STATISTICS FOR DATA SCIENCE

What is Statistics ?

Statistics is a branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data. It involves systematic methods for gathering data, classifying it, representing it for easy interpretation, and conducting further analysis. Statistics is crucial in making informed decisions based on sample data collected through surveys or experiments.

Various fields such as psychology, sociology, economics, geology, medicine, and probability rely on statistical methods for research and practical applications.

What is Mathematical Statistics ?

Mathematical statistics focuses on the theoretical foundations and mathematical principles underlying statistical methods. It applies techniques from linear algebra, differential equations, probability theory, and mathematical analysis to derive meaningful insights from data.

Historically, statistics was associated with state affairs, used to analyze economic conditions, population data, and other governmental metrics. Today, it is widely applied across various domains, including business, healthcare, and engineering.

Mathematical statistics has two primary methods for analyzing data:

1. Descriptive Statistics
2. Inferential Statistics

Importance of Statistics

Statistics plays a vital role in various domains:

- **Business & Economics:** Market analysis, forecasting, risk assessment
- **Healthcare & Medicine:** Clinical trials, epidemiology, patient outcomes
- **Social Sciences:** Behavioral research, demographic studies
- **Engineering & Manufacturing:** Quality control, product testing
- **Artificial Intelligence & Machine Learning:** Data modeling, pattern recognition

By applying statistical methods, organizations and researchers can make data-driven decisions, optimize processes, and improve outcomes.

Descriptive Statistics	Inferential Statistics
Used to quantify the characteristics of the data.	Used to make conclusions about the population by using analytical tools on the sample data.
Measures of central tendency and measures of dispersion are important tools used.	Hypothesis testing and regression analysis are the analytical tools used.
Used to describe the characteristics of a known sample or population.	Used to make inferences about an unknown population.
Measures include variance, range, mean, median, etc.	Measures include t-test, z-test, linear regression, etc.

Descriptive Statistics

The study of numerical and graphical methods to describe and display data is called **descriptive statistics**. It helps summarize data, making it easier to understand the characteristics of a sample or population. Statisticians use graphical representations to visualize data, aiding in business trend analysis. Visual representation is often more effective than large numerical datasets.

Data can be described using three key dimensions:

1. **Measures of Central Tendency**
2. **Measures of Dispersion**
3. **Shape of the Data**

1. Measures of Central Tendency

Central tendency refers to the center of the data distribution. It describes the location where most data points are concentrated.
The three most common measures are:

1.1 Mean

The **mean** is the arithmetic average of the dataset.

- Formula: $\text{Mean} = \frac{X_1+X_2+...+X_n}{n}$

- Example:
  Data: 10,20,30,40,50
  Number of observations: 5
  $\text{Mean} = \frac{10+20+30+40+50}{5} = 30$

- Outliers significantly affect the mean.

1.2 Median

The **median** is the middle value of an ordered dataset. It divides the data into two equal halves. Unlike the mean, the median is not affected by outliers.

- **Odd number of observations:** The middle value is the median.
  Example: 10,20,30,40,50 → Median = **30**
- **Even number of observations:** The median is the average of the two middle values.
  Example: 10, 20, 30, 40, 50, 60 → Median = $\frac{30+40}{2} = 35$

1.3 Mode

The **mode** is the most frequently occurring value in a dataset. A dataset may have:

- **No mode** (if no value repeats)

- **One mode** (unimodal distribution)

- **Multiple modes** (bimodal or multimodal distribution)

- Example:
  Data: $1, 3, 4, 6, 7, 3, 3, 5, 10, 31, 3, 4, 6, 7, 3, 3, 5, 10, 3$→ Mode = **3** (appears 7 times)

- Outliers do not affect the mode.

