Statistics and Probability

1 Introduction to Statistics

- What is Statistics?
- Types of Statistics
  - Descriptive Statistics
  - Inferential Statistics
- Data and types of Data
- Population and Sample
- Statistical Data Analysis Steps

2 Descriptive Statistics

- Measures of Central Tendency
  - Mean
  - Median
  - Mode
- Measures of Dispersion
  - Interquartile Range(IQR)
  - Range
  - Variance
  - Standard Deviation
- Shape of the Data
  - Symmetry
  - Skewness
  - Kurtosis
- Frequency
- Graphical Representations
  - Boxplots
  - Histograms
  - Scatterplots
- Outliers and Understanding Their Impact
- Correlation and Covariance

3 Probability

- Basic Probability Concepts
  - Sample Space
  - Events
- Types of Events
  - Disjoint and Non-Disjoint Events
  - Independent and Dependent Events
- Conditional Probability
- Bayes' Theorem
- Probability Distributions
  - Random Variables and Its Types (Discrete and Continuous)
  - PMF and PDF
- Discrete Distributions
  - Bernoulli Distribution
  - Binomial Distribution
- Continuous Distributions
  - Uniform Distribution
  - Normal Distribution
    - Standard Normal Distribution
    - Standardization
    - Normalization
    - Empirical Rule

4 Inferential Statistics

- Relationship with Descriptive Statistics
- Point and Interval Estimation
- Confidence Interval (Z / T Distribution)
- Hypothesis Testing
  - Types of Hypotheses: Null and Alternate
  - Level of Significance and P-Value
  - Type I and Type II Errors
  - One-tailed and Two-tailed Tests
  - Types of Tests in Statistics
    - Z-test
    - T-test
    - ANOVA
    - Chi-square Test

Introduction to 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, classifying, representing, and analyzing data, crucial for informed decision-making based on sample data from surveys or experiments.

Applications: Psychology, Sociology, Economics, Geology, Medicine, Probability, Business, Healthcare, Engineering, Artificial Intelligence, and Machine Learning.

What is Mathematical Statistics?

Mathematical statistics focuses on the theoretical foundations and mathematical principles underlying statistical methods. Techniques from linear algebra, differential equations, probability theory, and mathematical analysis are used to derive insights from data.

Historical Note: Originally linked with state affairs (e.g., population, economics), statistics is now applied across many domains.

Two Primary Methods in Statistics

1. **Descriptive Statistics**
2. **Inferential Statistics**

Importance of Statistics

- **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
- **AI & Machine Learning:** Data modeling, pattern recognition

Descriptive Statistics vs Inferential Statistics

Descriptive Statistics	**Inferential Statistics**
Quantifies characteristics of data	Makes conclusions about the population using sample data
Measures: mean, median, variance, range, etc.	Analytical tools: t-test, z-test, regression, etc.
Describes known sample or population	Infers about unknown population

What is Data?

In data science, **data** refers to raw information collected from different sources, such as databases, sensors, websites, surveys, and more. This data is processed and analyzed using statistical and machine learning techniques to derive meaningful insights and make data-driven decisions.

Data Categories in Data Science

Data can be broadly categorized into three types:

1. **Structured Data**
   - **Definition:** Highly organized data stored in a predefined format, usually in tables or databases.
   - **Examples:** Excel sheets, CSV files, SQL databases, customer records, financial transactions.
2. **Unstructured Data**
   - **Definition:** Data that does not follow a fixed format and is often free-form.
   - **Examples:** Social media posts, images, videos, emails, medical reports, text from articles and blogs.
3. **Semi-Structured Data**
   - **Definition:** Data with some organization but not fitting traditional tabular structures.
   - **Examples:** JSON and XML files, NoSQL databases like MongoDB.

Other Classifications of Data

- **By Structure**
  - **Structured:** Organized in rows and columns (e.g., relational databases, spreadsheets).
  - **Unstructured:** Lacks a predefined model or format (e.g., emails, images, videos, web pages).
- **By Time Dimension**
  - **Cross-sectional:** Collected at a single point in time across multiple subjects (e.g., survey data, test scores on a specific date).
  - **Time Series:** Collected over time intervals (e.g., monthly sales, daily stock prices).
- **By Variables**
  - **Univariate:** Consists of a single variable.
  - **Multivariate:** Involves two or more variables.

