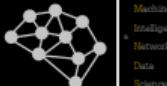


# ECE 302: Probabilistic Methods in ECE

---

Abolfazl Hashemi, Purdue ECE



# Table of contents i

1. Lecture 1: Probability Models and Axioms
2. Lecture 2: Conditioning and Bayes' Rule
3. Lecture 3: Independence
4. Lecture 4: Counting
5. Lecture 5: Discrete Random Variables Part I  
Probability Mass Functions
6. Lecture 6: Discrete Random Variables Part II  
Expectation; Variance; Conditioning
7. Lecture 7: Discrete Random Variables Part III  
Multiple Random Variables; Conditioning on a Random Variable; Independence of r.v.'s
8. Lecture 8: Continuous Random Variables Part I  
Probability Density Functions



## Table of contents ii

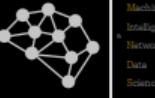
- 9. Lecture 9: Continuous Random Variables Part II  
Conditioning on an Event; Multiple Continuous r.v.'s
- 10. Lecture 10: Continuous Random Variables Part III  
Conditioning on a Random Variable; Independence; Bayes' Rule
- 11. Lecture 11: Derived Distributions
- 12. Lecture 12: Sums of Independent Random Variables; Covariance and Correlation
- 13. Lecture 13: Conditional expectation and variance revisited; Sum of a random number of independent r.v.'s
- 14. Lecture 14: Bi-variate and Multivariate Normal
- 15. Lecture 15: Transforms and Moment Generating Functions (MGFs)
- 16. Lecture 16: Introduction to Bayesian Inference
- 17. Lecture 17: Linear Models With Normal Noise

# Table of contents iii

18. Lecture 18: Least Mean Squares (LMS) Estimation
19. Lecture 19: Linear Least Mean Squares (LLMS) Estimation
20. Lecture 20: Inequalities, Convergence, and the Weak Law of Large Numbers
21. Lecture 21: The Central Limit Theorem (CLT)
22. Lecture 22: Classical Statistics I
23. Lecture 23: Classical Statistics II
24. Lecture 24: The Bernoulli Process
25. Lecture 25: The Poisson Process Part I
26. Lecture 26: The Poisson Process Part II
27. Lecture 27: Markov Chains I
28. Lecture 28: Markov Chains II
29. Lecture 29: Markov Chains III



## 30. Lecture 30: Stationarity and Ergodicity

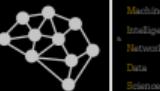


# Introduction to me

- Abolfazl Hashemi, Ph.D. (email: abolfazl@purdue.edu)
- Assistant Professor at The Elmore Family School of Electrical and Computer Engineering at Purdue University, Since Fall 2021
- Research Goal: advance the field of Large-Scale Optimization provides actionable insights from the perspective of this foundational field to innovate multiple domains within ML/AI
- Recent Applications: Federated Learning, Medical Image Analysis, NextG Manufacturing, and Cyber-Physical Systems
- Consistently using probabilistic methods, theorems, and arguments in my research.

# **Lecture 1: Probability Models and Axioms**

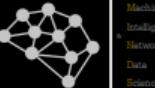
---



# Sample Space

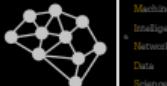
Two steps:

- Describe possible outcomes
- Describe beliefs about likelihood of outcomes



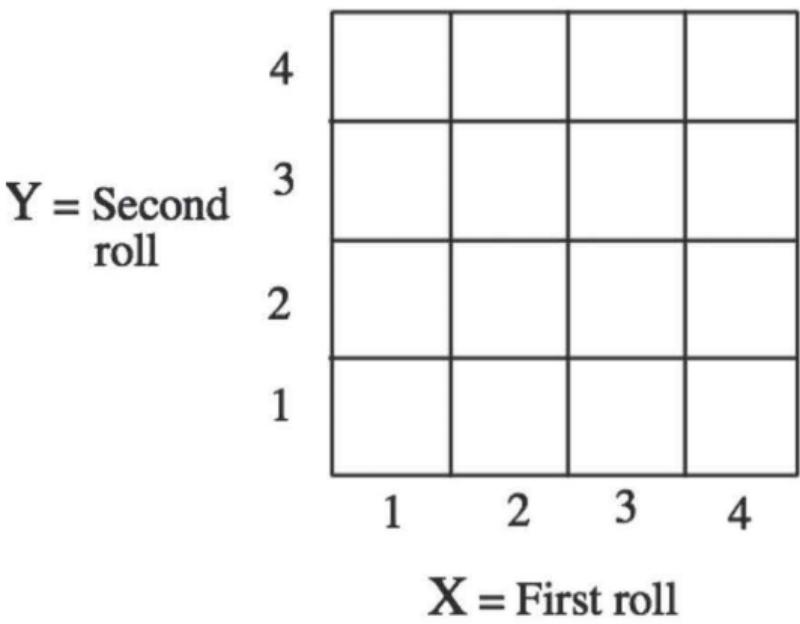
# Sample Space $\Omega$

- A list (set) of possible outcomes, denoted by  $\Omega$
- The list must be:
  - Mutually exclusive
  - Collectively exhaustive
- It must be at the “right” granularity