---

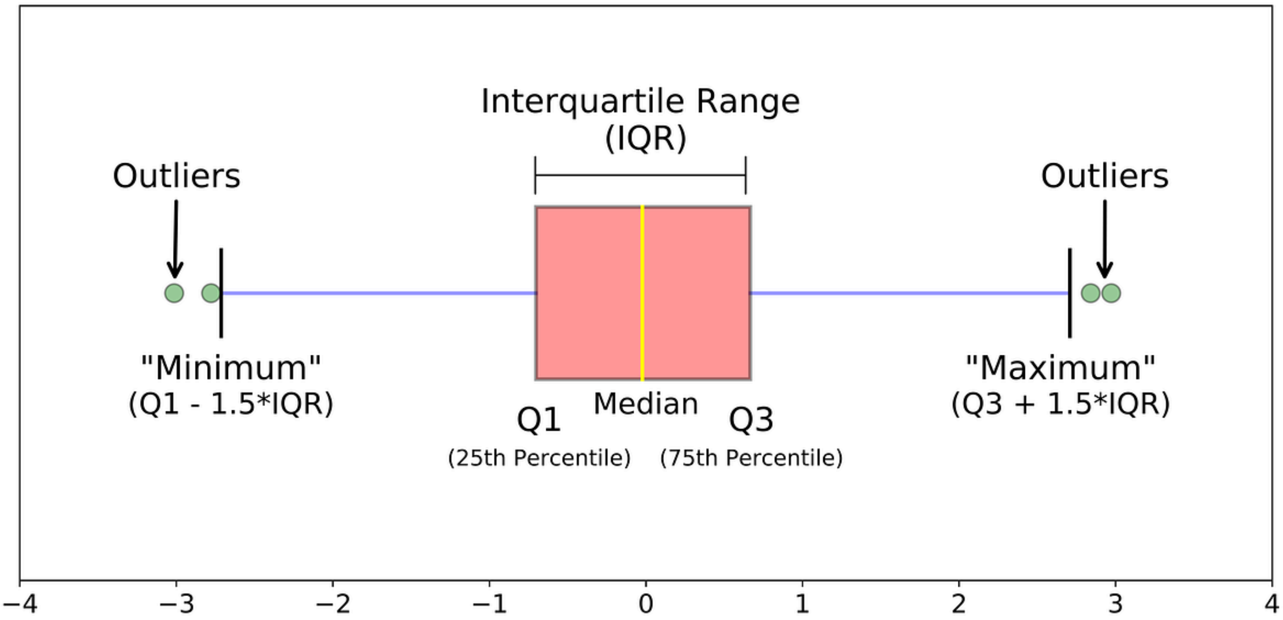
2. Measures of Dispersion

Dispersion refers to the **spread of data**. Some datasets have values concentrated near the mean, while others are widely spread. Measures of dispersion include:

2.1 Interquartile Range (IQR)

The IQR measures the spread of the middle 50% of the data. It is calculated as: $IQR = Q3 - Q1$
Where:

- **Q1 (1st quartile):** 25th percentile
- **Q2 (Median):** 50th percentile
- **Q3 (3rd quartile):** 75th percentile



2.2 Range

The **range** is the difference between the maximum and minimum values: $Range = Max - Min$

2.3 Standard Deviation (SD)

SD measures how far data points deviate from the mean. A higher SD indicates more dispersion.

- **Sample SD (s):** $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$
- **Population SD (σ):** $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$
- Where $\bar{x}$ is the sample mean, μ is the population mean, n is the sample size, and N is the population size.

2.4 Variance

Variance measures the average squared deviation from the mean.

- Population variance: σ^2
- Sample variance: s^2

---

3. Shape of the Data

The **shape** of the data distribution is crucial in probability and decision-making. It can be analyzed using:

3.1 Symmetry

A **symmetric** dataset has equal distribution on both sides. The mean and median are close together. The **normal distribution** forms a symmetric bell-shaped curve.

