PROBABILITY AND STATISTICS

Population - The entire group being studied (e.g., all students in India).

Sample - A subset of the population used for analysis.

Variable - A characteristic or property that can vary among subjects (e.g., height, age).

Data - Collected values of variables (can be qualitative or quantitative).

Types of Data

Term	Definition
Qualitative Data	Non-numerical data (e.g., gender, color).
Quantitative Data	Numerical data (e.g., income, weight).
Discrete Data	Countable, finite values (e.g., number of children).
Continuous Data	Measurable values on a scale (e.g., height, time).
Nominal Scale	Categories without any order (e.g., hair color).
Ordinal Scale	Categories with a meaningful order but no fixed interval (e.g., rankings).
Interval Scale	Ordered, equal intervals, but no true zero (e.g., temperature in Celsius).
Ratio Scale	Ordered, equal intervals with a true zero (e.g., weight, age).

Regular variable

Regular variable value doesn't change once value is assigned.

Random variable

A random variable, X, is a quantity that can have different values each time the variable is inspected,

X(value after rolling the dice)

Types of Random variable

There are two types of random variable

- Discrete random variable
- Continuous random variable

Discrete random variable

A discrete random variable is one which may take on only a countable number of distinct values.

For example, rolling dice can have values from the set {1, 2, 3, 4, 5, 6}.

Continuous random variable

A continuous random variable is one which takes an infinite number of possible values.

Predicting random variable value

Due to nature of random variable, it is impossible to predict the value.

Predicting randomness

While we might not be able to predict a specific value, it is often the case that some values might be more likely than others. We might be able to say something about how often a certain number will appear when drawing many examples.

An **Event** in probability is a **set of outcomes** (a subset of the sample space) from a random experiment.
It represents something that **can happen** when an experiment is performed.

- An event is **any collection of outcomes** from the **sample space (S)**.
- If S is the sample space, then:

Event $E \subseteq S$

Mutually Exclusive Events are events that cannot occur at the same time in a probability experiment.

If one event occurs, the other cannot happen.

Two events A and B are mutually exclusive if:

$P(A \cap B) = 0$

This means the intersection of A and B is empty.

If A and B are mutually exclusive:

$P(A \cup B) = P(A)+P(B)$ (Because $P(A \cap B)=0$)

Mutually exclusive ≠ Independent

Mutually exclusive events are dependent because knowing one occurs makes the other impossible.

Example

- Tossing a single coin:
  - Event A: “Head appears”
  - Event B: “Tail appears”
  - Both cannot happen together → **Mutually Exclusive**
- Rolling a die:
  - Event A: “Even number”
  - Event B: “Odd number”
  - Both cannot occur at the same time → **Mutually Exclusive**

Sample Space (S) is the **set of all possible outcomes** of a random experiment.

Sample Space S={all possible outcomes of an experiment}

🤔 Examples

1. **Toss a coin**

   S= {H, T}
2. **Roll a die**

S = {1,2,3,4,5,6}

Probability is a measure of how likely an event is to occur. It ranges between **0 and 1**:

$$0 \leq P(E) \leq 1$$

- **0** → Event will never happen
- **1** → Event will definitely happen

For an experiment with a **finite sample space** S and an event E:

$$P(E) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}} = \frac{n(E)}{n(S)}$$

- Toss a die:
  - Sample Space $S = \{1, 2, 3, 4, 5, 6\}$
  - Event $A = \{\text{even number}\} = \{2, 4, 6\}$

$$P(A) = \frac{3}{6} = 0.5$$

✅ Key Properties

1. $0 \leq P(E) \leq 1$
2. $P(S) = 1$ (Entire sample space)
3. For **complementary event** A':

$$P(A') = 1 - P(A)$$

4. For **mutually exclusive events** A, B:

$$P(A \cup B) = P(A) + P(B)$$

5. For **general case**:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Conditional Probability is the probability of an event **A** occurring **given that** another event **B** has already occurred.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Where:

- P(A∩B) = Probability that **A and B occur together**
- P(B) = Probability that event B occurs

🧠 Example 1: Coin Toss

Two coins are tossed.