# Sample Space: Discrete/Finite Example

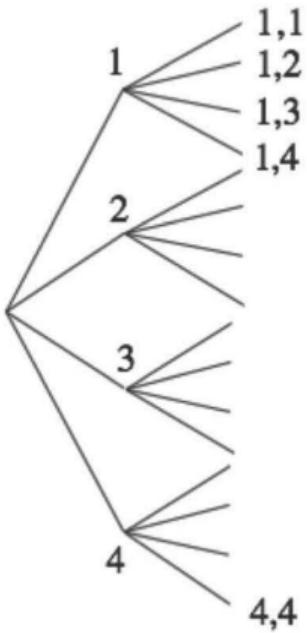
Two rolls of a tetrahedral die





# Sample Space: Discrete/Finite Example

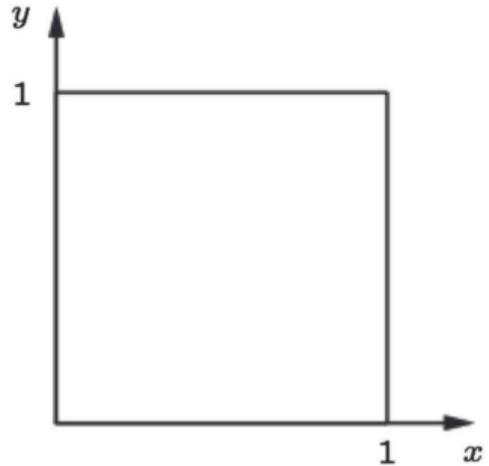
sequential description

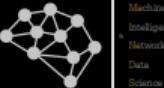




## Sample Space: Continuous Example

- An outcome is a pair  $(x, y)$  such that  $0 \leq x, y \leq 1$ .





# Probability Axioms

- An **event** is a subset of the sample space.
- Probability is assigned to events.

## Axioms

- **Nonnegativity:**  $P(A) \geq 0$  for any event A.
- **Normalization:**  $P(\Omega) = 1$ .
- **(Finite) Additivity:** If  $A \cap B = \emptyset$ , then  $P(A \cup B) = P(A) + P(B)$  (we can extend it by induction as we shall see).

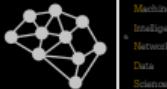
# Some Simple Consequences of the Axioms

## Axioms

- $P(A) \geq 0$
- $P(\Omega) = 1$
- If  $A, B$  are disjoint,  $P(A \cup B) = P(A) + P(B)$ .

## Consequences

- $P(A) \leq 1$
- $P(\emptyset) = 0$
- $P(A) + P(A^c) = 1$
- If  $A, B, C$  are disjoint,  
 $P(A \cup B \cup C) = P(A) + P(B) + P(C)$ .
- $P(\{s_1, \dots, s_k\}) = P(s_1) + \dots + P(s_k)$



# Some Simple Consequences of the Axioms

## Axioms

- $P(A) \geq 0$
- $P(\Omega) = 1$
- If  $A, B$  are disjoint,  $P(A \cup B) = P(A) + P(B)$ .

## Some Simple Consequences of the Axioms

If  $A$ ,  $B$ ,  $C$  are disjoint,  $P(A \cup B \cup C) = P(A) + P(B) + P(C)$ .

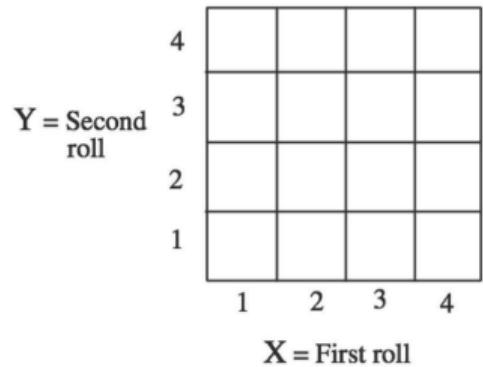
$$P(\{s_1, \dots, s_k\}) =$$

## More Consequences of the Axioms

- If  $A \subset B$ , then  $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cup B) \leq P(A) + P(B)$  (Union Bound)
- $P(A \cup B \cup C) = P(A) + P(A^c \cap B) + P(A^c \cap B^c \cap C)$

