Overview: Probability in Data Science

Terminology

- R**andom Experiment:**
  - (i) It has more than one possible outcome.
  - (ii) It is not possible to predict the outcome in advance
  - eg. **Tossing a coin**
- **Trial**:
  - Single execution of a random experiment.
  - Each trial produces an outcome.
  - eg. **Tossing a coin for 1 time → H/T**
- **Outcome**:
  - Outcome refers to a single possible result of a trial.
- **Sample Space:**
  - Sample Space of a random experiment is the **set of all possible outcomes** that can occur.
  - Generally, one random experiment will have one set of sample space.
  - eg. {H,T} , {1,2,3,4,5,6}
- **Event:**
  - Event is a **specific set of outcomes** from a random experiment or process.
  - subset of the sample space.
  - An event can include a single outcome, or it can include multiple outcomes.
  - One random experiments can have multiple events.

- e.g., rolling a die and getting a "3"

Example

Rolling a die:

1. **Random Experiment**: "Rolling a fair 6-sided die."
2. **Trial**: "One roll of the die."
3. **Outcome**: "The result of the roll, such as rolling a '3'."
4. **Sample Space**: "The set of all possible outcomes, $\{1, 2, 3, 4, 5, 6\}$."
5. **Event**: "Rolling an even number, which is the event $\{2, 4, 6\}$."

Tossing a coin:

1. **Random Experiment:**
   "Tossing a fair coin twice."
2. **Trial:**
   "One toss of the coin."
3. **Outcome:**
   "The result of the toss, such as 'Heads' or 'Tails'."
4. **Sample Space:**
   "The set of all possible outcomes, $\{HH, HT, TH, TT\}$, where H is Heads and T is Tails."
5. **Event:**
   "Getting at least one Head, which is the event $\{HH, HT, TH\}$."

Types of Events:

- **Simple Event**: An event that consists of exactly one outcome (e.g., rolling a die and getting a "3").

- **Compound Event**: An event that consists of two or more outcomes (e.g., rolling a die and getting an even number).
- **Impossible Event**: An event that cannot occur (e.g., rolling a 7 on a standard 6-sided die).
- **Certain Event**: An event that will always occur (e.g., rolling a number between 1 and 6 on a standard die).
- **Independent Events**: Two events are **independent** if the occurrence of one event does not affect the probability of the other event occurring.
    - Imagine flipping a coin and rolling a die:
        1. **Event A**: Getting **Heads** on the coin flip.
        2. **Event B**: Rolling a **3** on the die
- **Dependent Events:** Two events are **dependent** if the outcome of one event affects the probability of the other event occurring.
    - Imagine drawing two cards from a deck without replacement:
        1. **Event A**: Drawing an Ace on the first draw.
        2. **Event B**: Drawing an Ace on the second draw.
- **Mutually Exclusive Events:** Cannot happen at the same time.
    - "Heads" and "Tails" when tossing a coin.
- **Exhaustive Events:**
    - Events are **exhaustive** if, together, they cover all possible outcomes of an experiment.
    - In other words, at least one of the events must occur.

What is Probability

- Probability is a measure of the likelihood that a particular event will occur.
- A probability of 0 means that an event will not happen.
- A probability of 1 means that an event will certainly happen.

- A probability of 0.5 means that an event will happen half the time.

Empirical Probability Vs Theoretical Probability

Empirical Probability:

- Empirical probability, also known as experimental probability, is a probability measure that is based on observed data, rather than theoretical assumptions.
- It's calculated as the ratio of the number of times a particular event occurs to the total number of trials.
- ***eg. Suppose that, in our 100 tosses, we get heads 55 times and tails 45 times. What is the empirical probability of getting a head?***
    - ***Ans:*** 55/100

Theoretical Probability

- Theoretical (or classical) probability is used when each outcome in a sample space is equally likely to occur.
- ***Theoretical Probability of Event A = Number of Favourable Outcomes (that is, outcomes in Event A) / Total Number of Outcomes in the Sample Space***
- ***eg. Theoretical probability of getting 3 on a dice roll is 1/6.***

Random Variable

- Misleading Name
- It's a function. Not a variable.
- In the context of probability theory, a random variable is a function that **maps the outcomes of a random process** (k**nown as the sample space**) to a set of **real numbers**.
- eg. {H, T} → {1, 2}
    - {red, green, blue} → {1,2,3}

- Denoted by a capital number like X

- eg. Rolling 2 dice & event is getting a sum of 7
    - $X = \{1, 2, 3, \ldots 12\}$
    - 👆 **Logic**: To add the numbers.

Types of Random Variables:

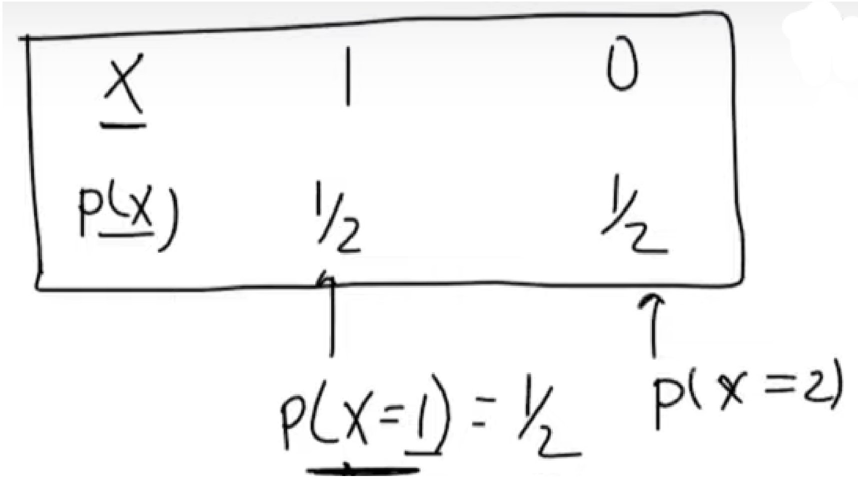
1. **Discrete Random Variable**: Takes on a finite or countably infinite number of possible values.
    - **Example**: The number of heads when flipping a coin 3 times. It can take values like 0, 1, 2, or 3.
2. **Continuous Random Variable**: Takes on an infinite number of possible values within a given range. These values are uncountable and can be measured on a continuous scale.
    - **Example**: The height of a person. It can take any value between a minimum and maximum (e.g., 5.5 feet, 5.55 feet, 5.555 feet, etc.).