- Sample Space $S = \{HH, HT, TH, TT\}$
- Event A: "At least one Head" = {HH, HT, TH}
- Event B: "First coin is Head" = {HH, HT}

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/4}{2/4} = 1$$

- A person is known to be a student (Event B).
- Probability that this student studies engineering (Event A) given they are a student:

$$P(\text{Engineering}|\text{Student}) = \frac{\text{Number of engineering students}}{\text{Total students}}$$

Bayes Theorem

Bayes' Theorem is a mathematical formula for determining conditional probability. Conditional probability is the likelihood of an outcome occurring based on a previous outcome in similar circumstances. Thus, Bayes' Theorem provides a way to revise or update an existing prediction or theory given new evidence.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Where:

- $P(A|B)$ = **Posterior probability** (probability of A given B)
- $P(B|A)$ = **Likelihood** (probability of B given A)
- $P(A)$ = **Prior probability** of A
- $P(B)$ = **Marginal probability** of B

If there are multiple events $A_1, A_2, \ldots, A_n$:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{k=1}^{n} P(B|A_k) \cdot P(A_k)}$$

Problem:

A company manufactures chips in 3 factories:

- Factory 1: Produces **50%** of chips, defect rate = **2%**
- Factory 2: Produces **30%** of chips, defect rate = **3%**
- Factory 3: Produces **20%** of chips, defect rate = **5%**

If a chip is selected at random and is found to be defective, **what is the probability it came from Factory 3?**

Events:

- A_1: Factory 1
- A_2: Factory 2
- A_3: Factory 3
- B: Chip is defective

$$P(A_1) = 0.5,\ P(A_2) = 0.3,\ P(A_3) = 0.2$$

$$P(B|A_1) = 0.02,\ P(B|A_2) = 0.03,\ P(B|A_3) = 0.05$$

Compute $P(B)$:

$$P(B) = (0.02)(0.5) + (0.03)(0.3) + (0.05)(0.2) = 0.01 + 0.009 + 0.01 = 0.029$$

Now apply Bayes' theorem:

$$P(A_3|B) = \frac{P(B|A_3) \cdot P(A_3)}{P(B)} = \frac{0.05 \cdot 0.2}{0.029} = 0.3448 \approx 34.5\%$$

Medical Test Problem

A certain disease affects **1% of the population**. A test for the disease is:

- **95% accurate** for people with the disease (True Positive Rate).
- **90% accurate** for people without the disease (True Negative Rate).

If a person tests **positive**, what is the probability that they actually **have the disease**?

Step 1: Define Events

- A: Person **has the disease**
- A': Person **does not have the disease**
- B: Test result is **positive**

Given:

$$P(A) = 0.01,\ P(A') = 0.99$$

$$P(B|A) = 0.95 \text{ (True Positive Rate)}$$

$$P(B|A') = 0.10 \text{ (False Positive Rate)}$$

Step 2: Compute $P(B)$

$$\begin{aligned} P(B) &= P(B|A)P(A) + P(B|A')P(A') \\ &= (0.95)(0.01) + (0.10)(0.99) \\ &= 0.0095 + 0.099 = 0.1085 \end{aligned}$$

Step 3: Apply Bayes' Theorem

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{0.95 \times 0.01}{0.1085} \\ &= \frac{0.0095}{0.1085} \approx 0.0876 \end{aligned}$$

Spam Email Detection

An email filtering system classifies emails as **spam** or **not spam**.

- 80% of all emails are **not spam**, 20% are **spam**.
- The filter correctly identifies spam **90%** of the time
  (P(Filter says spam | Spam)=0.9

- The filter incorrectly marks **1% of non-spam emails as spam** (P(Filter says spam | Not spam)=0.01

Question: If the filter says an email is spam, what is the probability that it is actually spam?

Step 1: Define Events

- S: Email is spam → $P(S) = 0.2$
- N: Email is not spam → $P(N) = 0.8$
- F: Filter says "spam"

Given:

$$P(F|S) = 0.9,\ P(F|N) = 0.01$$

Step 2: Compute $P(F)$

$$\begin{aligned} P(F) &= P(F|S)P(S) + P(F|N)P(N) \\ &= (0.9)(0.2) + (0.01)(0.8) \\ &= 0.18 + 0.008 = 0.188 \end{aligned}$$

Step 3: Apply Bayes' Theorem

$$\begin{aligned} P(S|F) &= \frac{P(F|S)P(S)}{P(F)} \\ &= \frac{(0.9)(0.2)}{0.188} \\ &= \frac{0.18}{0.188} \approx 0.957 \end{aligned}$$

About 95.7% chance the email is actually spam **if the filter says spam.**

Joint Probability

Joint Probability is the probability that **two or more events happen at the same time** (intersection of events).

