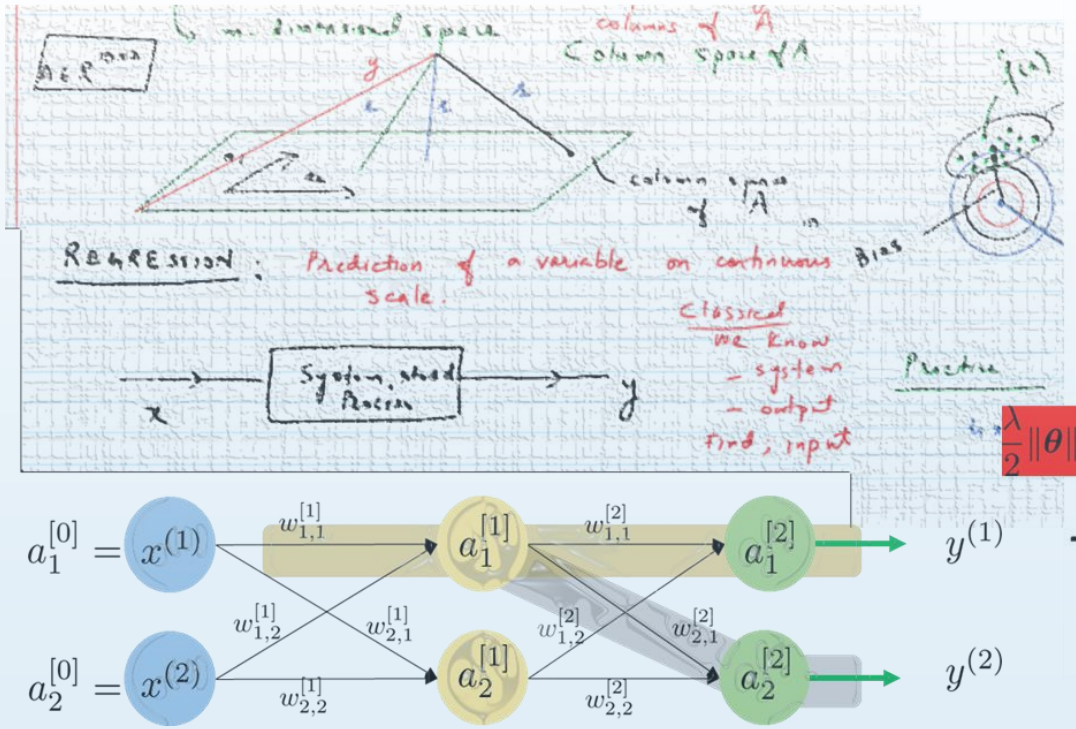
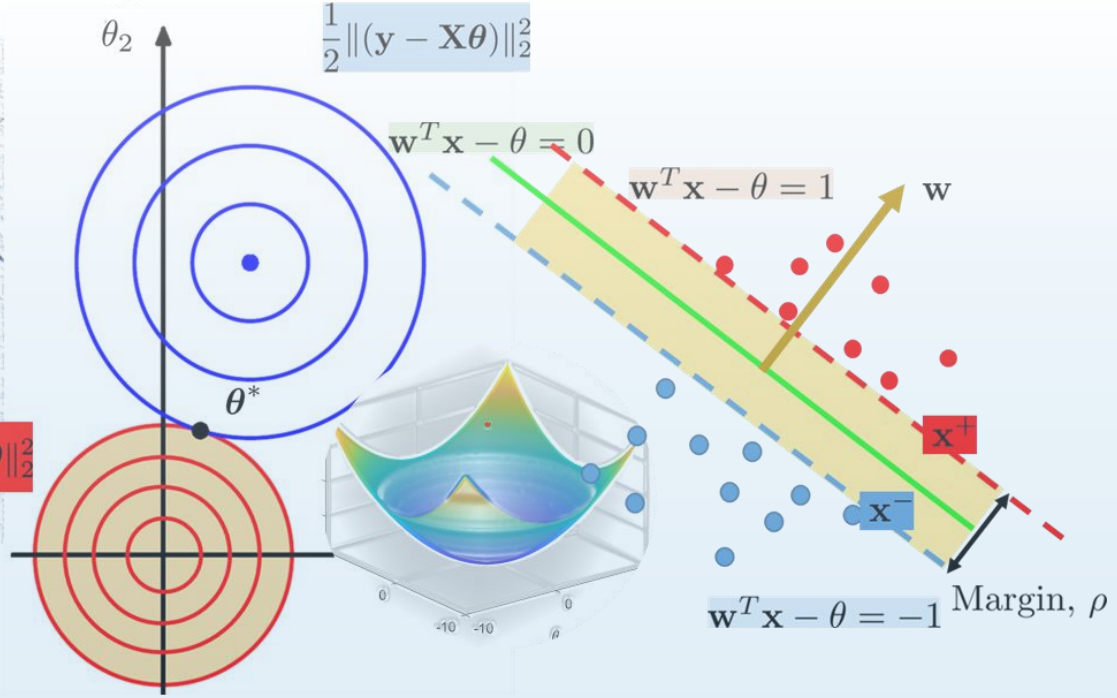



# AI-501 Mathematics for AI

## Probability Theory - Overview

Zubair Khalid
School of Science and Engineering

https://www.zubairkhalid.org/ai501_2024.html

LUMS
A Not-for-Profit University

# Why Probability Theory is Crucial for AI?

## Probability – Significance:

At the core of AI's ability to make **decisions, predict outcomes**, and **learn from data** lies a foundational pillar:

**PROBABILITY**

- By leveraging probability, AI systems gain the ability to
  - navigate uncertainty
  - make data-driven predictions, and
  - adapt effectively to ever-changing environments

# Why Probability Theory is Crucial for AI?

## Probability – Significance:

- **Handling Uncertainty**
  - Real-world data is noisy and incomplete
  - Probability theory provides a mathematical framework to reason about uncertainty

- **Foundation for Probabilistic Models**
  - Core of models like Bayesian networks, Hidden Markov Models, and Gaussian Mixture Models
  - Allows us to encode prior knowledge and update beliefs based on evidence

- **Bayesian Inference**
  - Key in machine learning for parameter estimation and model selection
  - Supports decision-making under uncertainty

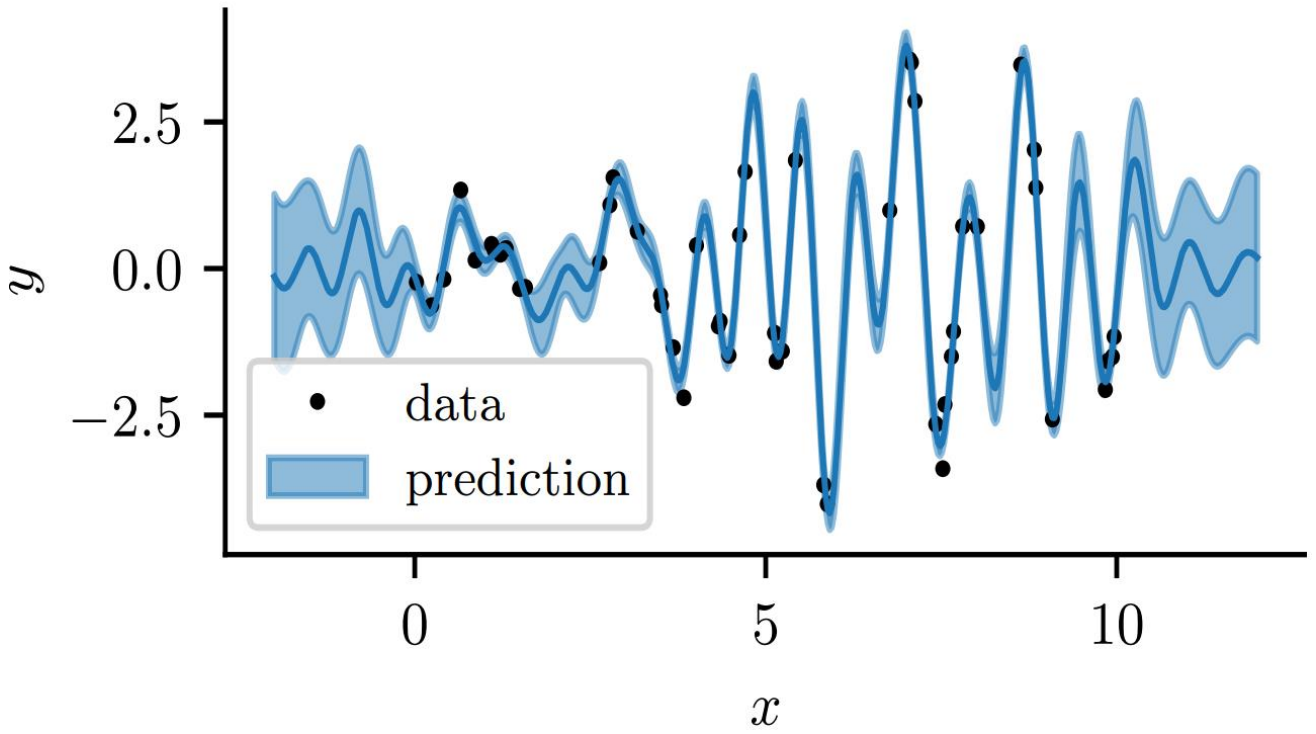
# Probabilistic AI

## Examples – Uncertainty Matters:

Given data of $N$ observations $D = \{(x_i, y_i)\}_{i=1}^{N}$

Find the best fit model $f$ that depends on the parameters $\theta$: $y = f(x, \theta)$

