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).

Types of Experiments

1. **Deterministic** – Always produces the same outcome when repeated under identical conditions.
2. **Probabilistic / Random** – Outcome is uncertain and may vary even if the experiment is repeated under identical conditions.

Sample Space

The sample space is the set of all possible outcomes of a probabilistic experiment.

Examples:
- **Rolling a Die** → { 1, 2, 3, 4, 5, 6 }
- **Tossing 2 Coins** → { (HH), (HT), (TH), (TT) }

For Coins
- Number of possible outcomes = $\mathbf{2^n}$
(where **n** is the number of coins)

For Dice
- Number of possible outcomes = $\mathbf{6^n}$
(where **n** is the number of dice)

For Cards
A standard deck contains **52 cards**:

- **26 Red Cards**
  - 13 ♥ Hearts → A, 2–10, J, Q, K
  - 13 ♦ Diamonds → A, 2–10, J, Q, K
- **26 Black Cards**
  - 13 ♠ Spades → A, 2–10, J, Q, K
  - 13 ♣ Clubs → A, 2–10, J, Q, K

Conclusion

Understanding the basic types of experiments and the concept of sample space is essential for solving probability problems. Knowing the number of possible outcomes for coins, dice, and cards helps in calculating probabilities accurately. Mastery of these fundamentals forms the foundation for more advanced topics in probability and statistics.

Questions

1. You roll a fair six-sided die. What is the probability of getting an even number?

---

2. Two coins are tossed. Find the sample space associated with this random experiment.

---

3. Two six-sided dice are rolled. What is the probability that the sum is 7?

---

4. A bag contains 3 red, 2 blue, and 5 green balls. What is the probability of randomly selecting a red ball?

---

5. A card is drawn at random from a deck. Find the probability of:
- Getting a heart
- Getting a face card

Basic Rules of Probability

1. Important Definitions

In probability, understanding the basic definitions is crucial for grasping more complex concepts.

- **Experiment**: An action or process that leads to one or more outcomes (e.g., rolling a die). Deterministic experiments have predictable outcomes, while probabilistic or random experiments have uncertain outcomes.
- **Outcome:** A possible result of an experiment (e.g., rolling a 3 on a die).
- **Probability (P):** A measure of the likelihood that an event will occur, ranging from 0 (impossible) to 1 (certain).
- **P(AUB):** The probability of event A or event B occurring.
- **P(A∩B):** The probability of both events A and B occurring.
- **Sample Space (S):** The set of all possible outcomes of an experiment.
- **Event (E):** A subset of the sample space.
- **Mutually Exclusive Events:** Two events that cannot occur at the same time (e.g., rolling a 2 and rolling a 5 on a die). $P(A \cap B) = 0$
- **Independent Events:** Two events where the occurrence of one does not affect the other. $P(A \cap B) = P(A) \cdot P(B)$
- **Certain Event:** An event that is guaranteed to happen, with a probability of 1.
- **Impossible Event:** An event that cannot happen, with a probability of 0.
- **Exhaustive Events:** A set of events that cover the entire sample space, meaning at least one of them must occur.

Example:

Tossing a die:

- Sample space: S = {1, 2, 3, 4, 5, 6}
- Event (E): Rolling an even number = {2, 4, 6}

---

2. The Complement Rule

The complement of an event A is the event that A does not occur.

- Notation: A^c
- Rule: $P(A^c) = 1 - P(A)$

Example:

If the probability of rain today is 0.3, the probability it won't rain is: - P(No Rain) = 1 − 0.3 = 0.7

---

3. The Addition Rule

Used to calculate the probability of the union of two events.

For general events:

P(A or B) = P(A) + P(B) − P(A and B)

For mutually exclusive events:

P(A or B) = P(A) + P(B)

Example:

- P(A) = 0.4, P(B) = 0.5, P(A and B) = 0.2
- P(A or B) = 0.4 + 0.5 − 0.2 = 0.7

4. The Multiplication Rule

Used to find the probability that two events occur together.

For Independent events:

P(A and B) = P(A) × P(B)

Example: A event A is rolling a 3 on a first throw of a die, and event B is rolling an even number on second throw.

- P(A) = 1/6 (rolling a 3)
- P(B) = 3/6 (rolling a 2, 4, or 6)
- P(A and B) = P(A) × P(B) = (1/6) × (3/6) = 1/12

For Dependent events:

P(A and B) = P(A) × P(B | A)

Example: A bag contains 5 red and 3 blue balls. Two balls are drawn **without replacement**. Let:

- Event A be drawing a red ball first.
- Event B be drawing a red ball second.
- P(A) = 5/8 (5 red balls out of 8)
- After removing one red ball, there are 4 red left out of 7 total: P(B/A) = 4/7
- P(A and B) = (5/8) X (4/7)= 20/56 = 5/14

So the probability of drawing two red balls **without replacement** is 5/14

5. Independent vs Dependent Events - More examples

- **Independent:** The outcome of one event does not affect the other.
    - Example: Tossing two coins.
- **Dependent:** One event affects the probability of the other.
    - Example: Drawing two cards without replacement.

---

6. Mutually Exclusive Events

Two events are **mutually exclusive** if they cannot happen at the same time.

Example:

- Event A: Rolling a 2
- Event B: Rolling a 5
- These are mutually exclusive because a die can't show both at once.

Summary

This lesson covered key probability rules and concepts that are foundational for more advanced topics like distributions and statistical inference.

- Complement Rule: $P(A^c) = 1 - P(A)$
- Addition Rule for unions
- Multiplication Rule for intersections
- Understanding independence and exclusivity

Homework / Practice

1. A card is drawn from a standard deck. What is the probability of drawing a red card or a queen?
2. If two dice are rolled, what is the probability that both show even numbers?
3. Think of an example from your daily life where two events are dependent.