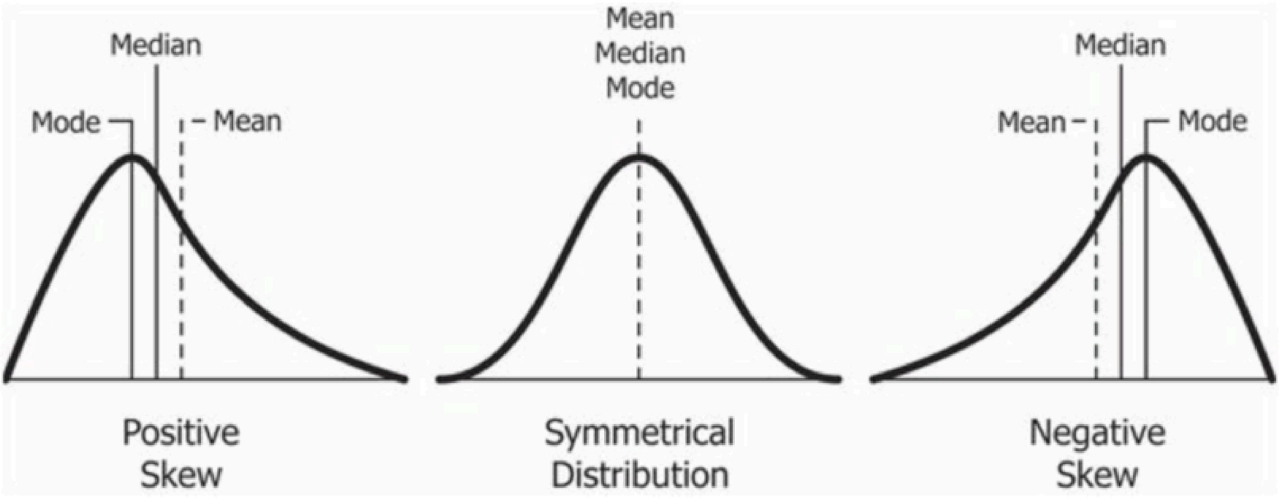
3.2 Skewness - Measure of Asymmetry

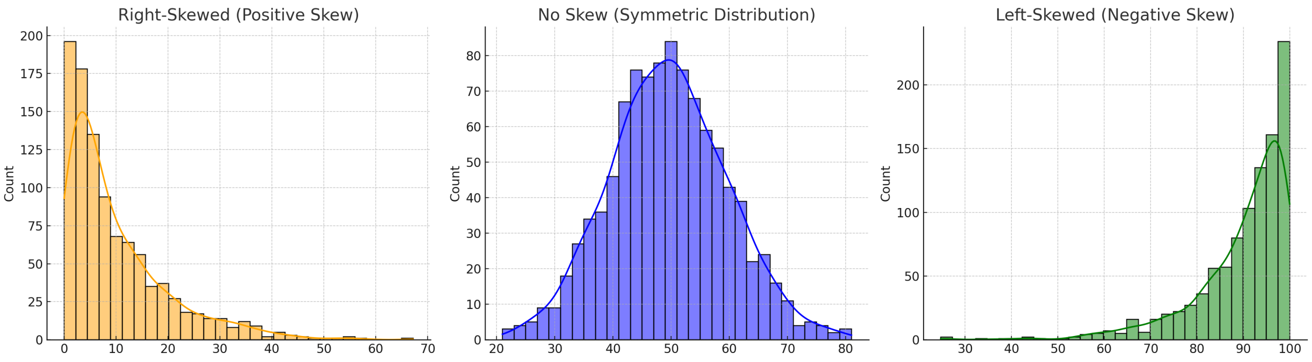
Skewness tells whether the data is symmetrical or leaning toward one side.

- **Positive Skew (Right-Skewed):** Tail is longer on the right (e.g., income data).

- **Negative Skew (Left-Skewed):** Tail is longer on the left (e.g., exam scores).

- **Zero Skew (Symmetrical):** Data is evenly distributed.