For two events A and B:

P(A∩B)=Probability that both A and B occur

✅ Formula

- For **independent events**:

$$P(A \cap B) = P(A) \times P(B)$$

- For **dependent events**:

$$P(A \cap B) = P(A) \times P(B|A) = P(B) \times P(A|B)$$

✅ Example 1 (Independent Events)

A coin is tossed and a die is rolled:

- Event A: Getting **Head** → $P(A) = \frac{1}{2}$
- Event B: Getting **4** on die → $P(B) = \frac{1}{6}$

Since they are independent:

$$P(A \cap B) = P(A) \times P(B) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$$

✅ Example 2 (Dependent Events)

A bag has 3 red and 2 blue balls. Two balls drawn without replacement:

- Event A: First ball is red → $P(A) = \frac{3}{5}$
- Event B: Second ball is red given first is red → $P(B|A) = \frac{2}{4}$

So:

$$P(A \cap B) = \frac{3}{5} \times \frac{2}{4} = \frac{6}{20} = 0.3$$

Probability Density (or Distribution)

Probability distributions describe how probabilities are distributed over values of a random variable. They are mainly divided into Discrete and Continuous types.

1. Discrete Probability Distributions

Applicable when the random variable takes **countable values**.

Distribution	Key Points	Example
Bernoulli	O0058ne trial, two outcomes (success/failure).	Tossing a coin (Head/ Tail)
Binomial	Number of successes in **n independent trials** with probability pp.	Number of heads in 10 coin tosses
Poisson	Number of events in a fixed time/space with constant rate.	Number of emails per hour
Geometric	Trials until first success.	First success in coin toss
Negative Binomial	Trials until **k successes**.	5th success in basketball shots

Hypergeometric	Sampling without replacement.	Picking red balls from a box

2.Continuous Probability Distributions

Applicable when the random variable takes **uncountably infinite values**.

Distribution	**Key Points**	**Example**
Uniform	Equal probability over an interval.	Random time between 0 and 10 mins
Normal (Gaussian)	Bell-shaped, defined by mean μ\mu and SD σ\sigma.	Heights, IQ scores
Exponential	Time between events in a Poisson process.	Time until next call
Gamma	Generalization of exponential.	Life span of devices
Beta	Used in Bayesian statistics (0 to 1).	Probability modeling
Chi-square	Based on variance of samples.	Hypothesis testing
Student's t	For small sample mean tests.	Confidence intervals

✅ Key Differences

Feature	Discrete	Continuous
Values	Countable	Infinite
Function	PMF (Probability Mass Function)	PDF (Probability Density Function)
Example	Number of heads	Height of students

Hypothesis

A **proposed statement or assumption** about a population parameter (e.g., mean, proportion) that we test using sample data.

- **Null Hypothesis (H_0)**: No effect or no difference (status quo).
- **Alternative Hypothesis (H_a)**: Represents a change, difference, or effect.

Example:
H0:μ=50 vs Ha:μ≠50

Bias

A **systematic error** that causes an estimate to deviate from the true value.
Types of bias:

Parameter

A **numerical characteristic of a population** (e.g., population mean, population proportion).

Statistic

A **numerical characteristic of a sample** (e.g., sample mean $\bar{x}$, sample proportion).

Significance Level (α)

The **probability of rejecting the null hypothesis when it is true** (Type I error). Common values: **0.05, 0.01**.

p-value

The **probability of obtaining a result as extreme or more extreme than the observed**, assuming H_0 is true.

- If **p-value < α**, reject H_0.

Confidence Interval

A range of values that likely contains the population parameter with a certain confidence (e.g., 95%).

Errors in Hypothesis Testing

- **Type I Error (α):** Reject H_0 when it is true.
- **Type II Error (β):** Fail to reject H_0 when it is false.

Power of a Test

The probability of correctly rejecting a false null hypothesis (1 − β).

Hypothesis Testing

Hypothesis Testing is a **statistical method** used to make decisions or inferences about a **population parameter** using **sample data**.
It determines whether there is enough evidence to **accept or reject** a claim (hypothesis).

- **Null Hypothesis** (H_0): The default assumption (e.g., "No difference," "No effect").
  Example: $H_0 : \mu = 50$
- **Alternative Hypothesis** (H_a **or** H_1): The opposite claim (e.g., "There is a difference").
  Example: $H_a : \mu \neq 50$
- **Test Statistic**: A value computed from sample data to decide whether to reject H_0.
- **Significance Level** (α): Probability of making a **Type I error** (rejecting H_0 when it's true). Common values: 0.05, 0.01.
- **p-value**: Probability of observing a test statistic as extreme as the one observed, assuming H_0 is true.
- **Type I Error** (α): Rejecting a true H_0.
- **Type II Error** (β): Failing to reject a false H_0.