- Uncertainties:
  - Measurement noise in the data
  - Uncertainty in the values of estimated parameters
  - Uncertainty in the structure of the model
    - E.g., polynomial fit or neural network

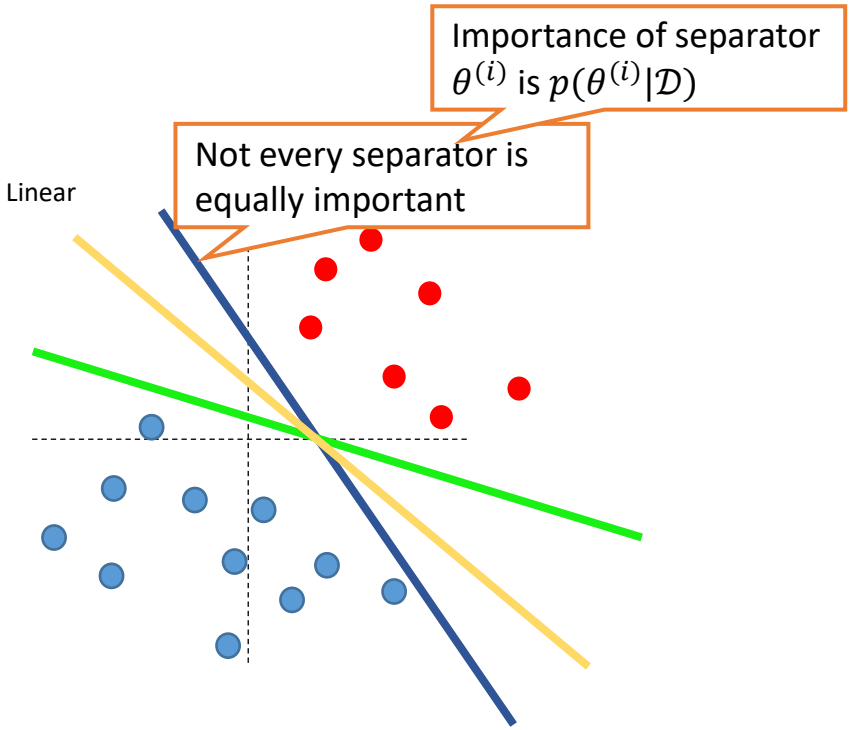



Probabilistic Machine Learning – Advanced Topics - Figure 18.23

# Probabilistic AI

## Uncertainty Types – Epistemic or Model Uncertainty – Example:

- Epistemic uncertainty is related to the model: both the structure and the parameters

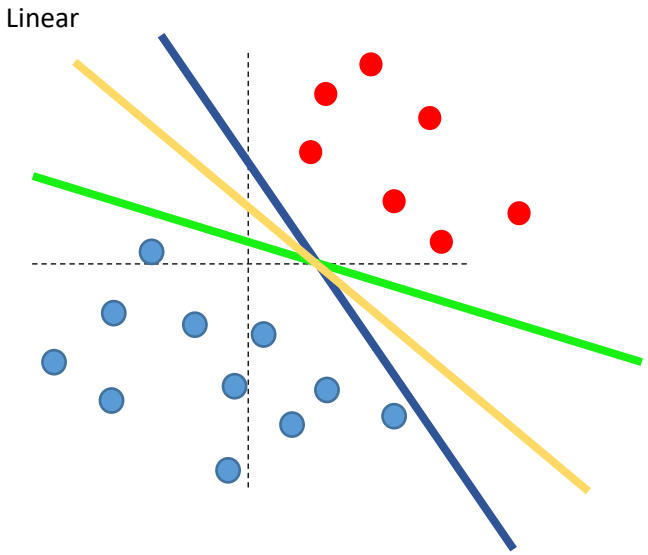



Linear separator but we have *uncertainty* about the weights

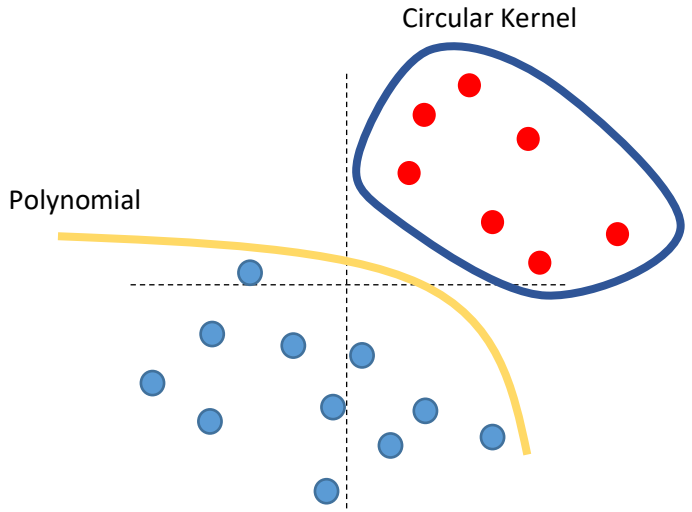
- Consider a linear classification model for 2-dim inputs
- Classifier weight will be a 2-dim vector $\theta = [\theta_1, \theta_2]$
- Its posterior will be some 2-dim distribution $p(\theta|\mathcal{D})$
- Sampling from this distribution will generate 2-dim vectors
- Each vector will correspond to a linear separator (left fig)
- Thus, the posterior in this case is equivalent to a “collection” or “ensemble” of weights, each representing a different linear separator

# Probabilistic AI

## Uncertainty Types – Epistemic or Model Uncertainty:




Linear separator but we have ***uncertainty*** about the weights




***Uncertainty*** about the model structure as well

Probabilistic approach to formulate model uncertainty:

Model structure or parameter distribution conditioned on data, for example:

$$p(\theta|D)$$

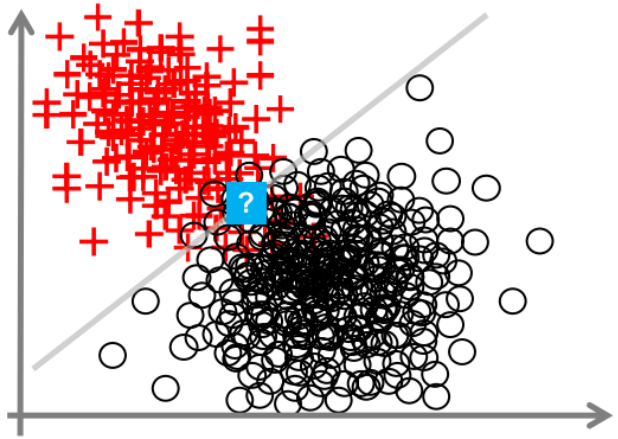
Also referred to as 'Posterior distribution' and is hard to compute, in general, but we will look at some methods to compute this.

*Model uncertainty is usually reducible with the increase in the amount of data.*

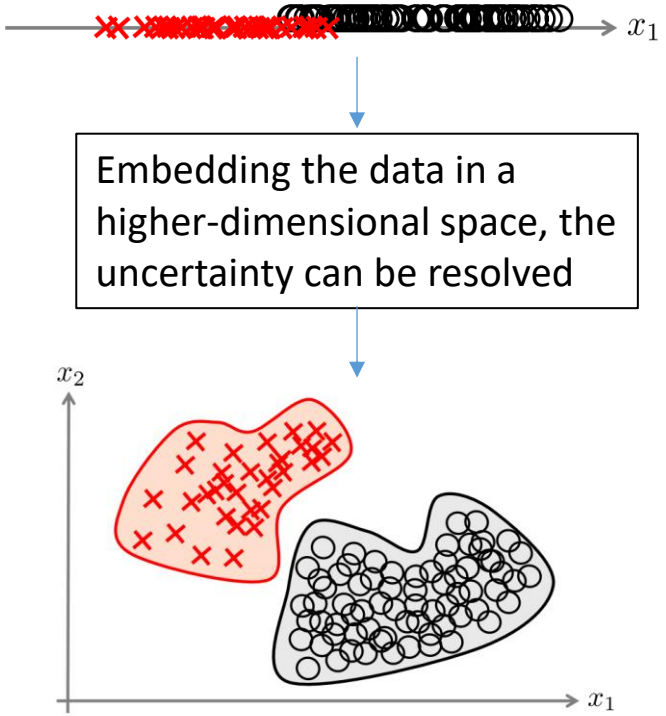
# Probabilistic AI

## Uncertainty Types – Aleatoric or Data Uncertainty:

- Aleatoric uncertainty is related to the data: noisy measurements, overlapping of classes, incorrect labelling, etc.