Practice Questions

1. A box contains 4 red, 3 green, and 2 blue balls. One ball is picked at random.
   What is the probability that the ball is **not green**?

2. Two coins are tossed. What is the probability of getting **at least one head**?

3. A number is chosen at random from the set {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}.
   What is the probability that the number is a **multiple of 3 or 5**?

4. A bag contains 6 white and 4 black balls. Two balls are drawn **with replacement**.
   What is the probability that **both are white**?

5. A student knows the answer to 70% of the questions on a test.
   If one question is selected at random, what is the probability that the student **doesn't know** the answer?

Conditional Probability - An Intuitive Introduction

What is Conditional Probability?

Conditional probability is the probability of an event **A** occurring, given that another event **B** has already occurred. It helps us update our understanding based on new information.

We denote it as:

P(A/B) = P(A and B) / P(B)

This is read as: "The probability of A given B."

Law of Total Probability

The **Law of Total Probability** helps you find the probability of an event `A` by breaking it down based on other known events `B1`, `B2`, ..., `Bn`.

Formula:

P(A) = P(A/B1) * P(B1) + P(A/B2) * P(B2) + ... + P(A/Bn) * P(Bn)

Conditions to Use This Law:

1. **Partition of the Sample Space**
   The events `B1`, `B2`, ..., `Bn` must:
    1. Be **mutually exclusive** (no overlap):
       No two events can happen at the same time.

2. Be **collectively exhaustive**:
   Together, they cover the entire sample space.

2. **Known Probabilities**
   You must know:
   1. P(Bi) for each event in the partition.
   2. P(A/Bi), the conditional probability of A given Bi.
3. **P(Bi) > 0**
   The probability of each Bi must be greater than 0, since you can't condition on an impossible event.

Real-life Analogy

Suppose you have a box containing:

- 5 red balls
- 3 blue balls
- 2 green balls

Now, you're told that the ball picked is **not green**. What is the probability that it's red?

Let:

- A = ball is red
- B = ball is not green

Now, out of the 8 balls that are not green (5 red + 3 blue), 5 are red.

So: **P(Red / Not Green) = 5 / 8**

Formula Breakdown

If we know:

- P(A and B): The probability that both A and B happen
- P(B): The probability that B happens

Then: **P(A / B) = P(A and B) / P(B)**

This means: among all the outcomes where B occurs, how many also include A?

Example Problem

A class has 60% boys and 40% girls. 30% of the boys and 10% of the girls play football. If a randomly selected student is a football player, what is the probability that the student is a boy?

Let:

- B = student is a boy → P(B) = 0.6
- G = student is a girl → P(G) = 0.4
- F = student plays football
    - P(F / B) = 0.3
    - P(F / G) = 0.1

Now calculate the total probability that a student plays football:

P(F) = P(F / B) * P(B) + P(F / G) * P(G)

P(F) = 0.3 * 0.6 + 0.1 * 0.4 = 0.18 + 0.04 = 0.22

Now use conditional probability to find P(B / F):

P(B / F) = (P(F / B) * P(B)) / P(F)

P(B / F) = 0.18 / 0.22 ≈ 0.818

So, there's about an 81.8% chance that a football player is a boy.

Common Mistakes to Avoid

- Confusing P(A / B) with P(B / A)
- Forgetting to divide by P(B)
- Assuming P(A / B) is the same as P(B / A) — they are not

Bayes' Theorem

Bayes' Theorem is written as:

P(A / B) = [P(B / A) * P(A)] / P(B)

where:

- **P(A / B)** is the probability of event A occurring given that event B has occurred.
- **P(B / A)** is the probability of event B occurring given that event A has occurred.
- **P(A)** is the prior probability of event A.

This can also be interpreted as: P(A / B) = [P(B / A) * P(A)] / [P(B / A) * P(A) + P(B / No A) * P(No A)]

It calculates the probability of event A given that event B has occurred, using the known conditional probability of B given A.

Proof of Bayes' Theorem

We start with the definition of conditional probability:

1. **P(A / B) = P(A and B) / P(B)**
2. **P(B / A) = P(A and B) / P(A)**

From equation (2), we can write:

P(A and B) = P(B / A) * P(A)

Now substitute this into equation (1):

P(A / B) = [P(B / A) * P(A)] / P(B)

This is Bayes' Theorem.

Example: Medical Test Problem

Suppose a disease affects 1% of the population. A test for the disease has the following characteristics:

- If a person **has** the disease, the test is **positive** 99% of the time → P(Positive / Disease) = 0.99
- If a person **does not have** the disease, the test is **positive** 5% of the time → P(Positive / No Disease) = 0.05
- The chance of having the disease → P(Disease) = 0.01
- The chance of not having the disease → P(No Disease) = 0.99

A person takes the test and gets a **positive** result. What is the probability they actually have the disease?

Let:

- A = Disease
- B = Positive test

We apply Bayes' Theorem:

P(Disease / Positive) = [P(Positive / Disease) * P(Disease)] / P(Positive)

First, calculate P(Positive):

P(Positive) = P(Positive / Disease) * P(Disease) + P(Positive / No Disease) * P(No Disease)

P(Positive) = 0.99 * 0.01 + 0.05 * 0.99

P(Positive) = 0.0099 + 0.0495 = 0.0594

Now calculate P(Disease / Positive):

P(Disease / Positive) = 0.0099 / 0.0594 ≈ 0.1667

Conclusion: Even though the test is 99% accurate, the probability that a person actually has the disease given a positive result is only **16.67%**. This is because the disease is very rare, and false positives affect the result significantly.