Types of Variables

Variable Type	**Description**	**Examples**
Nominal	Categories or labels with no inherent order	Gender, Colors
Ordinal	Categories with a meaningful order, but unequal intervals	Education levels, Satisfaction ratings
Categorical	Data falling into categories (nominal or ordinal)	Product categories, Eye color
Numerical	Measurable quantities (discrete or continuous)	Temperature, Income, Age
Interval	Equal intervals between values, no true zero	IQ scores, Celsius temperature
Ratio	Equal intervals with a true zero point	Height, Weight, Income

Types of Data

1. **Qualitative (Categorical) Data**
   - Represents characteristics that cannot be measured numerically.
   - Used for classification.
     - **Nominal Data:** Unordered categories (e.g., Gender, Eye color, Nationality).
     - **Ordinal Data:** Ordered categories without equal intervals (e.g., Satisfaction ratings, Education levels).

2. **Quantitative (Numerical) Data**
    - Consists of measurable quantities.
        - **Discrete Data:** Countable, whole numbers (e.g., Number of students, Number of cars).
        - **Continuous Data:** Measurable and can take infinite values within a range (e.g., Height, Weight, Temperature).

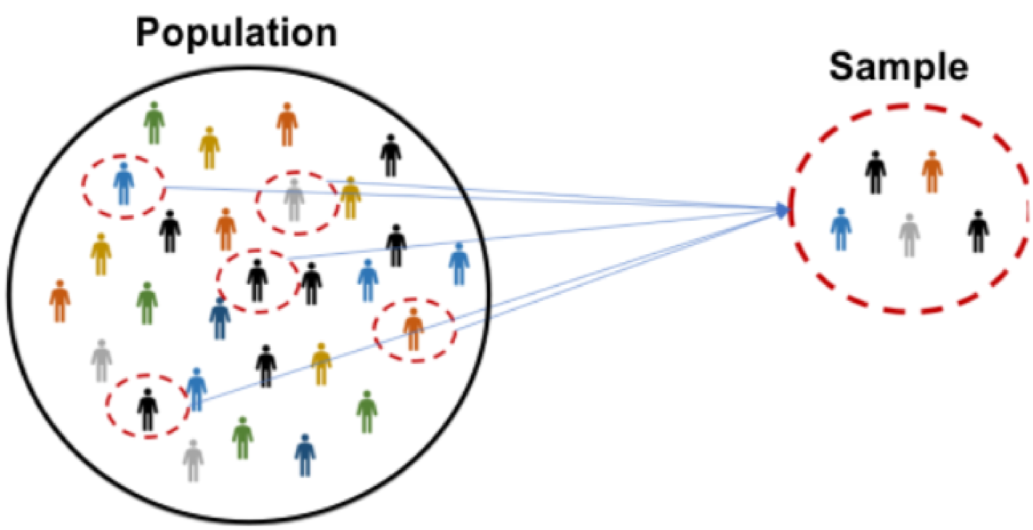
Scales of Measurement (Levels of Data)

Scale	Order	Equal Intervals	True Zero	Examples
Nominal	✗	✗	✗	Gender, Nationality
Ordinal	✓	✗	✗	Satisfaction Ratings, Education Level
Interval	✓	✓	✗	Temperature (°C, °F)
Ratio	✓	✓	✓	Height, Weight, Income

Population and Sample

Introduction

In statistics, the concepts of **population** and **sample** are fundamental. They help us analyze data and make inferences about a large group using a smaller, manageable subset.



Definition of Population

A **population** refers to the entire group of individuals, objects, or events that share a common characteristic and are the subject of a statistical study.

Examples of Populations:

- All students in a university.
- All patients suffering from diabetes in India.
- All smartphones manufactured by a company in a year.

Characteristics of a Population:

- Includes every individual that fits the criteria.
- Population size can be finite (e.g., all employees in a company) or infinite (e.g., all stars in the universe).
- Studying the entire population is often impractical due to time, cost, and feasibility constraints.

Definition of Sample

A **sample** is a subset of the population selected for analysis. The goal is to make inferences about the population based on observations from the sample.