*Aleatoric uncertainty:* the prediction at the query point is uncertain




Probabilistic approach to formulate data uncertainty:

The distribution of data being modeled conditioned on model parameters and other inputs, for example:

$$p(y|\theta, x)$$

*Data uncertainty is mostly irreducible (even with infinite amount of data).*

*Sometimes reduced by adding more features or using more complex model*

# Probabilistic AI

## Overview:

Review the foundations of machine learning from the probabilistic and Bayesian perspective

We will answer *fundamental* questions:

- How do we set up a probabilistic model for a given machine learning problem?
- How do we quantify uncertainty in the process of estimation and prediction of parameters?
- What are the estimation and inference algorithms to learn the parameters of the model?

# Outline

- Overview of sets
- Probability Models
- Axioms of Probability
- Conditional Probability
- Independence
- Combinatorics
- Binomial Probability

# Probability Theory – Overview

## Nomenclature:

**Relative frequency**

- Consider an experiment that can result in $M$ possible outcomes $O_1$ , $O_2$ … $O_M$
- Let $N_n(O_i)$ denotes the number of times ,$O_i$ occurred in $n$ trials
- Relative frequency of outcome: $\frac{N_n(O_i)}{n}$
- When number of trials $n$ becomes large, the relative frequency converge to some limiting value.
- This behaviour is known as statistical regularity.

# Probability Theory – Overview

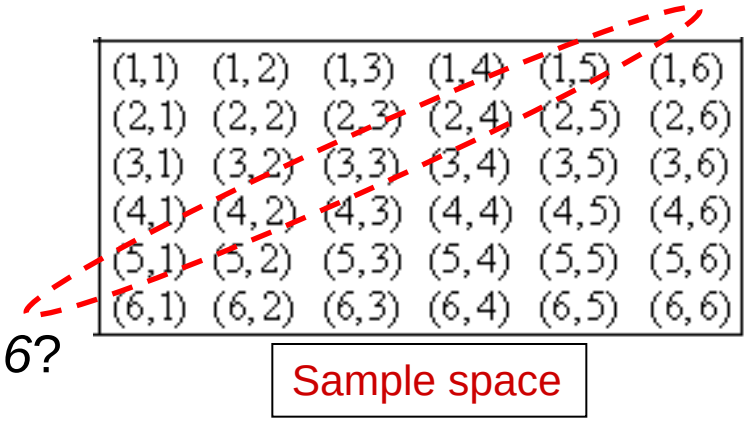
## Nomenclature:

- **Sample Space:** set of all possible distinct **outcomes** of an experiment.
  - Outcome: $\omega$
  - Sample space: $\Omega$

- **Events:** Collection of outcomes are called events. Usually denoted by capital letters. Every event is a subset of sample space.

Examples:

1. Rolling two dies together.
   - sample space?
   - event: the sum of numbers on two dies = *6*?



Sample space

2. Noise voltage can be between *0* and *5* volts.
   1. sample space?
   2. event: the noise voltage is between *2* and *3* volts?

# Probability Theory – Overview

## Language of Sets:

- $\omega \in \Omega$ , $A \subset \Omega$ . ($A$ is the event).
- Events $A$ and $B$ are equal, i.e., $A = B$ if $A \subset B$ and $B \subset A$.
- $A^c$ : Complement of $A$. Given by $A^c := \{\omega \in \Omega : \omega \notin A\}$.
- Empty set or Null set is denoted by Ø and represents no point in $\Omega$ .
- Union of two sets: $A \cup B := \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$.
- Intersection of two sets: $A \cap B := \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$
- Disjoint or mutually exclusive sets: $A \cap B = \varnothing$
- Set difference operation: $B \setminus A := B \cap A^c$

# Probability Theory – Overview

## Laws of Sets:

- Associative laws:

$$A \cup (B \cup C) = (A \cup B) \cup C$$

$$A \cap (B \cap C) = (A \cap B) \cap C$$

- Distributive laws:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

- De Morgan's laws:

$$(A \cap B)^c = A^c \cup B^c$$

$$(A \cup B)^c = A^c \cap B^c.$$

# Probability Theory – Overview

## Venn Diagram:

- Graphical representation of relationship between sets
- Rectangle represents sample space
- Ellipsoids represents events

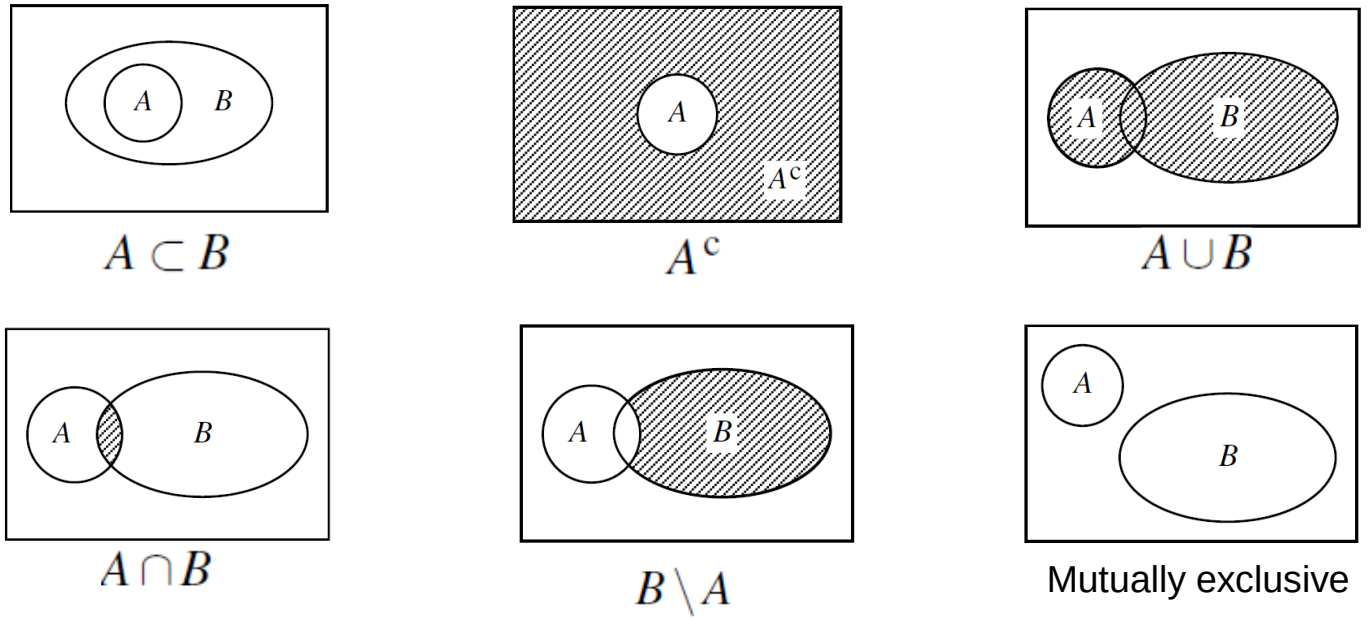



Fig 5: Venn diagram illustration

# Probability Theory – Overview

## Concept of Countable and Uncountable Sets:

- |*A*| denotes the number of points in set |*A*| and is called cardinality of set *A.*
- A nonempty set *A* is said to be countable if elements of *A* can be enumerated;

$$A = \bigcup_{k=1}^{\infty} \{a_k\}$$

- Every finite set is countable (No requirement that $a_k$ to be distinct).

Example: *A* = {a,e,i,o,u}.

- Empty set is also countable.
- If the set is not finite, we call it countably infinite. Examples?
- If the set is not countable, we call it uncountable or uncountably infinite.

# Probability Theory – Overview

## Probability Models:

- Mathematical modelling of the problem.
- Basic model if each outcome is equally likely: $P(A) = \frac{|A|}{|\Omega|}$
- Example: Fair die

**Properties:**

$P(\emptyset) = 0$

$P(A) \geq 0$

if $P(A \cap B) = \emptyset$, then
$P(A \cup B) = P(A) + P(B)$

$P(\Omega) = 1$

- If outcomes are not equally likely;

$$P(A) = \frac{|A|}{|\Omega|} = \sum_{\omega \in A} \frac{1}{|\Omega|} = \sum_{\omega \in A} p(\omega)$$

- $p(\omega)$ probability for each outcome $\omega \in \Omega$

# Probability Theory – Overview

## Probability Models:

***Example* 1.10.** Construct a sample space $\Omega$ and probability $\mathsf{P}$ to model an unfair die in which faces 1–5 are equally likely, but face 6 has probability $1/3$. Using this model, compute the probability that a toss results in a face showing an even number of dots.

