STATISTICS

Defination:

Statistics is the science of collecting, analyzing, and interpreting data to uncover patterns and make decisions. In data science, it acts as the backbone for understanding data and building reliable models.

- Summarizes data using measures like mean, median, and variance
- Models uncertainty with probability and distributions
- Tests hypotheses (e.g., A/B testing)
- Finds relationships through regression and correlation

Data:

Data refers to raw facts, figures, or information collected from various sources, which by themselves may not have meaning.
When this data is processed, cleaned, and analyzed, it helps to uncover patterns, insights, and support decision-making.

Applications of Statistics :

1. **Data Exploration and Summarization**

- Involves collecting, organizing, and summarizing data.
- Helps understand trends, patterns, and relationships.
- Example: Using mean, median, mode, and visualizations like histograms or box plots.

2. **Model Building and Validation**

- Statistics helps create **predictive or explanatory models**.
- Validation ensures the model works well on new/unseen data.
- Example: Regression models, classification models, etc.

3. **Statistical Analysis**

- Involves analyzing **sample data** to make conclusions about the **population**.
- Example: Using inferential statistics like confidence intervals and hypothesis tests.

4. **Hypothesis Testing**

- Used to test assumptions or claims about data.
- Helps in decision-making using probability and statistical evidence.
- Example: Testing whether a new drug is more effective than the old one.

5. **Optimization and Efficiency**

- ○ Statistics is used to optimize processes and improve performance.
- ○ Example: Reducing manufacturing defects, maximizing profit, or minimizing costs.

6. **Reporting**

- ○ Presenting analyzed data clearly through reports, dashboards, and visualizations.
- ○ Example: A data analyst summarizing insights for business decisions.

Types of Statistics:

1. Descriptive Statistics

Definition:
Descriptive statistics involves **organizing, summarizing, and presenting data** in a meaningful way.
It helps describe the **basic features** of a dataset.

Main Components:

a. **Measures of Central Tendency**
   - Mean
   - Median
   - Mode

     → These represent the *center* or *average* of data.

b. **Measures of Dispersion**
   - Variance
   - Standard Deviation

     → These show *how spread out* the data is.

c. **Data Distribution**
   - Histogram
   - Box Plot
   - Pie Chart
   - PDF (Probability Density Function) / PMF (Probability Mass Function)

     → These show how data values are distributed.

d. **Summary Statistics**
   - Five-number summary: **Minimum, Q1, Median (Q2), Q3, Maximum**

     → Gives a quick snapshot of dataset spread and extremes.

2. Inferential Statistics

Definition:
Inferential statistics helps in making **predictions or generalizations** about a **population** using data from a **sample**.
It involves drawing conclusions and testing hypotheses.

Key Concepts:

- **Sample → Population (Inference or Conclusion)**

Main Components:

a. **Hypothesis Testing** – Checking if assumptions about population are true.
b. **P-value** – Probability that results happened by chance.
c. **Confidence Interval** – Range within which the true population parameter likely falls.
d. **Statistical Analysis Tests:**
    - Z-test
    - t-test
    - ANOVA (F-test)
    - Chi-square test

Population And Sample Data

Population	sample
defination: a population is the entire set of individuals or objects of intrest in a particular study. it includes all members of a defined group that we are studying or collecting information on.	defination: a sample is a subset of the population that is used to represent the entire group. sampling involves selecting a group of individual or observations from the population to draw conclusions about the whole population.
characteristics: 1} complete set: contains all the observation of interest. 2} parameter: a numerical value summarizing the entire population. ● population mean ● population variance	characteristics: 1} subset: represent a portion of the population. 2} statistics: a numerical value summarizing the sample data. ● sample mean ● sample varience ● random sampling : sample should be randomly selected to avoid bias.
examples: 1}Population in a school study ● All students enrolled in a school. ● Determine the avrage height of student, population mean. 2} Population in Market Research ● All consumers in a city. ● To understand the purchasing behaviour of all consumers. 3} population in a medical study ● All the patients with a specific disease. ● To study the effectiveness of a drug.	examples 1} Sample in a school study ● A group of 50 students from school. ● Usecase: estimate the average height of students in a school. 2} Sample in a Market Research ● 500 consumers from the city. ● behavious → population. 3} Sample in a medical Study ● 200 patients ● Test the effectiveness of the drug.

<u>Types of sampling Techniques</u>

1. Probability Sampling - In **probability sampling**, every member of the population has a **known and equal chance** of being selected.
   → It reduces bias and allows statistical inference about the population.