Examples of Samples:

- 500 students randomly selected from a university.
- A group of 1,000 diabetes patients from various hospitals.
- 100 randomly selected smartphones from a production batch.

Characteristics of a Sample:

- Always smaller than the population.
- Should accurately represent the entire population.
- Selection should follow specific techniques to avoid bias.

Differences Between Population and Sample

Feature	**Population**	**Sample**
Definition	The entire group of interest	A subset of the population
Size	Usually large	Smaller than the population
Representation	Complete data	Should be representative of the population
Data Collection	More time-consuming and expensive	Less time-consuming and cost-effective
Accuracy	Higher accuracy (if measured)	Subject to sampling errors

Types of Sampling Techniques

To ensure that a sample accurately represents the population, various sampling techniques are used:

A. Probability Sampling (Random Selection)

- **Simple Random Sampling (SRS):** Every individual has an equal chance of being selected.
  Example: Randomly picking 100 students from a university.
- **Stratified Sampling:** The population is divided into subgroups (strata), and samples are drawn from each stratum.
  Example: Selecting students based on year (1st year, 2nd year, etc.).
- **Cluster Sampling:** The population is divided into clusters, and a few entire clusters are randomly selected.
  Example: Selecting entire classrooms from a school.
- **Systematic Sampling:** Every nth individual is selected from a list.
  Example: Choosing every 5th patient from a hospital record.

B. Non-Probability Sampling (Non-Random Selection)

- **Convenience Sampling:** Selecting individuals based on ease of access.
  Example: Surveying people at a shopping mall.
- **Judgmental (Purposive) Sampling:** Selection is based on expert judgment.
  Example: Choosing patients who fit a certain medical profile.
- **Quota Sampling:** Sampling is done to meet a pre-set quota.
  Example: Selecting equal numbers of male and female participants.
- **Snowball Sampling:** Existing participants recruit future participants.
  Example: Research on drug addicts where one addict introduces another.

Importance of Sampling

- Saves time and reduces costs.
- Reduces workload for researchers.
- Provides accurate insights when proper techniques are used.
- Enables research on large populations using limited data.

Statistical Data Analysis

Statistical data analysis is the systematic process of collecting, organizing, exploring, transforming, and interpreting data to discover useful information, support decision-making, and draw meaningful conclusions.

Key Steps in Statistical Data Analysis

1. **Define the Problem or Research Question**
   - Clearly articulate the objective or problem statement.
   - Define what you are trying to solve or understand.
   - This step sets the direction for the entire analysis.
   - *Examples:*
     - What factors influence student performance?
     - Does a new drug reduce symptoms better than the standard?
2. **Data Collection**
   - Gather relevant data to answer the research question.
   - **Methods:** Surveys, questionnaires, observational studies, controlled experiments, public/proprietary datasets.
   - Ensure data is reliable, valid, and representative.
3. **Data Cleaning (Preprocessing)**
   - Prepare data for analysis by:
     - Removing duplicates
     - Handling missing values (e.g., imputation)
     - Detecting and treating outliers
     - Resolving inconsistencies and data entry errors
   - Ensures data quality and improves accuracy.

4. **Exploratory Data Analysis (EDA)**
    - Gain initial insights using:
        - Descriptive statistics (mean, median, mode, standard deviation, etc.)
        - Data visualization (histograms, box plots, scatter plots, pair plots)
    - Identify patterns, relationships, and anomalies.
    - Understand data structure and distribution.
5. **Data Transformation**
    - Modify data to meet model assumptions or improve interpretability.
    - **Common transformations:** Normalization, standardization, log transformation, encoding categorical variables.
    - Helps improve model performance and interpretability.
6. **Hypothesis Formulation**
    - Construct formal hypotheses:
        - **Null Hypothesis (H_0):** No effect or relationship.
        - **Alternative Hypothesis (H_1):** There is an effect or relationship.
    - Hypothesis testing determines statistical significance.
7. **Statistical Testing**
    - Choose the right test based on data type, sample size, and assumptions.
    - **Common tests:** T-tests, Chi-square test, ANOVA, correlation, regression, Z-test, Mann-Whitney U.
    - Use statistical software (R, Python, SPSS, Excel) for calculations.
8. **Interpretation of Results**
    - Analyze outputs:
        - **P-value:** Statistical significance
        - **Confidence intervals:** Range for true value
        - **Effect size:** Practical significance
    - *Example:* If p-value < 0.05, reject the null hypothesis (statistically significant result).
9. **Draw Conclusions**
    - Accept or reject hypotheses.
    - Summarize key findings and address the research question.
    - Consider both statistical and practical significance.
10. **Documentation and Reporting**
    - Clearly document data sources, cleaning steps, tools, methods, assumptions, interpretations, and limitations.
    - Prepare a well-structured report including:
        - Introduction
        - Methodology
        - Results
        - Visuals
        - Conclusion
        - Recommendations
    - Ensure the report is reproducible and easy to follow.