# Probability Theory – Overview

## Probability Models:

***Example* 1.12.** A single card is drawn at random from a well-shuffled deck of playing cards. Find the probability of drawing an ace. Also find the probability of drawing a face card.

# Probability Theory – Overview

## Probability Models:

***Example* 1.13.** Suppose that we have two well-shuffled decks of cards, and we draw one card at random from each deck. What is the probability of drawing the ace of spades followed by the jack of hearts? What is the probability of drawing an ace and a jack (in either order)?

# Probability Theory – Overview

## Axioms of Probability:

Given a nonempty set $\Omega$, called the sample space, and a function $P$ defined on the subsets of $\Omega$, we say $P$ is a probability measure if the following four axioms are satisfied:

1. The empty set $\emptyset$ is called the impossible event. $P(\emptyset) = 0$.

2. For any event $A \subset \Omega, P(A) \geq 0$.

3. If $A1$, $A2$, ... are events that are mutually exclusive, that is, $A_n \cap A_m = \emptyset$ for $n \neq m$, then

$$P\left(\bigcup_{n=1}^{N} A_n\right) = \sum_{n=1}^{N} P(A_n).$$

4. $P(\Omega) = 1$, $\Omega$ is a sure event.

# Probability Theory – Overview

## Properties of Probability:

- Finite version of axiom 3:

$$P\left(\bigcup_{n=1}^{N} A_n\right) = \sum_{n=1}^{N} P(A_n).$$

- $P(A^c) = 1 - P(A)$

- $A \subset B \Rightarrow P(A) \leq P(B)$ Monotonicity property

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

# Probability Theory – Overview

## Conditional Probability:

• **Motivation:** Conditional probability, important concept in probabilistic modeling, allows us to update probabilistic models when additional information is revealed.

• Probability of event $A$ given event $B$;

$$P(A|B) = \frac{\text{outcomes in A and B}}{\text{outcomes in B}} = \frac{\text{outcomes in A and B}}{\text{total}} \; \frac{\text{total}}{\text{outcomes in B}} = \frac{P(A \cap B)}{P(B)}$$

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

• **The law of total probability**

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

• **Baye's Rule**

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|B^c)P(B^c)}$$

# Probability Theory – Overview

## Law of total probability:

Example:

Due to an Internet configuration error, packets sent from Karachi to Quetta are routed through Lahore with probability 3/4. Given that a packet is routed through Lahore, suppose it has conditional probability 1/3 of being dropped. Given that a packet is not routed through Lahore, suppose it has conditional probability 1/4 of being dropped. Find the probability that a packet is dropped.

# Probability Theory – Overview

## Bayes Rule:

Example:
Find the conditional probability that a packet is routed through Lahore given that it is not dropped.

# Probability Theory – Overview

## Generalized Laws:

- **The law of total probability:**

If $B1, B_2, \ldots$ denotes a sequence of pairwise mutually exclusive sets and $\sum_n P(B_n) = 1$, then,

$$P(A) = \sum_n P(A|B_n)P(B_n)$$

- **Baye's Rule**

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_n P(A|B_n)P(B_n)}$$

# Next

- Independence
- Combinatorics
- Examples
- Binomial Probabilities

# Independence

In probability theory, if events $A$ and $B$ satisfy $\mathsf{P}(A|B) = \mathsf{P}(A|B^c)$, we say $A$ does not depend on $B$. This condition says that

$$\Rightarrow \qquad \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(A \cap B^c)}{\mathsf{P}(B^c)}$$

Using $\mathsf{P}(A) = \mathsf{P}(A \cap B) + \mathsf{P}(A \cap B^c)$ and $\mathsf{P}(B^c) = 1 - \mathsf{P}(B)$

$$\Rightarrow \qquad \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(A) - \mathsf{P}(A \cap B)}{1 - \mathsf{P}(B)}$$

$$\Rightarrow \qquad \mathsf{P}(A \cap B)[1 - \mathsf{P}(B)] = \mathsf{P}(B)[\mathsf{P}(A) - \mathsf{P}(A \cap B)]$$

$$\Rightarrow \qquad \mathsf{P}(A \cap B) - \mathsf{P}(A \cap B)\mathsf{P}(B) = \mathsf{P}(A)\mathsf{P}(B) - \mathsf{P}(A \cap B)\mathsf{P}(B)$$

$$\Rightarrow \qquad \boxed{\mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B)}$$

# Independence

- Mutually exclusive events not to be confused with independent events. **(Different)**

- Interpretation of independence between events:

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

- If $A$ and $B$ are independent events, $A^c$ and $B$, $A$ and $B^c$, $A^c$ and $B^c$ are also independent events.

- Independence of more than 2 events:

$$\mathsf{P}(A_i \cap A_j \cap A_k) = \mathsf{P}(A_i)\,\mathsf{P}(A_j)\,\mathsf{P}(A_k)$$

- Mutually independent

- Pairwise independent

- Mutually independence implies pairwise independence but not the other way.

# Independence

Example: An internet packet travels through three stages:

1. From its source to Router 1.
2. From Router 1 to Router 2.
3. From Router 2 to its destination.

Each router drops packets independently with a probability $p$. What is the probability that the packet successfully reaches its destination without being dropped at any stage?

# Combinatorics

• Combinatorics: branch of mathematics that deals with systematic counting and arrangement methods.

- Counting problems:
  - ordered sampling with replacement
  - ordered sampling without replacement
  - unordered sampling without replacement

# Ordered sampling with replacement

**Example**

Let $A$, $B$, and $C$ be finite sets. How many triples are there of the form $(a, b, c)$, where $a \in A$, $b \in B$, and $c \in C$?

***Solution.*** Since there are $|A|$ choices for $a$, $|B|$ choices for $b$, and $|C|$ choices for $c$, the total number of triples is $|A| \cdot |B| \cdot |C|$.

Similar reasoning shows that for $k$ finite sets $A_1, \ldots, A_k$, there are $|A_1| \cdots |A_k|$ $k$-tuples of the form $(a_1, \ldots, a_k)$ where each $a_i \in A_i$.

**If $|A_1| \ldots |A_k| = n$, there are $n^k$ number of $k$- tuples.**

# Ordered sampling without replacement

- Means that the item chosen from any set is not replaced. The choice from the group affects the remaining choices.

**Example (Ordered Sampling Without Replacement)**

A computer virus erases files from a disk drive in random order. If there are $n$ files on the disk, in how many different orders can $k \leq n$ files be erased from the drive?

***Solution.*** There are $n$ choices for the first file to be erased, $n-1$ for the second, and so on. Hence, there are

$$n(n-1)\cdots(n-[k-1]) = \frac{n!}{(n-k)!}$$

different orders in which files can be erased from the disk.

Given a set $A$, we let $A^k$ denote the set of all $k$-tuples $(a_1, \ldots, a_k)$ where each $a_i \in A$. We denote by $A^k_*$ the subset of all $k$-tuples with *distinct* entries. If $|A| = n$, then $|A^k| = |A|^k = n^k$, and $|A^k_*| = n!/(n-k)!$.

- Often stated as: The number of **permutations** of $k$ items chosen from $n$ items.

# Unordered sampling without replacement

- Unordered; when order is not important: (1,2,3) is same as (3,2,1) or (2,1,3)…
- Given $n$ elements and choose $k$ tuples, we have: $\frac{n!}{(n-k)!}$
- Out of $\frac{n!}{(n-k)!}$ tuples, each $k$- tuple have $k!$ same arrangement or permutation.

- Total $k$- tuples with different permutation: $\frac{n!}{k!(n-k)!} = \binom{n}{k}$

- Often stated as: The number of **combinations** of $k$ items chosen from $n$ items.

# Combinatorics - Summary

- **Draw sample size of *k* out of *n* objects.**

| Method | No. of outcomes |
|---|---|
| **ordered** samples **with** replacement | $n^k$ |
| **ordered** samples **without** replacement | $\frac{n!}{(n-k)!}$ |
| unordered samples **without** replacement | $\frac{n!}{k!(n-k)!}$ |

# Combinatorics - Examples

***Problem***

In a pick-4 lottery game, a player selects four digits, each one from $0, \ldots, 9$. If the four digits selected by the player match the random four digits of the lottery drawing in any order, the player wins. If the player has selected four distinct digits, what is the probability of winning?

# Binomial Probabilities

- When there are only two events, win or loss, success or failure.