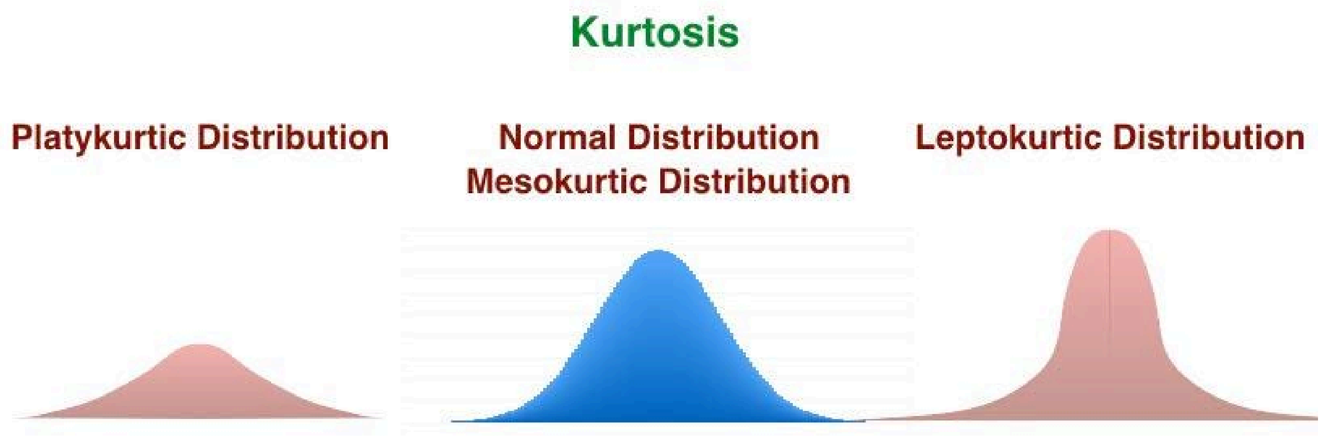
Example:

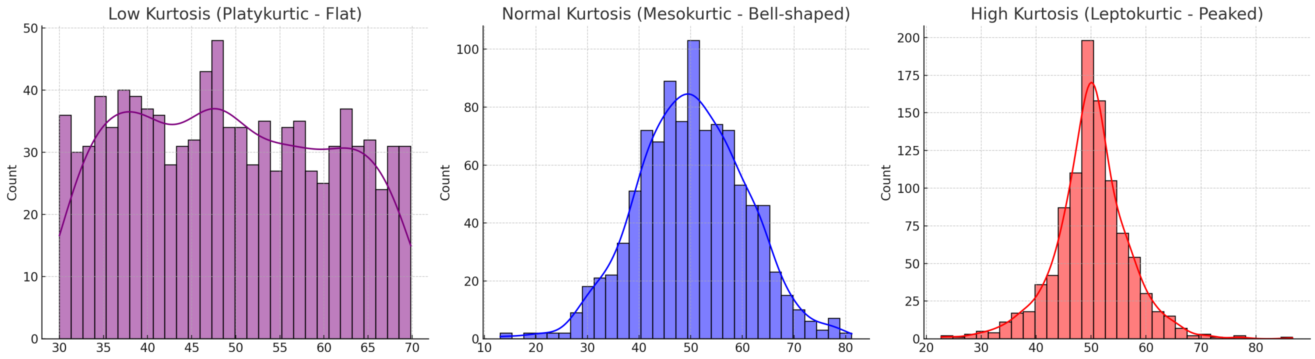
- **Right-Skewed:** 2, 3, 5, 8, 12, 20, 50 (long right tail)
- **Left-Skewed:** 50, 20, 12, 8, 5, 3, 2 (long left tail)

3.3 Kurtosis - Measure of Peakedness

Kurtosis describes how peaked or flat a distribution is.

- **High Kurtosis (Leptokurtic):** Very peaked, sharp peak (e.g., financial crashes).
- **Low Kurtosis (Platykurtic):** Flat distribution, spread out (e.g., uniform data).
- **Normal Kurtosis (Mesokurtic):** Standard bell-shaped curve.





Example:

- **Leptokurtic:** Exam scores with most students scoring around the average.
- **Platykurtic:** Heights of randomly selected individuals.