Descriptive Statistics

Descriptive statistics summarize data using numerical and graphical methods, making it easier to understand characteristics of a sample or population.

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.

- **Mean (Arithmetic Average):**

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\text{Mean} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

  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.

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

  Median is not affected by outliers.

- **Mode:** The most frequently occurring value in a dataset.
  - No mode (if no value repeats)
  - One mode (unimodal)
  - Multiple modes (bimodal/multimodal)

  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.

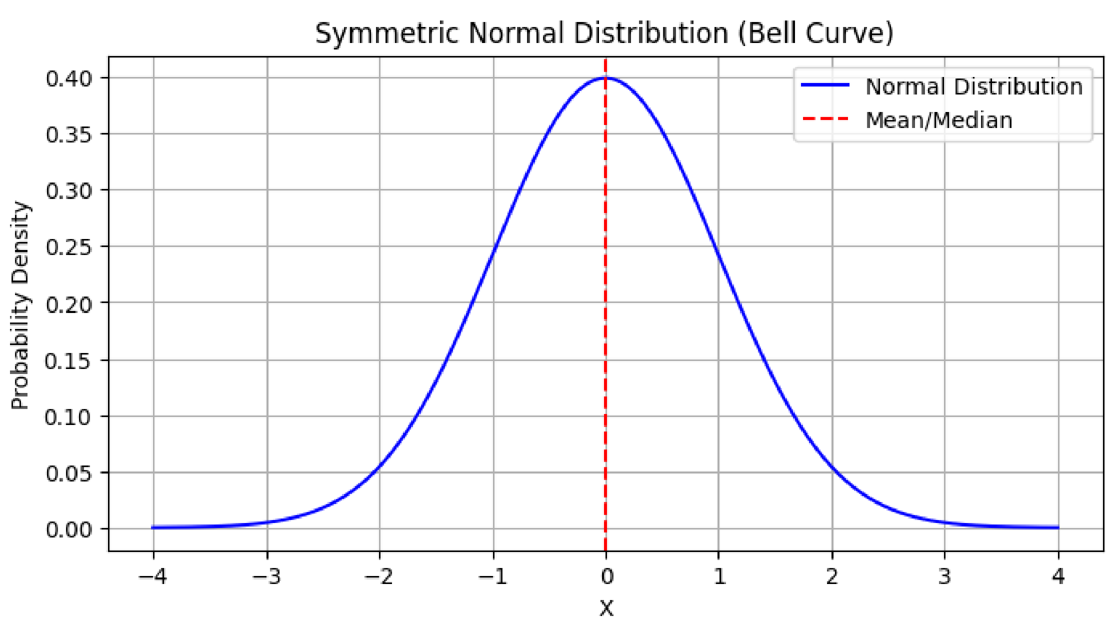
2. Measures of Dispersion

Dispersion refers to the spread of data.

