Introduction to Probability in Data Science

1. What is Probability?

Probability is a measure of the likelihood that a particular event will occur. It ranges from 0 (impossible) to 1 (certain). Probability is used to model uncertainty with data and make informed decisions based on that uncertainty.

Examples:

- Tossing a coin
- Rolling a die
- Predicting customer churn

---

2. Role of Probability in Data Science

Probability forms the foundation of statistical inference, machine learning, and decision-making under uncertainty. Data scientists use it to:

- Estimate unknown values from sample data
- Evaluate risks and uncertainties
- Make predictions using probabilistic models

---

3. Types of Probability

Theoretical Probability

Based on known possible outcomes.

- Example: Probability of getting heads in a fair coin = 0.5

Empirical (Experimental) Probability

Based on observed data.

- Example: Probability that a user clicks an ad based on 10,000 past impressions.

---

4. Key Concepts

- **Sample Space (S):** All possible outcomes of an experiment.
- **Event (E):** A subset of the sample space; one or more outcomes.

Example:

For rolling a die:

- Sample space: {1, 2, 3, 4, 5, 6}
- Event (even number): {2, 4, 6}

---

5. Real-World Business Scenarios

- Predicting customer conversion rate from a marketing campaign
- Estimating product failure rate in manufacturing
- Calculating insurance risks
- Detecting fraud in financial transactions

Calculating Probability

Formula:

P(E) = Number of favorable outcomes / Total number of outcomes

Where:

- P(E) is the probability of event E
- Number of favorable outcomes is the count of outcomes that satisfy event E
- Total number of outcomes is the count of all possible outcomes in the sample space

Example:

If you roll a die, the probability of rolling a 3 is:

P(rolling a 3) = Number of favorable outcomes (1) / Total number of outcomes (6) = 1/6 ≈ 0.1667

Quick Quiz

1. What is the probability of rolling a 5 on a fair six-sided die?
2. Two fair six-sided dice are rolled. What is the probability that at least one of the dice shows a 4?

Homework / Practice

1. List three real-world problems in Data Science where probability is involved.
2. Identify the sample space and possible events for each.
3. Try calculating a simple empirical probability using past data (if available).