# Probability Calculation: Discrete Example

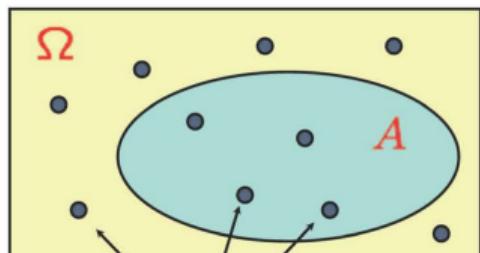
- Two rolls of a tetrahedral die
- Let every possible outcome have probability  $1/16$ .
- $P(X = 1) = ?$
- Let  $Z = \min(X, Y)$ .
- $P(Z = 4) = ?$
- $P(Z = 2) = ?$



# Discrete Uniform Law

If  $\Omega$  consists of  $n$  equally likely elements and event  $A$  consists of  $k$  elements, then:

$$P(A) = \frac{\text{Number of elements in } A}{\text{Total number of elements in } \Omega} = \frac{k}{n}$$

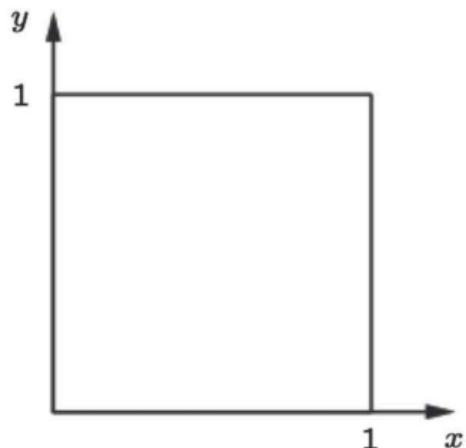


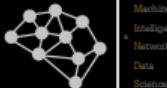
$$\text{prob} = \frac{1}{n}$$

# Probability Calculation: Continuous Example

The sample space is  $(x, y)$  such that  $0 \leq x, y \leq 1$ .

- Uniform probability law on the unit square: Probability = Area
- $P(\{(x, y) | x + y \leq 1/2\}) = ?$
- $P(\{(0.5, 0.3)\}) = ?$



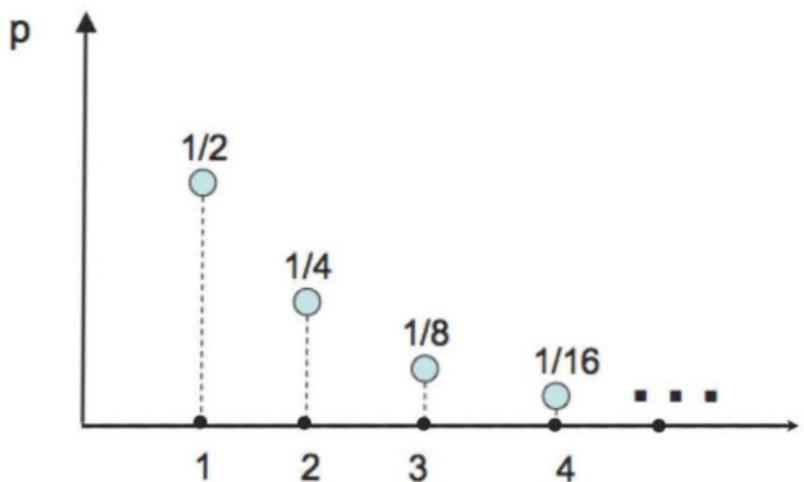


# Probability Calculation Steps

1. Specify the sample space ( $\Omega$ ).
2. Specify a probability law.
3. Identify an event of interest (A).
4. Calculate the probability of the event.

# Probability Calculation: Discrete Infinite Example

- Sample space:  $\{1, 2, 3, \dots\}$
- We are given  $P(n) = \frac{1}{2^n}$  for  $n = 1, 2, \dots$
- What is  $P(\text{outcome is even})$ ?



# The Countable Additivity Axiom

This axiom strengthens the finite additivity axiom.

## Countable Additivity Axiom

If  $A_1, A_2, A_3, \dots$  is an infinite sequence of **disjoint** events, then:

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

# Mathematical Subtleties

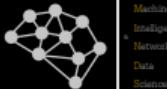
## Countable Additivity Axiom

If  $A_1, A_2, A_3, \dots$  is an infinite sequence of **disjoint** events, then:

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

- The Countable Additivity Axiom holds only for countable sequences of events.
- The unit square (and the real line) is not countable, meaning its elements cannot be arranged in a sequence.
- Area is a legitimate probability law on the unit square, but care must be taken as we cannot assign probabilities/areas to very strange sets.