- Examples:
  - Tossing of a coin: head (win) or tail (loss)
  - Rolling die: getting 4 (win) or getting other than 4 (loss)

## Binomial Probabilities

A certain coin has probability $p$ of turning up heads. If the coin is tossed $n$ times, what is the probability that $k$ of the $n$ tosses result in heads? Assume tosses are independent.

$$\binom{n}{k} p^k (1-p)^{n-k}.$$

# Binomial Probabilities

***Example***

The probability of a student to score 90% in AI501 is 0.7 (**Assume it is true**). If 7 students appear in an exam, find the probability of exactly 4 getting 90% marks.

# Outline

- Discrete random variable
- Probability mass function
- Multiple random variable
- Independence of random variables
- Joint probability mass function
- Expectation

# Discrete Random Variable

- **Random Variable (Definition):**

  Random variable is a **function** which **maps** elements from the **sample space** to the **real line**.

- Random Variables are denoted by upper case letters ($X$ or $Y$).
- Individual outcomes for RV are denoted by lower case letters ($x$ or $y$).
- Mathematically, $X(\omega)$ is a real-valued function defined for $\omega \in \Omega$.
- For each element of an experiment's sample space, the random variable can take on exactly one value.
- **Discrete Random Variable: A RV that can take on only a finite or countably infinite set of outcomes.**

# Example 1

A random variable $X(\omega)$ = number of heads if three coins are tossed at the same time

Sampe space: $\Omega := \{\text{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH}\}$

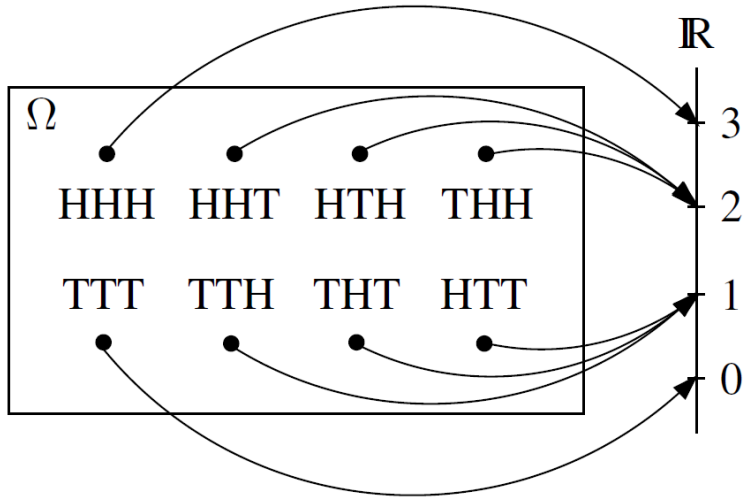



Fig 1: Illustration of RV mapping

$$X(\omega) := \begin{cases} 0, & \omega = \text{TTT}, \\ 1, & \omega \in \{\text{TTH, THT, HTT}\}, \\ 2, & \omega \in \{\text{THH, HTH, HHT}\}, \\ 3, & \omega = \text{HHH}. \end{cases}$$

# Example 2

A random variable $X(\omega)$ = number of girls in a family of 4 kids.

Lower case $x$ is a particular value of $X(\omega)$.

| $\omega$ | **Random Variable *X*** |
|---|---|
| BBBB | *x=0* |
| GBBB | *x=1* |
| BGBB | *x=1* |
| BBGB | *x=1* |
| BBBG | *x=1* |
| GGBB | *x=2* |
| GBGB | *x=2* |
| GBBG | *x=2* |
| BGGB | *x=2* |
| BGBG | *x=2* |
| BBGG | *x=2* |
| BGGG | *x=3* |
| GBGG | *x=3* |
| GGBG | *x=3* |
| GGGB | *x=3* |
| GGGG | *x=4* |

# Example 3

Random variable, $Y$ = Sum of the up faces of the two die.

Die 1 →

Die 2 ↓

| | 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 |

| y |
|---|
| 2 |
| 3 |
| 4 |
| 5 |
| 6 |
| 7 |
| 8 |
| 9 |
| 10 |
| 11 |
| 12 |

# Probability Mass Function (pmf)

- **Probability Mass Function:** Assigns probabilities (masses) to the individual outcomes. (Also referred as probability density function.)

- For a random variable $X$, its pmf is given by

$$p_X(x_i) := \mathrm{P}(X = x_i)$$

- By axioms of probability;
  - pmf is between 0 and 1 $$0 \leq p_X(x_i) \leq 1$$
  - sum of all probabilities equal to 1 $$\sum_i p_X(x_i) = 1$$

# Example 1

A random variable $X(\omega)$ = number of heads if three coins are tossed at the same time

$$p_X(0) = \mathrm{P}(X=0) = \mathrm{P}(\{\mathrm{TTT}\}) = \frac{|\{\mathrm{TTT}\}|}{|\Omega|} = \frac{1}{8}$$

$$p_X(1) = \mathrm{P}(X=1) = \mathrm{P}(\{\mathrm{HTT,THT,TTH}\}) = \frac{3}{8}$$

$$p_X(2) = \mathrm{P}(X=2) = \mathrm{P}(\{\mathrm{HHT,HTH,HHT}\}) = \frac{3}{8}$$

$$p_X(3) = \mathrm{P}(X=3) = \mathrm{P}(\{\mathrm{HHH}\}) = \frac{1}{8}$$

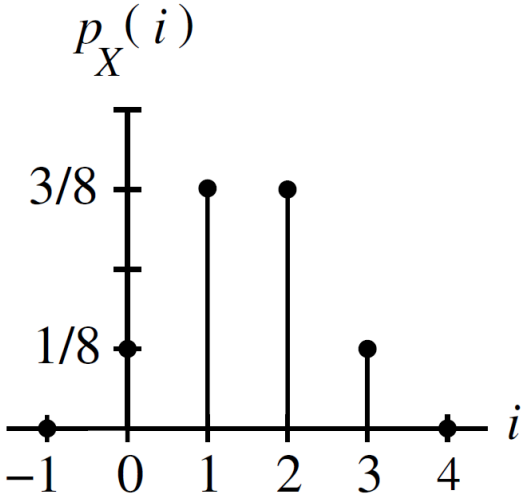



Fig 2: pmf of RV $X$

# Example 2

A random variable $X$ = number of girls in a family of 4 kids

| Number of Girls, $x$ | Probability, $p_X(x)$ |
|---|---|
| 0 | 1/16 |
| 1 | 4/16 |
| 2 | 6/16 |
| 3 | 4/16 |
| 4 | 1/16 |
| Total | 16/16=1.00 |

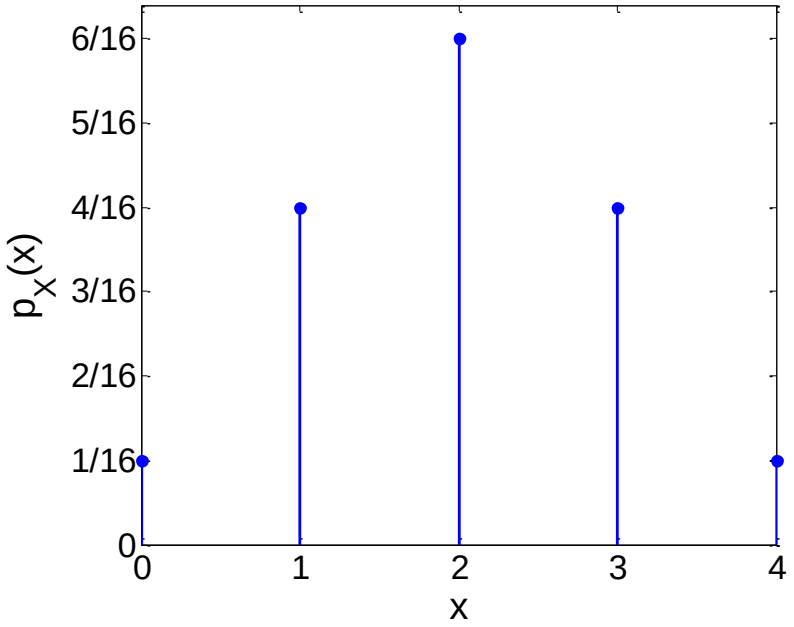



Fig 3: pmf of RV

What is the probability of exactly 3 girls in 4 kids?

What is the probability of at least 3 girls in 4 kids?