Probability Distribution of a Random Variable

- A **probability distribution** describes how probabilities are distributed over the values of a random variable.

Types of Probability Distributions

- A probability distribution is a list of all of the possible outcomes of a random variable along with their corresponding probability values.



- 👆 Sample space along with their probability for a coin toss.

- **Rolling 2 dice:**

Input
(Sample Space)

(a,b)	1	2	3	4	5	6
1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
2	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
3	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
4	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
5	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
6	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

Output

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Unique Numbers: **X = {2,3,4,5,6,7,8,9,10,11,12}**

X	2	3	4	5	6	7	8	9	10	11	12
Probabilities	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36

1. Discrete Probability Distribution:

- **Definition**: Describes the probabilities of a discrete random variable.
- **Examples**:
    - **Uniform Distribution**: All outcomes are equally likely (e.g., rolling a fair die).
    - **Binomial Distribution**: Number of successes in a fixed number of trials (e.g., number of heads in 10 coin flips).
    - **Poisson Distribution**: Number of events in a fixed interval (e.g., number of emails received in an hour).

2. Continuous Probability Distribution:

- **Definition**: Describes the probabilities of a continuous random variable.
- **Examples**:
    - **Normal Distribution**: Symmetric, bell-shaped distribution (e.g., heights of people).
    - **Uniform Distribution**: All outcomes in a range are equally likely (e.g., time taken to complete a task).
    - **Exponential Distribution**: Time between events in a Poisson process (e.g., time between arrivals at a bus stop).

Probability Mass Function (PMF):

- **Definition**: Gives the probability that a discrete random variable is exactly equal to some value.
- **Example**: PMF of rolling a fair die:

$$P(X = x) = \frac{1}{6} \quad \text{for} \quad x = 1, 2, 3, 4, 5, 6$$

Probability Density Function (PDF):

- **Definition**: Describes the relative likelihood of a continuous random variable taking on a specific value.
- **Example**: PDF of a normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Cumulative Distribution Function (CDF):

- **Definition**: Gives the probability that a random variable is less than or equal to a certain value.
- **Example**: CDF of a normal distribution:

$$F(x) = P(X \leq x)$$

Mean of a Random Variable

- The **expected value** or **average value** of a random variable over many trials.
- You roll a die for 1000 times and calculate the mean.
  - 3+5+5+1+3+2+4+1+3.....*(1000 values)* /1000
- **Interpretation**: Represents the central tendency or "center of mass" of the random variable's distribution.

For a Discrete Random Variable:

$$E(X) = \sum_i x_i \cdot P(X = x_i)$$

- x_i: Possible values of the random variable.
- $P(X = x_i)$: Probability of x_i

Example: Rolling a fair die.

- Possible values: $\{1, 2, 3, 4, 5, 6\}$.
- Probabilities: $\frac{1}{6}$ for each value.
- Mean:

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

So, the mean of the random variable X (the outcome of rolling the die) is 3.5.

For a Continuous Random Variable:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

- $f(x)$: Probability density function (PDF) of the random variable.
- The integral is taken over the entire range of values that X can take.

Variance of a Random Variable

What is Variance?

- **Definition**: A measure of how spread out the values of a random variable are **around the mean**.
- **Interpretation**: A higher variance means the values are more spread out; a lower variance means they are closer to the mean.

Variance of a Random Variable

The **variance** of a random variable measures the **spread** or **dispersion** of its values around the mean (expected value).

$$\boxed{Var(x) = E[X^2] - (E[X])^2} \rightarrow \checkmark \quad \boxed{\text{cont} \quad \text{discrete}}$$

$$\boxed{Var(x) = E[(X - E[X])^2]} \checkmark$$

For a Discrete Random Variable:

$$\mathrm{Var}(X) = \sum_i (x_i - \mu)^2 \cdot P(X = x_i)$$

- x_i: Possible values of the random variable.
- μ: Mean of the random variable.
- $P(X = x_i)$: Probability of x_i.

- **Example**: Rolling a fair die.
  - Possible values: $\{1, 2, 3, 4, 5, 6\}$.
  - Probabilities: $\frac{1}{6}$ for each value.
  - Mean (μ): 3.5.
  - Variance:

$$\mathrm{Var}(X) = (1 - 3.5)^2 \cdot \frac{1}{6} + (2 - 3.5)^2 \cdot \frac{1}{6} + \cdots + (6 - 3.5)^2 \cdot \frac{1}{6} = 2.9167$$

For a Continuous Random Variable:

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)\, dx$$

- $f(x)$: Probability density function (PDF) of the random variable.
- μ: Mean of the random variable.