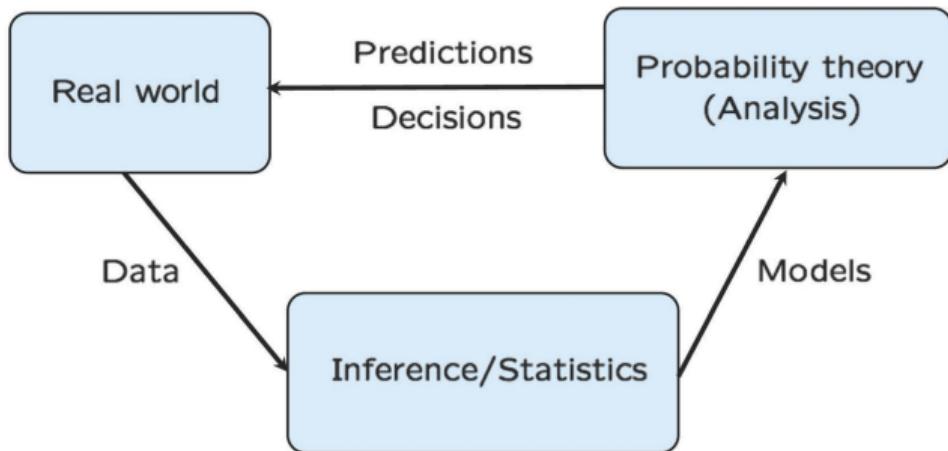
# Interpretations of Probability Theory

- A narrow view sees it as a branch of mathematics built on axioms.
- From axioms we establish theorems
- Are probabilities equivalent to frequencies?
  - $P(\text{coin toss is heads}) = 1/2$  can be seen as a frequency.
  - $P(\text{the president will be reelected}) = 0.7$  is not based on frequency.
- Probabilities are often interpreted as:
  - A description of beliefs
  - Betting preferences

# The Role of Probability Theory

It is a framework for analyzing phenomena with uncertain outcomes.

It provides rules for consistent reasoning, used for making predictions and decisions.



## **Lecture 2: Conditioning and Bayes' Rule**

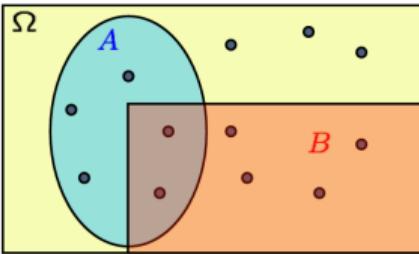
---

# The Idea of Conditioning

Conditional probability allows us to **use new information to revise a model**.

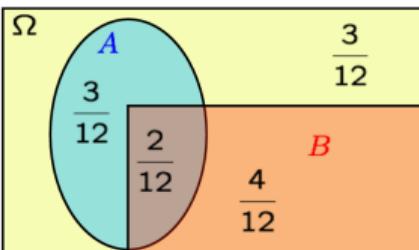
## Original Model

Assume 12 equally likely outcomes in  $\Omega$ .



$$P(A) = \frac{5}{12} \quad P(B) = \frac{6}{12}$$

$$P(A) = 5/12, P(B) = 6/12$$

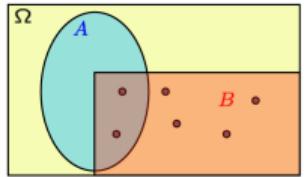


# The Idea of Conditioning

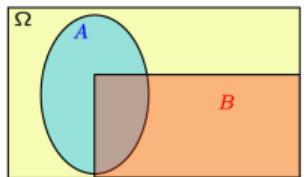
Conditional probability allows us to **use new information to revise a model.**

## After Information

We are told that event B occurred. The sample space effectively shrinks to B, where each of the 6 outcomes is now equally likely with probability 1/6.



