Probability Distributions

By: **Annamalai A** with **GPT** assist

Random Variable (RV)

A **random variable** is a numerical value assigned to the outcomes of a random experiment.

It is a *function* that maps outcomes from a sample space to real numbers.

Example

- Experiment: Tossing a coin
- Sample space: {H, T}
- Random variable mapping:
  - H → 1
  - T → -1

Thus, the RV assigns real values to possible outcomes.

Types of Random Variables

Discrete Random Variable

A **discrete random variable** takes **countable** values (0, 1, 2, 3, ...).

Examples:

- Number of cars in a parking lot
- Number of students who pass an exam
- Number of defects on a product

Continuous Random Variable

A **continuous random variable** takes **uncountable real values**.

Examples:

- Time to complete a task (e.g., 0 to 60 minutes)
- Age of a fossil
- Mileage (MPG) of a car

Continuous values lie on an **interval**, not a countable set.

Expectation (Mean)

The **expectation** or **expected value** $E(X)$ of a random variable is the average value we expect over many repetitions of an experiment.

Here, let X be a list of random variables $x_1, x_2, \ldots, x_k$, p_i be the probability of x_i and $p_1 + p_2 + p_3 + \ldots + p_k = 1$. Let $f(x)$ be probability density function

Then when data is:

- Discrete — weighted average of possible outcomes: $E(X) = \sum_{i=1}^{k} x_i\, p_i$
- Continuous — area under curve: $E(X) = \int x\, f(x)\, dx$

Variance

Variance measures how much values of a random variable deviate from the mean. It is the expected value of the squared deviation from the mean of X.

$$\operatorname{Var}(X) = E[(X - E(X))^2] = E[X^2] - E[X]^2$$

A higher variance means more spread in the data.

Covariance

Covariance measures how two random variables vary together.

$$\operatorname{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E[XY] - E[X]E[Y]$$

Interpretation:

- Positive → variables increase together
- Negative → one increases while the other decreases
- Zero → no linear relationship

Formula Summary

	Discrete Case	Continuous Case
Expectation $E(x)$	$E(X) = \sum_i x_i \, p_i$	$E(X) = \int_{-\infty}^{\infty} x \, f(x) \, dx$
Variance	$\mathrm{Var}(X) = \sum_i (x_i - E(X))^2 p_i$	$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) \, dx$
Mean Deviation from Mean	$\mathrm{MD} = \sum_i$	$x_i - E(X)$
Covariance	$\mathrm{Cov}(X, Y) = \sum_i (x_i - E(X))(y_i - E(Y)) p_i$	$\mathrm{Cov}(X, Y) = \int_{-\infty}^{\infty} (x - E(X))(y - E(Y)) f(x, y) \, dx \, dy$

Probability Distribution

A **probability distribution** describes how probabilities are assigned to different outcomes of a random variable.

Examples:

- Height of a population
- Survey results
- Weather conditions

Two types:

- For discrete RV → **Probability Mass Function (PMF)**
- For continuous RV → **Probability Density Function (PDF)**

Probability Mass Function (PMF)

PMF gives the probability that a discrete random variable takes a specific value.

Properties:

- $P(X = x) \geq 0$
- $\sum P(X = x) = 1$

Graph is usually a bar chart where probabilities sum to 1.

Probability Density Function (PDF)

PDF describes the likelihood of a continuous random variable taking a value.

Properties:

- $f(x) \geq 0$
- Total area under the curve = 1

Note:

For continuous variables,

$$P(X = x) = 0$$

Only intervals have probability.

Cumulative Distribution Function (CDF)

CDF gives the probability that the random variable is **less than or equal to x**.

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

Mean, Median, Mode

- **Mean** → arithmetic average
- **Median** → middle value
- **Mode** → most frequent value

Relation between Mean, Median, Mode

In a symmetric distribution:

$$\text{Mean} = \text{Median} = \text{Mode}$$

Empirical relation:

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

Bernoulli Distribution

A **Bernoulli distribution** models experiments with **two outcomes** (success/failure).

Let:

- Success = 1
- Failure = 0
- Probability of success = p

PMF:

$$P(X = x) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\}$$

Examples:

- Coin toss
- Yes/No response
- Pass/Fail experiment

Bernoulli = Binomial distribution with n = 1.

Binomial Distribution

A **binomial distribution** models the number of successes in *n independent Bernoulli trials*.

Parameters:

- n = number of trials
- p = probability of success in each trial

PMF:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Where $\binom{n}{k}$ means ${}^nC_k$

Examples:

- Number of heads in 10 coin tosses
- Number of defective items in a batch

Normal Distribution (Gaussian)

The **normal distribution** is a continuous distribution shaped like a bell curve.

PDF:

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

Where:

- μ = mean
- σ^2 = variance

Standard Normal Distribution:

- Mean = 0
- Variance = 1

Used widely in:

- Statistics
- Machine learning
- Finance
- Natural phenomena

Homoscedasticity vs Heteroscedasticity

Homoscedasticity

A sequence of variables has **constant variance**.

Example:

- Equal spread of points across values of X

Heteroscedasticity

Variance **changes** with the value of X.

Example:

- Spread increases as X increases
- Common in real-world data (e.g., income vs expenses)

Important for linear regression assumptions.