- **Interquartile Range (IQR):**
$$\text{IQR} = Q_3 - Q_1$$
Where Q_1 is the 25th percentile, Q_2 is the median, and Q_3 is the 75th percentile.
- **Range:**
$$\text{Range} = \text{Max} - \text{Min}$$
- **Standard Deviation (SD):**
  - **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 sample size, N is population size.
- **Variance:**
  - **Population Variance:**
$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$
  - **Sample Variance:**
$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

3. Shape of the Data

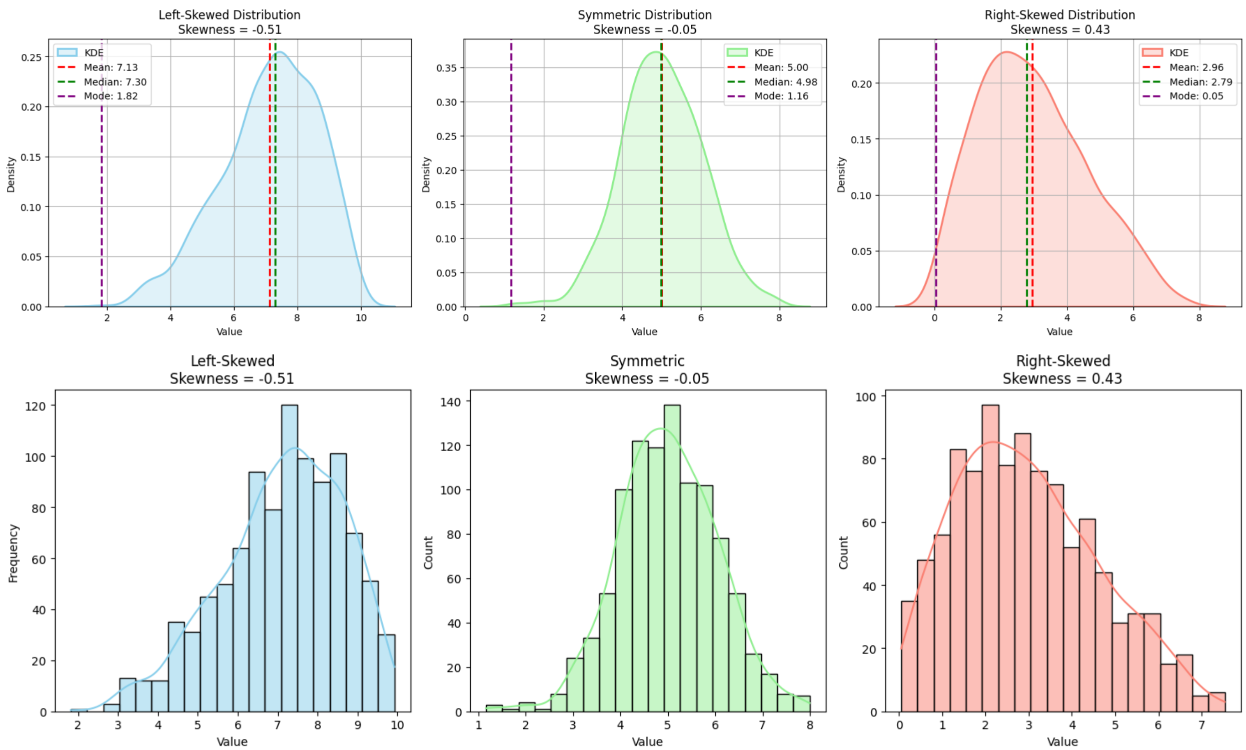
- **Symmetry:** Symmetric datasets have equal distribution on both sides. Mean and median are close. The normal distribution is symmetric (bell-shaped).



- **Skewness:** Measures asymmetry.
  - **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.

Examples:

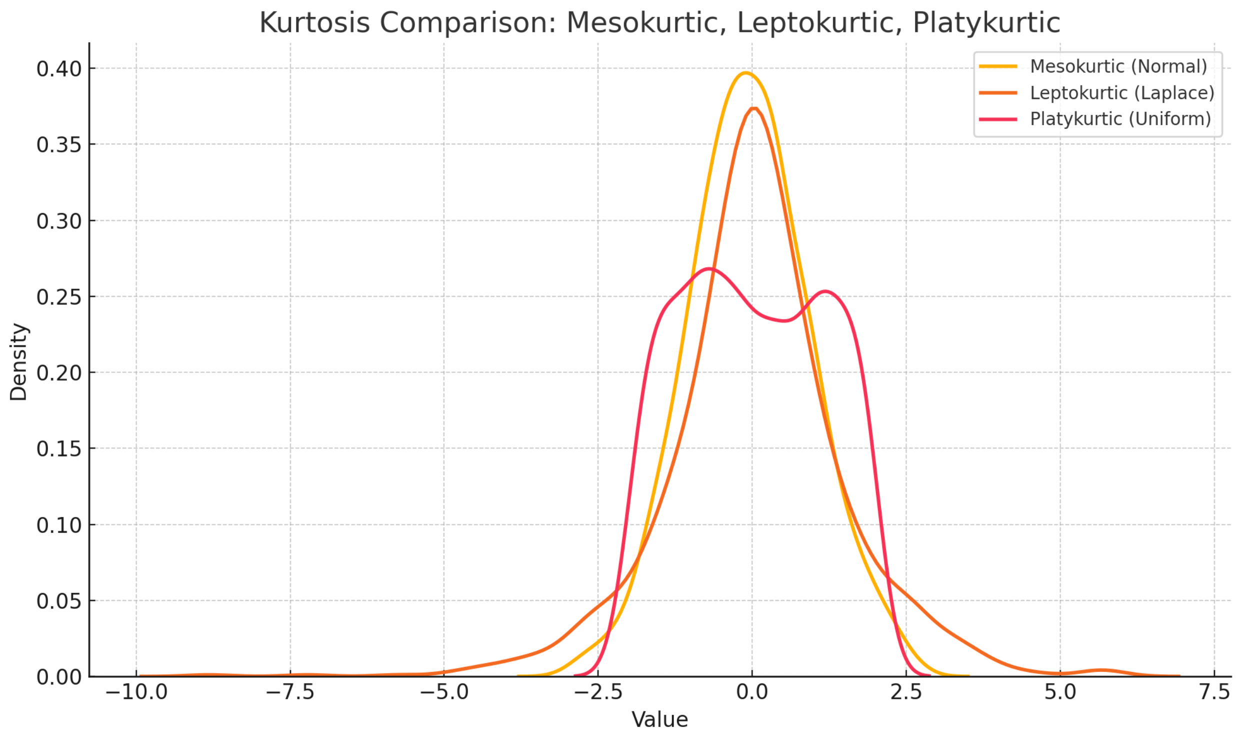
  - Right-Skewed: 2, 3, 5, 8, 12, 20, 50
  - Left-Skewed: 50, 20, 12, 8, 5, 3, 2

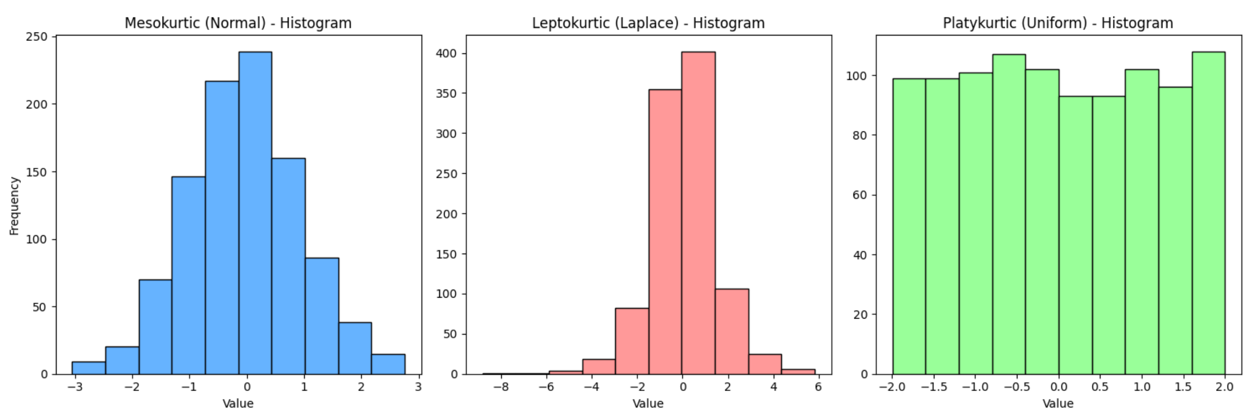


- **Kurtosis:** Measures the peakedness of a distribution.
  - **High Kurtosis (Leptokurtic):** Very peaked, sharp peak (e.g., financial crashes).
  - **Low Kurtosis (Platykurtic):** Flat, spread out (e.g., uniform data).
  - **Normal Kurtosis (Mesokurtic):** Standard bell-shaped curve.