# Example 3

Random variable, $Y$ = Sum of the up faces of the two die.

| $y$ | $p_Y(y)$ |
|---|---|
| 2 | 1/36 |
| 3 | 2/36 |
| 4 | 3/36 |
| 5 | 4/36 |
| 6 | 5/36 |
| 7 | 6/36 |
| 8 | 5/36 |
| 9 | 4/36 |
| 10 | 3/36 |
| 11 | 2/36 |
| 12 | 1/36 |

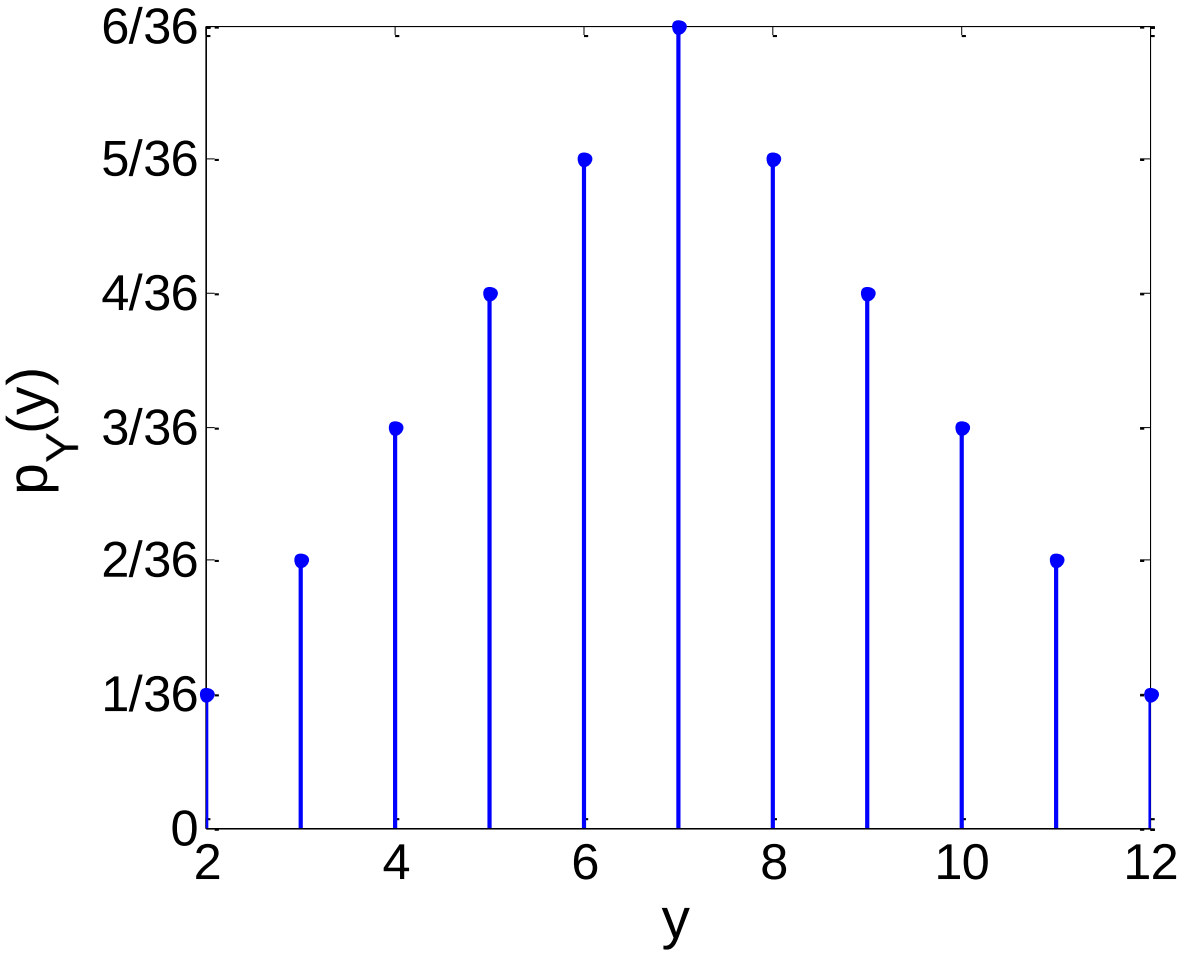



Fig 4: pmf of RV

# Important Random Variables

1. **Bernoulli Random Variable**

   If there are only two outcomes of an experiment, the experiment is modeled with uniform random variable. For example, the tossing of coin is modeled with Bernoulli random variable.

   - It is most common to associate {0,1} to the outcomes of an experiment.
   - pmf is given by,

$$p_X(0) = p$$
$$p_X(1) = 1 - p$$

# Important Random Variables

**2. Uniform Random Variable:**

If the outcomes of an experiment are finite, and are equally likely, the experiment is modeled with uniform random variable.

- If there are *n* outcomes of an experiment, probability of each outcome = $\frac{1}{n}$ .
- If outcomes are indexed, *k=1, 2,..., n*, $\mathrm{P}(X = k) = \frac{1}{n}, \quad k = 1, \ldots, n$
- pmf is given by,

$$p_X(k) = \begin{cases} 1/n, & k = 1, \ldots, n, \\ 0, & \text{otherwise.} \end{cases}$$

# Important Random Variables

**3. Poisson Variable:**

A random variable $X$ is said to have a Poisson probability mass function with parameter $\lambda > 0$, denoted by $X \sim \text{Poisson}(\lambda)$, if

$$p_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots.$$

- Parameter $\lambda$ fully characterizes the distribution.

- Used in modelling of physical phenomenon arising in different applications:
- arrival of photons at a telescope
- distribution of nodes in wireless sensor networks
- telephone calls arriving in a system
- arrival of network messages in a queue for transmission

# Important Random Variables - Examples

***Example* 2.6.** Ten neighbors each have a cordless phone. The number of people using their cordless phones at the same time is totally random. Find the probability that more than half of the phones are in use at the same time.

***Example* 2.7.** The number of hits to a popular website during a 1-minute interval is given by a Poisson($\lambda$) random variable. Find the probability that there is at least one hit between 3:00 am and 3:01 am if $\lambda = 2$. Then find the probability that there are at least 2 hits during this time interval.

***Example* 2.16.** A light sensor uses a photodetector whose output is modeled as a Poisson($\lambda$) random variable $X$. The sensor triggers an alarm if $X > 15$. If $\lambda = 10$, compute $\mathrm{P}(X > 15)$.

# Multiple Random Variable

- When events are defined by more than one random variable.
- Let $X$ represent one variable and $Y$ represent another random variable, which maps elements of sample space to real line, but can be different, then the event involving both $X$ and $Y$ is described as

$$\{X \in B, Y \in C\} := \{\omega \in \Omega : X(\omega) \in B \text{ and } Y(\omega) \in C\}$$

- This is taken as an event that $X$ belongs to $B$ and $Y$ belongs to $C$.
- **Very important to understand the concept:** the event above is a function of two random variable and is comprised of only those points on the real line which are common between $B$ and $C$, that is,

$$\{X \in B, Y \in C\} = \{X \in B\} \cap \{Y \in C\}$$

# Multiple Random Variable – Probability mass function

- The joint probability involving two random variables is given by the probability of the joint event

$$\begin{aligned} \mathsf{P}(X \in B, Y \in C) &:= \mathsf{P}(\{X \in B, Y \in C\}) \\ &= \mathsf{P}(\{X \in B\} \cap \{Y \in C\}) \end{aligned}$$

- taking $B = \{x_i\}$ and $C = \{y_j\}$, define joint probability mass function,

$$p_{XY}(x_i, y_j) := \mathsf{P}(X = x_i, Y = y_j)$$

- **Interpretation:** $p_{XY}(x_i, y_j)$ gives the probability that the RV $X = x_i$ and RV $Y = y_i$ at the same time.

- **Marginal probability mass function:** We can obtain $p_X(x_i)$ and $p_Y(y_j)$

$$p_X(x_i) = \sum_j p_{XY}(x_i, y_j)$$

$$p_Y(y_j) = \sum_i p_{XY}(x_i, y_j)$$

# Multiple Random Variable – Concept of Independence

- When RVs $X$ and $Y$ are independent events, we can write the joint probability as

$$\mathsf{P}(X \in B, Y \in C) = \mathsf{P}(X \in B)\mathsf{P}(Y \in C)$$

$$\mathsf{P}(X = x_i, Y = y_j) = \mathsf{P}(X = x_i)\,\mathsf{P}(Y = y_j)$$

- Equivalently, we can write in terms of joint pmf and individual pms of RVs:

$$p_{XY}(x, y) = p_X(x)p_Y(y)$$

- The concepts presented for two random variables are also valid for more than two random variables.

# Multiple Random Variable – Examples

***Example* 2.8.** On a certain aircraft, the main control circuit on an autopilot fails with probability $p$. A redundant backup circuit fails independently with probability $q$. The aircraft can fly if at least one of the circuits is functioning. Find the probability that the aircraft cannot fly.

***Example* 2.9.** Let $X$, $Y$, and $Z$ be the number of hits at a website on three consecutive days. Assuming they are i.i.d. Poisson($\lambda$) random variables, find the probability that on each day the number of hits is at most $n$.