Steps in Hypothesis Testing

1. **State Hypotheses**
   - Null Hypothesis (H_0)
   - Alternative Hypothesis (H_a) → One-tailed or Two-tailed
2. **Set Significance Level (α)**
   - Common: 0.05 or 0.01
3. **Choose the Right Test**
   - Z-test, t-test, Chi-square, ANOVA (depends on data type, sample size, variance known/unknown)
4. **Compute Test Statistic**

$$\text{Example: For Z-test: } Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

5. **Find p-value or Critical Value**
   - Compare p-value with α, or check if test statistic falls in rejection region.
6. **Make Decision**
   - If p-value < α → Reject H_0
   - If p-value ≥ α → Fail to reject H_0

4. Types of Tests

✅ Based on Tail:

- **One-tailed test**: Directional hypothesis (e.g., $H_a : \mu > 50$)
- **Two-tailed test**: Non-directional (e.g., $H_a : \mu \neq 50$)

✅ Common Tests:

- **Z-Test**: Population variance known, n > 30
- **t-Test**: Population variance unknown, n < 30
- **Chi-Square Test**: For categorical data (goodness of fit, independence)
- **ANOVA**: Compare more than 2 means

Claim: A factory claims the average weight of a product = 500g.
Sample: n = 36, $\bar{x} = 490g$, σ = 20g, α = 0.05
Is the claim valid?

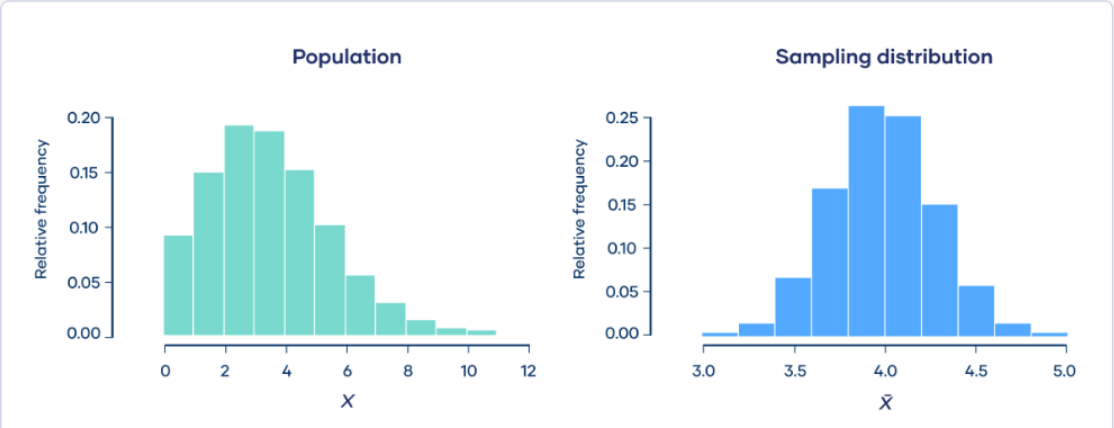
Solution:

- $H_0 : \mu = 500$, $H_a : \mu \neq 500$
- Z = $(490 - 500)/(20/\sqrt{36}) = (-10)/(20/6) = (-10)/3.33 = -3.0$
- Critical Z for α = 0.05 (two-tailed) = ±1.96
  Since |Z| > 1.96 → Reject H_0.
  Conclusion: Average weight is significantly different from 500g.

Central Limit Theorem (CLT)

The **central limit theorem** states that if you take sufficiently large samples from a population, the samples' means will be normally distributed, even if the population isn't normally distributed.

A **population** follows a **Poisson distribution** (left image). If we take 10,000 **samples** from the population, each with a sample size of 50, the sample means follow a normal distribution, as predicted by the **central limit theorem** (right image).



✅ Mathematical Statement

If:

- Population mean = μ
- Population standard deviation = σ
- Sample size = n

Then for large n:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

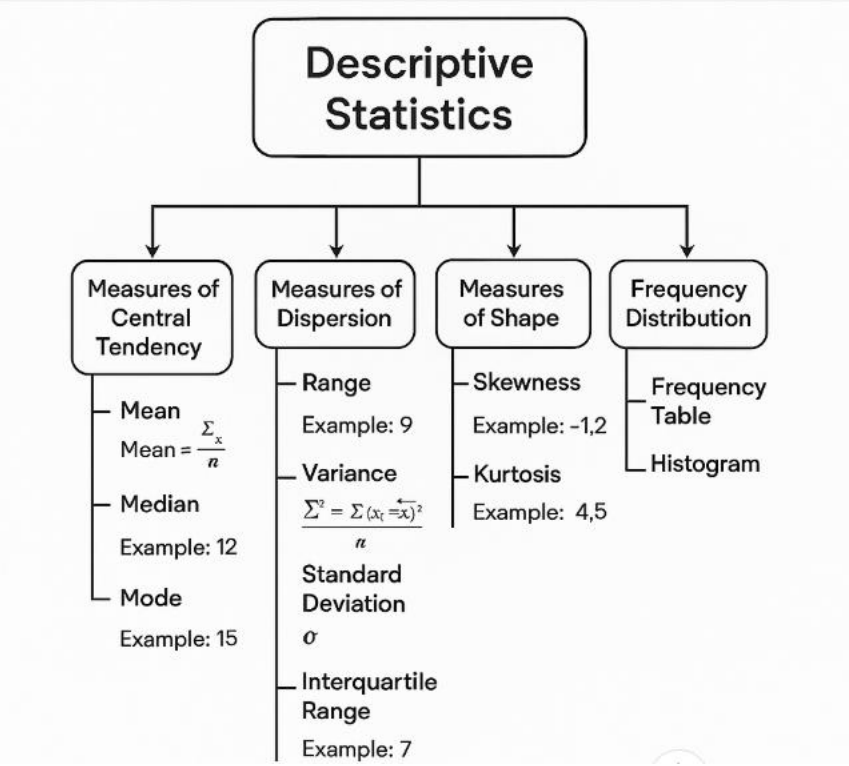
Where:

- Mean of sampling distribution = $\mu_{\bar{X}} = \mu$
- Standard error = $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

Descriptive Statistics

Descriptive statistics summarize and organize characteristics of a data set. A data set is a collection of responses or observations from a sample or entire population.

Types of descriptive statistics



🔹 1. Measures of Central Tendency

These describe the **center or average** of the data:

- **Mean**: Arithmetic average

$$\text{Mean} = \frac{\sum x_i}{n}$$

- **Median**: Middle value when data is sorted.
- **Mode**: Most frequently occurring value.

✅ 1. Measures of Central Tendency

- **Mean**:
  Example: Heights of 5 people = {150, 160, 170, 180, 190}

$$\text{Mean} = \frac{150 + 160 + 170 + 180 + 190}{5} = 170$$

- **Median**:
  Sorted data = {150, 160, 170, 180, 190} → Median = 170
- **Mode**:
  Example: {2, 3, 3, 4, 5} → Mode = 3

🔹 2. Measures of Dispersion (Variability)

These indicate **how spread out** the data is:

- **Range** = Maximum − Minimum
- **Variance** (σ^2):

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

- **Standard Deviation** (σ) = Square root of variance
- **Interquartile Range (IQR)**: Q3 − Q1
- **Coefficient of Variation (CV)** = $\frac{\sigma}{\bar{x}} \times 100\%$

✅ 2. Measures of Dispersion

- **Range**:
  Data = {5, 8, 12, 20} → Range = 20 − 5 = 15
- **Variance**:
  For {2, 4, 6}, Mean = 4

$$\sigma^2 = \frac{(2-4)^2 + (4-4)^2 + (6-4)^2}{3} = \frac{4+0+4}{3} = 2.67$$

- **Standard Deviation**:
  $\sigma = \sqrt{2.67} \approx 1.63$
- **Interquartile Range (IQR)**:
  Data = {1, 3, 5, 7, 9, 11, 13, 15}
  Q1 = 4, Q3 = 12 → IQR = 12 − 4 = 8

🔹 3. Measures of Shape

These describe the **distribution shape**:

- **Skewness**: Measure of symmetry
  - Positive skew (right tail longer)
  - Negative skew (left tail longer)
- **Kurtosis**: Measure of peakness (flat or peaked distribution)

✅ 3. Measures of Shape

- **Skewness**:
  Data {2, 3, 4, 50} → Right skewed (positive skew)
- **Kurtosis**:
  Normal distribution kurtosis ≈ 3
  Leptokurtic (high peak) > 3, Platykurtic (flat) < 3

🔹 4. Frequency Distribution

Organizes data into **classes or intervals** and shows how many values fall into each category:

- Frequency tables
- Relative frequency
- Cumulative frequency

✅ 4. Frequency Distribution

- **Example**:
  Student marks grouped in intervals:

Marks	Frequency
0–10	2
10–20	5
20–30	8