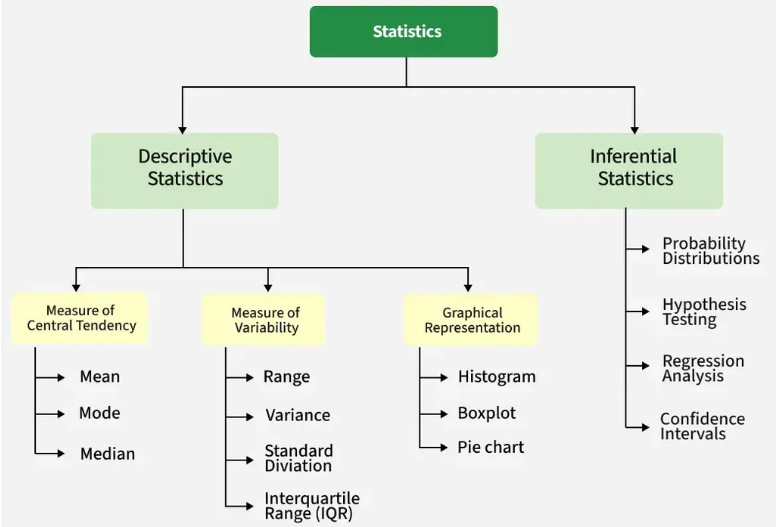
**Statistics :**

Statistics is the branch of mathematics that deals with the collection, organization, analysis, interpretation, and presentation of data. It provides methods to understand data patterns, summarize information, and make informed decisions based on data.

Use of Statistics in Data Science :

1. **Data Collection & Organization** : Helps in designing surveys, experiments, and sampling methods to gather relevant data.
2. **Data Summarization** : Uses measures like mean, median, mode, variance, and standard deviation to describe data patterns.
3. **Understanding Data Distributions** : Identifies how data is spread (e.g., normal distribution, skewness, kurtosis).
4. **Making Predictions** : Applies inferential statistics to draw conclusions from samples and predict outcomes for larger populations.
5. **Hypothesis Testing** : Tests assumptions about data to support decision-making (e.g., A/B testing).
6. **Data-Driven Decision Making** : Guides business and scientific decisions with quantitative evidence.
7. **Model Building** : Forms the foundation for machine learning algorithms by providing probability models, regression analysis, and feature evaluation.

****

**Inferential Statistics**

Definition: Inferential statistics allows us to make predictions or inferences about a population from a sample. It uses probability theory to generalize the results from a sample to the entire population.

Example:  
If we want to know the average salary of data scientists in India, we can survey 500 data scientists (sample) instead of the entire workforce (population), and then estimate the population's average salary.

**Common Methods in Inferential Statistics**

* **Confidence Intervals** – Range within which a population parameter is likely to fall.
* **Hypothesis Testing** – Procedures like t-tests, chi-square tests, ANOVA.
* **Regression Analysis** – Modeling relationships between variables.
* **Correlation Analysis** – Measuring the strength and direction of relationships.

**Role in Data Science**

* Helps make predictions from limited data.
* Guides decision-making with statistical evidence.
* Forms the basis for machine learning model validation and A/B testing.

**Population (N) and Sample (n)**

Population (N): The full set of items or individuals of interest.  
Sample (n): A subset of the population used for analysis.

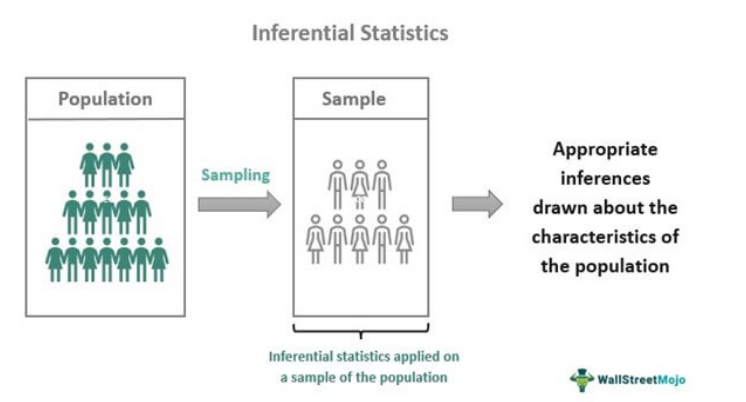
Why Sampling?

* Saves time and cost
* Makes large-scale analysis feasible

Example:  
Population: All marks of students in a university.  
Sample: Marks of 50 randomly selected students.

Table: Population vs Sample

| **Aspect** | **Population** | **Sample** |
| --- | --- | --- |
| Size | N (large) | n (small subset) |
| Data Availability | Complete | Partial |
| Purpose | True parameters | Estimate parameters |



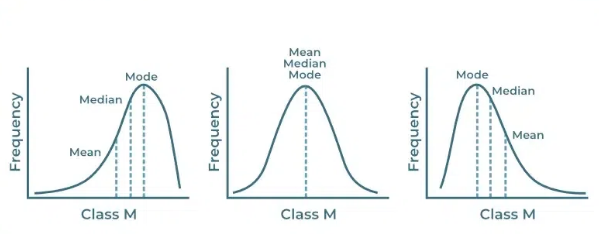
**Measures of Central Tendency**

These measure the "center" of data.

Mean:  
Formula: mean = sum(values) / n  
Example: [2, 4, 6] → Mean = 4

Median:  
Definition: The middle value in sorted data.  
Example: [3, 5, 8] → Median = 5  
If even count, average the two middle values.

Mode:  
Definition: The most frequent value.  
Example: [2, 2, 3, 4] → Mode = 2



**Measures of Dispersion**

These measure how spread out the data is.

**Variance:**  
Population variance: σ² = Σ(x - μ)² / N  
Sample variance: s² = Σ(x - x̄)² / (n - 1) [Bessel's Correction]

Why n-1 for samples?  
When calculating sample variance, we use the sample mean instead of the true population mean. This underestimates variability. Dividing by n-1 corrects this bias.

Example:  
Population: [2, 4, 6] → Variance ≈ 2.67  
Sample: [2, 4] → Variance without correction = 1.0; with correction = 2.0

**Standard Deviation:**  
Square root of variance. Example: Variance = 4 → SD = 2.

**Variables**

A variable can take different values.

Types:

1. Quantitative (Numeric)
   * Discrete: Countable (e.g., number of students)
   * Continuous: Any value in a range (e.g., height)
2. Qualitative (Categorical)
   * Nominal: No order (e.g., colors)
   * Ordinal: Ordered categories (e.g., small, medium, large)

**Random Variable**

A variable whose values are outcomes of a random event.

Types:

* Discrete: Countable values (e.g., dice roll)
* Continuous: Any value in a range (e.g., time to finish a task)

[Insert Image: Probability distribution graph]

**Histograms**

A graphical representation showing frequencies of data values grouped into bins.

Example:  
Exam scores binned into 0–10, 11–20, etc. The height of each bar = number of scores in that range.

[Insert Image: Histogram of exam scores]

1. Percentiles and Quartiles  
   Percentile: A value below which a certain % of observations fall.  
   Example: 90th percentile = score above which only 10% scored higher.

Quartiles:

* Q1: 25th percentile
* Q2: Median
* Q3: 75th percentile

**Five Number Summary**

* Minimum
* Q1
* Median
* Q3
* Maximum.

Example:  
Data: [2, 4, 6, 8, 10] → Min=2, Q1=4, Median=6, Q3=8, Max=10

Use: Often visualized using a box plot to detect distribution and outliers.