# Expectation of a Random Variable

- Expectation of a random variable is defined as;

$$\mathsf{E}[X] := \sum_i x_i \mathsf{P}(X = x_i)$$

$$\boxed{\mathsf{E}[X] = \sum_i x_i \, p_X(x_i)}$$

- Expectation of a random variable gives an average value of the values $x_1, x_2, \ldots,$ a random variable can take with probabilities $\mathsf{P}(X = x_1), \mathsf{P}(X = x_2), \ldots$.

- Expectation is a linear operator: $\mathsf{E}[aX + bY] = \mathsf{E}[aX] + \mathsf{E}[bY] = a\mathsf{E}[X] + b\mathsf{E}[Y]$

- **law of the unconscious statistician (LOTUS):** If another RV $Y$ is a function of RV $X$ given by, $Y = f(X)$, the expected value of RV $Y$ is given in terms of pmf of the RV $X$ as

$$\boxed{\mathsf{E}[Y] = \mathsf{E}[f(X)] = \sum_i f(x_i) p_X(x_i)}$$

# Expected Values of Discrete RV's

- Mean : Long-run average value a RV.
- Variance – Average squared deviation between a realization of a RV and its mean. Quantifies the spread around
- Standard Deviation – Positive square root of variance , measure of spread.
- Notation:
  - Mean: $\mathsf{E}[X] = m$
  - Variance: $\text{var}(X) = \mathsf{E}[(X - m)^2] = \sigma^2$
  - Standard Deviation: $\sigma$

# Moments of Random Variable

**Moments:**

- $n-$th momemt of a RV $X$ is defined as $\mathsf{E}[X^n]$.
- Mean, $x = \mathsf{E}[X]$ is the first moment.

**Central Moments - Moments around center (mean):**

- $n-$th central momemt of a RV $X$ is defined as $\mathsf{E}[(X-m)^n]$.
- Variance, $\text{var}(X) = \mathsf{E}[(X-m)^2] = \sigma^2$ is the second central moment.
- Skewness: $\mathsf{E}[(X-m)^3]/\sigma^3$
- Kurtosis: $\mathsf{E}[(X-m)^4]/\sigma^4$

# Variance of Random Variable

- Variance, $\text{var}(X) = \mathsf{E}[(X-m)^2] = \sigma^2$, is often computed as

$$\text{var}(X) = \mathsf{E}[X^2] - (\mathsf{E}[X])^2$$

Derivation:

$$\begin{aligned}
\text{var}(X) &:= \mathsf{E}[(X-m)^2] \\
&= \mathsf{E}[X^2 - 2mX + m^2] \\
&= \mathsf{E}[X^2] - 2m\mathsf{E}[X] + m^2, \\
&= \mathsf{E}[X^2] - m^2 \\
&= \mathsf{E}[X^2] - (\mathsf{E}[X])^2.
\end{aligned}$$

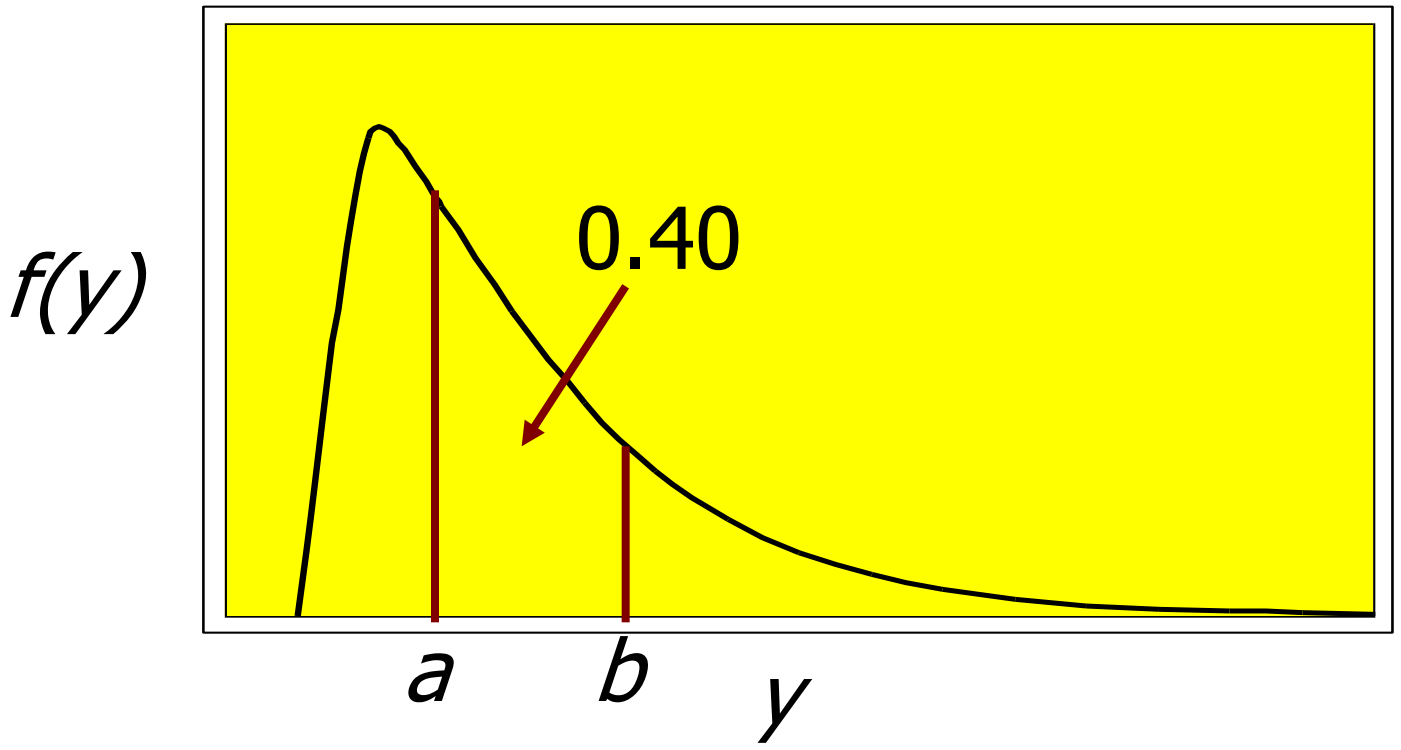
# Continuous Random Variable

# Continuous Random Variable

- A continuous random variable is one for which the outcome can be any value in an interval of the real number line.
- There are always infinitely many sample points in the sample space.
- For **discrete** random variables, only the value listed in the **pmf** have positive probabilities, all other values have probability zero.
- For continuous random variables, the probability of every specific value is zero. Probability only exists for an interval of values for continuous RV., that is, for continuous RV $Y$,
  - We don't calculate $P(Y = y)$, we calculate $P(a < Y < b)$, where $a$ and $b$ are real numbers.
  - For a continuous random variable $P(Y = y) = 0$.

# Probability density function

- The **probability density function (pdf)** denotes a curve against the possible values of random variable and the area under an interval of the curve is equal to the probability that random variable is in that interval.
- For example if *f(y)* denotes the pdf of RV *Y*, we calculate $P(a < Y < b)$,



# pmf versus pdf

- For a discrete random variable, we have probability mass function (pmf).
- The pmf looks like a bunch of spikes, and probabilities are represented by the heights of the spikes.
- For a continuous random variable, we have a probability density function (pdf).
- The pdf looks like a curve, and probabilities are represented by areas under the curve.

# Characteristics of pdf

- Given $Y$ is a continuous random variable with pdf is $f(x)$ .
- By axioms of probability, $f(x)$ must satisfy the following conditions:

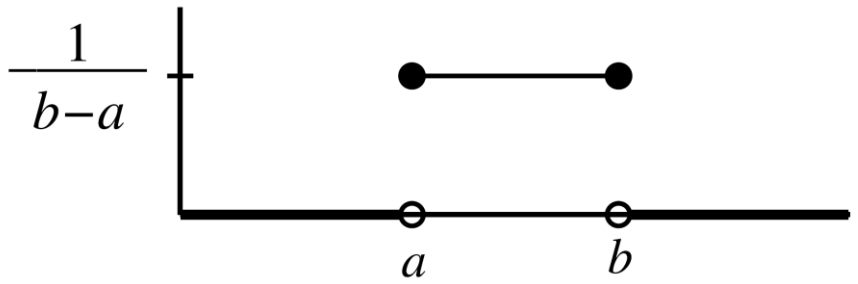
1. $f(x) \geq 0$ for all $x \in R$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

# Important Continuous RVs

- **Uniform random variable:** used to model the experiments in which outcome is constrained to lie in a known interval, say [$a$,$b$] and all possible outcomes are equally likely.
- Define uniform random variable $f \sim$ uniform[a,b] for $a<b$ with pdf

