – Comparison Table

Feature	Population	Sample
Meaning	Entire group with common characteristics	Subset of population chosen for study
Includes	Every unit of the group	Only a few selected units
Characteristic	Parameter	Statistic
Data Collection	Census / Complete Enumeration	Survey / Sampling
Focus	Identify true characteristics	Make inference about population
Size	Usually very large (N)	Smaller, finite (n)

Key Differences

1. Population includes **all elements**; sample includes **only some elements**.
2. A population parameter is fixed; a sample statistic varies.
3. Census is time-consuming and costly, while sampling is faster and efficient.
4. Sample helps make **generalizations** about the population.

Conclusion

Population and sample are deeply connected – a **sample is always derived from a population**.

- The **goal of sampling** is to make accurate inferences about the population.
- Larger, unbiased samples lead to **more reliable generalizations**.

In ML, we rarely use the entire population (impractical). Instead, we train models on a sample (training data) and evaluate them on another sample (test data) to estimate performance on the population.

Types of Statistics

Statistics is broadly divided into two categories, with extensions into advanced applications:

1. **Descriptive Statistics**

- Focuses on **summarizing and organizing data**.
- Includes measures like **mean, median, mode, variance, standard deviation, and visualizations**.
- Example → Finding the average marks of students in a class.

2. **Inferential Statistics**

- Helps us make **predictions or generalizations** about a population based on a sample.
- Uses techniques like **hypothesis testing, confidence intervals, and significance tests**.
- Example → Predicting election results by surveying a sample of voters.

3. **Advanced Statistics**

- Extends inferential methods with **complex models and multivariate analysis**.
- Includes techniques like **ANOVA, MANOVA, Chi-square tests, correlation, regression, and survival analysis**.

4. **Regression Analysis**

- A predictive modeling technique that studies the **relationship between dependent and independent variables**.
- Forms the basis of many ML algorithms.
- Types include **linear regression, logistic regression, ridge, lasso, and time series forecasting**.

Statistics Overview

Concept	Focus	Examples
Descriptive	Summarizing and describing data	Mean, Median, Mode, Standard Deviation, Plots
Inferential	Drawing conclusions about population	Hypothesis Testing, Regression, Confidence Intervals

---

Summary:

- Statistics is the science of collecting, analyzing, and interpreting data.
- It is widely used in **machine learning** for model building, evaluation, and decision-making.