Type	Description	Example
1. Convenience Sampling	Selecting whoever is easiest to reach.	Surveying people in a nearby mall.
2. Judgmental (Purposive) Sampling	Researcher selects based on their judgment or expertise.	Selecting only "experienced" teachers for an education study.
3. Snowball Sampling	Existing participants refer others (useful for hard-to-reach groups).	Interviewing drug users who refer other users.
4. Quota Sampling	Researcher fills quotas for specific categories (like gender, age group).	Selecting 50 males and 50 females for a survey.
5. Voluntary Sampling	Participants choose to take part themselves.	Online polls or feedback forms.

1. Non Probability Sampling - In **non-probability sampling**, not every member has a known or equal chance of being selected.
   → It's quicker, cheaper, but may involve **bias**.

Type	Description	Example
1. Convenience Sampling	Selecting whoever is easiest to reach.	Surveying people in a nearby mall.
2. Judgmental (Purposive) Sampling	Researcher selects based on their judgment or expertise.	Selecting only “experienced” teachers for an education study.
3. Snowball Sampling	Existing participants refer others (useful for hard-to-reach groups).	Interviewing drug users who refer other users.
4. Quota Sampling	Researcher fills quotas for specific categories (like gender, age group).	Selecting 50 males and 50 females for a survey.
5. Voluntary Sampling	Participants choose to take part themselves.	Online polls or feedback forms.

Types of data:

<u>Qualitative Data (Categorical Data)</u>

→ *Describes qualities, categories, or attributes, not numbers.*
→ *It answers questions like “What type?” or “Which category?”*

Type	Description	Examples
a. Nominal Data	Data that names or labels categories **without order** or ranking.	Gender (Male/Female), Color (Red/Blue/Green), City names (Delhi, Mumbai)
b. Ordinal Data	Data that has **order or ranking**, but **differences between ranks are not measurable**.	Education level (High school < Graduate < Postgraduate), Rating (Poor < Good < Excellent), Class rank (1st, 2nd, 3rd)

<u>Quantitative Data (Numerical Data)</u>

→ *Represents numbers or measurable quantities.*
→ *Answers questions like "How much?" or "How many?"*

Type	Description	Examples
a. Discrete Data	Data that can take **only whole, countable values**.	Number of students (30), Cars in parking (15), Goals scored (3)
b. Continuous Data	Data that can take **any value within a range** (including decimals).	Height (165.5 cm), Weight (58.3 kg), Temperature (36.8°C), Time (2.5 hrs)

Summary table:

Main Type	**Subtype**	**Nature**	**Examples**
Qualitative	**Nominal**	Categorical, no order	Gender, Color, City
Qualitative	**Ordinal**	Categorical, ordered	Rating, Education level
Quantitative	**Discrete**	Countable, whole numbers	Students, Cars, Books
Quantitative	**Continuous**	Measurable, decimals possible	Height, Weight, Time, Temperature

Scales of Measurement of data

There are **4 main scales of measurement** (developed by S.S. Stevens):

1. **Nominal Scale**
2. **Ordinal Scale**
3. **Interval Scale**
4. **Ratio Scale**

1. Nominal Scale

Definition:

Used for **labeling or naming** variables without any order or ranking.
It simply **classifies data into categories**.

Characteristics:

- Categories are **distinct**.
- **No order** or ranking among them.
- **Numbers** used are just labels, not quantities.

Examples:

- Gender → Male / Female
- Blood group → A, B, AB, O
- Colors → Red, Green, Blue
- Marital Status → Single, Married, Divorced

2. Ordinal Scale

Definition:

Represents **ordered categories** — data can be ranked, but **differences between ranks are not measurable**.

Characteristics:

- Data has **order or ranking**.
- **Intervals between values are not equal or known.**
- Only **comparisons** (like greater or smaller) are meaningful.

Examples:

- Ratings → Poor, Good, Excellent
- Education level → High School < College < Graduate
- Customer satisfaction → 1-star to 5-star rating

3. Interval Scale

Definition:

Shows **ordered data with equal intervals** between values, but **no true zero point**.
Zero here does **not mean absence** of the quantity.

Characteristics:

- **Equal intervals** between measurements.
- **No absolute zero** (so ratios are meaningless).
- Can add and subtract, but can’t multiply or divide meaningfully.

Examples:

- Temperature (°C or °F): 0°C doesn't mean "no temperature."
- Calendar years: 2000, 2020 → difference is meaningful, but no true zero year.
- IQ scores

4. Ratio Scale

Definition:
Has all features of the interval scale **plus a true zero** point.
This allows **all arithmetic operations** (addition, subtraction, multiplication, division).

Characteristics:

- **Equal intervals** between values.
- Has **a true zero point** (means total absence).
- **All mathematical operations** are valid.
- **Ratios are meaningful** (twice, half, etc.).

Examples:

- Height, Weight, Age, Income, Distance, Time, Marks obtained