  Examples:

  - Leptokurtic: Exam scores clustered around the average.
  - Platykurtic: Heights of randomly selected individuals.





Frequency in Statistics

What is Frequency?

Frequency is the number of times a specific value or category appears in a dataset.
Example:
Dataset: A, B, A, C, B, A, B, C, A

- Frequency of A = 4
- Frequency of B = 3
- Frequency of C = 2

Types of Frequency in Statistics

1. Absolute Frequency

Definition: The number of times a specific value or category appears in a dataset.
Example:
If a survey records the favorite fruits of 10 people and we get:

```
[Apple, Banana, Apple, Orange, Banana, Apple, Banana, Apple, Orange, Banana]
```

Fruit	Absolute Frequency
Apple	4
Banana	4
Orange	2

2. Relative Frequency

Definition: The proportion (fraction or percentage) of the total number of observations that a particular value represents.
Formula:

$$\text{Relative Frequency} = \frac{\text{Absolute Frequency}}{\text{Total Observations}}$$

Example (from above): Total responses = 10

Fruit	Absolute Freq	Relative Freq
Apple	4	$4/10 = 0.40$
Banana	4	$4/10 = 0.40$
Orange	2	$2/10 = 0.20$

3. Cumulative Frequency

Definition: The sum of the frequencies of all values up to and including the current value (typically used for ordered data like numbers or grades).
Example: Scores of 10 students:

```
[50, 60, 60, 70, 70, 70, 80, 80, 90, 100]
```

Score	Frequency	Cumulative Frequency
50	1	1
60	2	$1 + 2 = 3$
70	3	$3 + 3 = 6$
80	2	$6 + 2 = 8$
90	1	$8 + 1 = 9$
100	1	$9 + 1 = 10$

4. Cumulative Relative Frequency

Definition: The running total of relative frequencies up to a certain point. It tells you the proportion of observations less than or equal to a certain value.

Example (from above):

Score	**Frequency**	**Rel. Freq**	**Cumulative Rel. Freq**
50	1	$1/10 = 0.10$	0.10
60	2	$2/10 = 0.20$	$0.10 + 0.20 = 0.30$
70	3	$3/10 = 0.30$	$0.30 + 0.30 = 0.60$
80	2	$2/10 = 0.20$	$0.60 + 0.20 = 0.80$
90	1	$1/10 = 0.10$	$0.80 + 0.10 = 0.90$
100	1	$1/10 = 0.10$	$0.90 + 0.10 = 1.00$

Frequency Table Example

Suppose you have test scores:

Scores: 2, 3, 2, 5, 3, 4, 4, 4, 5, 5

Score (x)	Frequency (f)	Relative Frequency (f/n)	Cumulative Frequency	Cumulative Relative Frequency
2	2	0.20	2	0.20
3	2	0.20	4	0.40
4	3	0.30	7	0.70
5	3	0.30	10	1.00

- Total frequency = 10 (number of data points)
- Relative frequency shows the proportion for each score.
- Cumulative frequency and cumulative relative frequency help track totals as you move through the dataset.

Uses of Frequency

- **Descriptive Statistics:** Summarizes how data is distributed.
- **Visualization:** Forms the basis of bar charts, histograms, and frequency polygons.
- **Probability:** Relative frequency can estimate probabilities.
- **Data Preparation:** Useful for sorting, binning, and detecting outliers.

Visualizing Frequency

1. **Bar Chart:** Best for categorical data (e.g., grades, colors). X-axis: Categories, Y-axis: Frequencies.
2. **Histogram:** Best for continuous or grouped data. Bars touch to show data intervals.
3. **Frequency Polygon:** Line graph using midpoints of intervals (X-axis) vs. frequency (Y-axis).

Frequency Distribution for Grouped Data

When data is large or continuous, group it into intervals (bins):

Example:

Scores: 23, 35, 42, 47, 51, 52, 60, 67, 71, 85, 87, 90, 95

Class Interval	Frequency
20 − 29	1
30 − 39	1
40 − 49	2
50 − 59	2
60 − 69	2
70 − 79	1
80 − 89	2
90 − 99	2