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}$$

- Plot of pdf

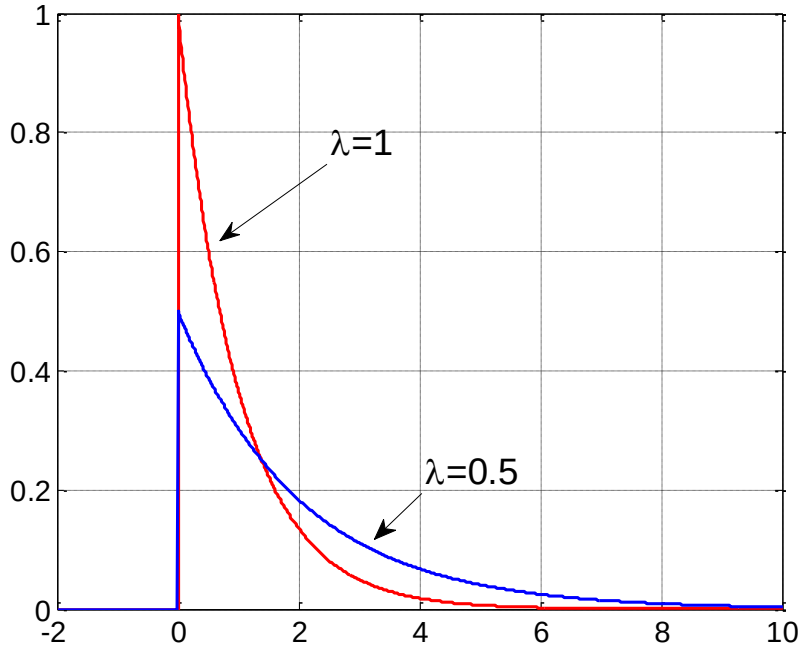


# Important Continuous RVs

- **Exponential random variable:** used to model lifetimes, such as
  - how long it takes before next phone call arrivrs
  - how long it takes a computer network to transmit a message
  - how long it takes a radioactive particle to decay

- Define $f \sim \exp(\lambda)$ for $\lambda > 0$ with pdf given by:

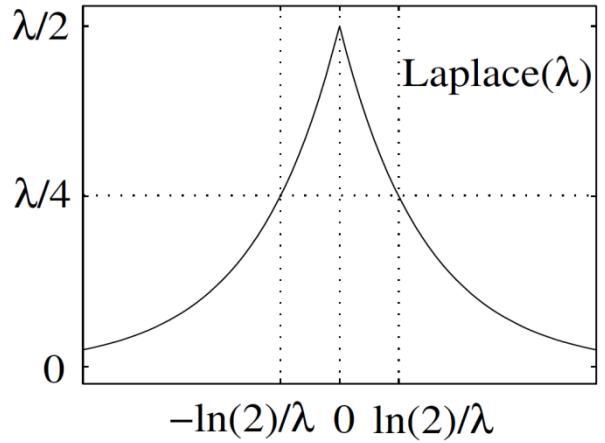
$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$



# Important Continuous RVs

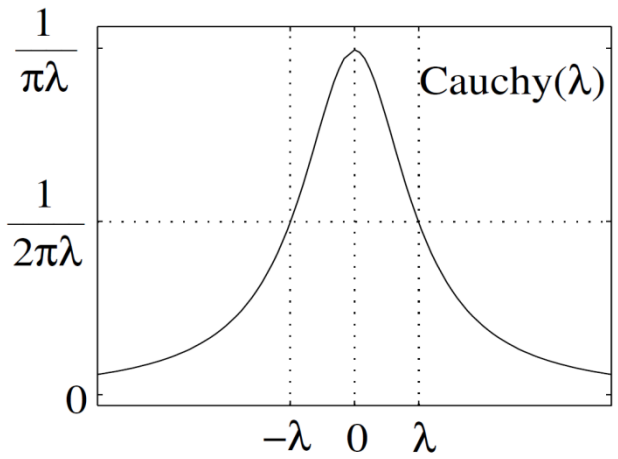
- **Laplace (double sided exponential) random variable :**
- Denoted by $f \sim \text{Laplace}(\lambda)$ for $\lambda > 0$ :

$$f(x) = \frac{\lambda}{2} e^{-\lambda |x|}$$




- **Cauchy random variable:**
- Denoted by $f \sim \text{Cauchy}(\lambda)$ for $\lambda > 0$ :

$$f(x) = \frac{\lambda/\pi}{\lambda^2 + x^2}$$

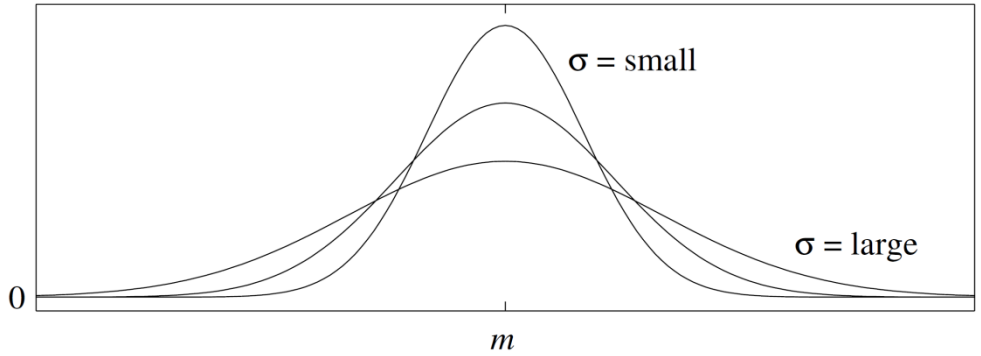


# Important Continuous RVs

- **Gaussian (Normal) random variable:**
- Define Gaussian RV $f \sim N(m, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right]$$

- center $m \in R$
- standard deviation, $\sigma^2 \in R^+$ , quantifies the spread of the pdf
- $N(0,1)$ is called standard normal density



# Some Examples

***Example* 4.1.** In coherent radio communications, the phase difference between the transmitter and the receiver, denoted by $\Theta$, is modeled as having a density $f \sim \text{uniform}[-\pi, \pi]$. Find $\mathsf{P}(\Theta \leq 0)$ and $\mathsf{P}(\Theta \leq \pi/2)$.

# Some Examples

***Example* 4.4.** An Internet router can send packets via route 1 or route 2. The packet delays on each route are independent exp($\lambda$) random variables, and so the difference in delay between route 1 and route 2, denoted by $X$, has a Laplace($\lambda$) density. Find

$$\mathsf{P}(-3 \leq X \leq -2 \text{ or } 0 \leq X \leq 3).$$

# Some Examples

***Problem.*** For Gaussian RV $f \sim N(m, \sigma^2)$, show that $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

You may use the following information to derive the result :

$$\int_{-\infty}^{\infty} e^{-y^2}\,dy = \sqrt{\pi}$$

# Expectation of Random Variable

• Law of the unconscious statistician (LOTUS) version for continuous random variable $X$ :

$$\mathsf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx$$

• Recall, mean or average, $m = \mathsf{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx$

• $\mathsf{var}(X) = \mathsf{E}[X^2] - (\mathsf{E}[X])^2$

**Question:** If $X$ is a uniform random variable on the interval $[a, b]$, find $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, and $\mathrm{Var}(X)$.

**Solution:** The PDF of $X \sim U(a, b)$ is:

$$f_X(x) = \frac{1}{b-a}, \quad a \le x \le b.$$

$$\mathbb{E}[X] = \int_a^b x f_X(x)\, dx = \frac{1}{b-a} \int_a^b x\, dx.$$

$$\mathbb{E}[X] = \frac{1}{b-a} \left[\frac{x^2}{2}\right]_a^b = \frac{a+b}{2}.$$

$$\mathbb{E}[X^2] = \int_a^b x^2 f_X(x)\, dx = \frac{1}{b-a} \int_a^b x^2\, dx.$$

$$\mathbb{E}[X^2] = \frac{1}{b-a} \left[\frac{x^3}{3}\right]_a^b = \frac{b^2+ab+a^2}{3}.$$

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Substitute values:

$$\mathrm{Var}(X) = \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

**Question:** If $X$ is an exponential random variable with parameter $\lambda = 1$, find all moments of $X$.

**Solution:** The probability density function (PDF) of an exponential random variable $X \sim \text{Exp}(\lambda)$ is:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

For $\lambda = 1$, the $n$-th moment of $X$ is given by:

$$\mathbb{E}[X^n] = \int_0^\infty x^n e^{-x}\, dx.$$

The integral of $x^n e^{-x}$ is related to the Gamma function:

$$\Gamma(n+1) = \int_0^\infty x^n e^{-x}\, dx.$$

$$\mathbb{E}[X^n] = \int_0^\infty x^n e^{-x}\, dx = \Gamma(n+1) = n!$$