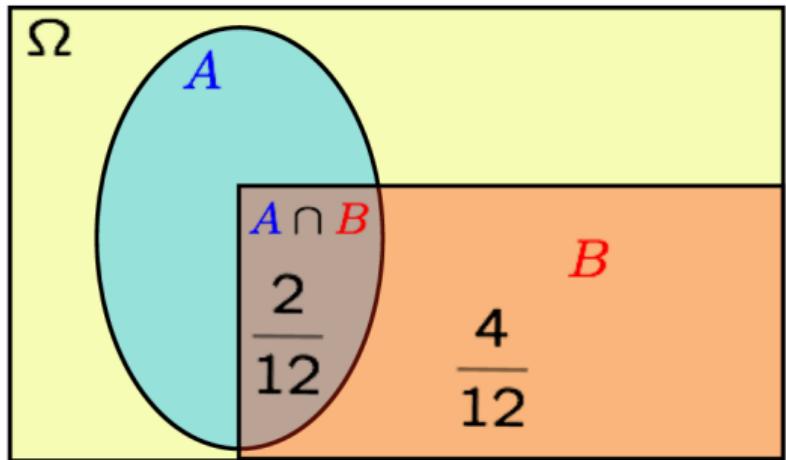
$$P(A | B) = \quad P(B | B) =$$



The new probability of A, given B, is:  $P(A|B) = \frac{2}{6} = \frac{1}{3}$ . Also,  $P(B|B) = 1$ .

# Definition of Conditional Probability

- $P(A|B)$  is the “probability of A, given that B occurred.”



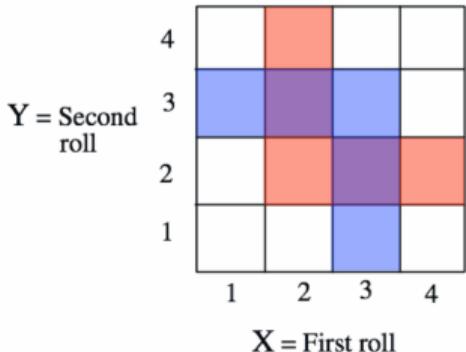
Definition:  $P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}$  This is defined only when  $P(B) > 0$ .

From our previous example:  $P(A|B) = \frac{2/12}{6/12} = \frac{2}{6} = \frac{1}{3}$ .

## Example: Two Rolls of a 4-Sided Die

The sample space consists of 16 equally likely outcomes, each with probability  $1/16$ .

- Let  $B$  be the event:  $\min(X, Y) = 2$ . This is the set  $\{(2, 2), (2, 3), (2, 4), (3, 2), (4, 2)\}$ . So  $P(B) = 5/16$ .
- Let  $M$  be:  $M = \max(X, Y)$ .



- $P(M = 1|B) =$
- $P(M = 3|B) =$

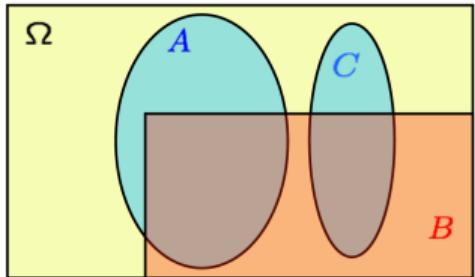
# Properties of Conditional Probability

Conditional probabilities satisfy the axioms of probability on the new universe B. (Assuming  $P(B) > 0$ ).

- **Nonnegativity:**  $P(A|B) \geq 0$ .
- **Normalization:**  $P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$ . Similarly,  $P(B|B) = 1$ .
- **Additivity:** If A and C are disjoint events ( $A \cap C = \emptyset$ ), then  $P(A \cup C|B) = P(A|B) + P(C|B)$

$$P(A \cup C|B) = \frac{P((A \cup C) \cap B)}{P(B)} = \frac{P((A \cap B) \cup (C \cap B))}{P(B)} = \frac{P(A \cap B) + P(C \cap B)}{P(B)}$$

This property extends to countable additivity as well.



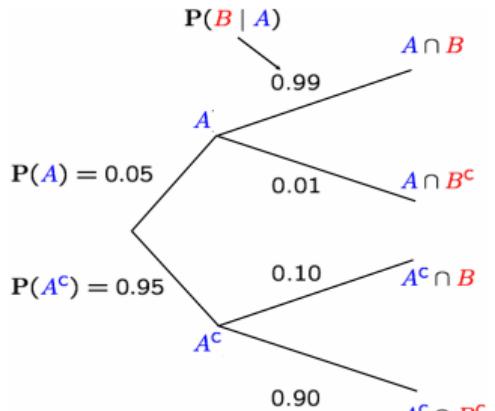
# Models Based on Conditional Probabilities

- Event A: Airplane is flying above.  $P(A) = 0.05$ .
- Event B: Something registers on radar screen.
- Model gives us:  $P(B|A) = 0.99$  (probability of detection) and  $P(B|A^c) = 0.10$  (probability of false alarm)

- $P(A \cap B)$

- $P(B)$

- $P(A|B)$



# The Multiplication Rule

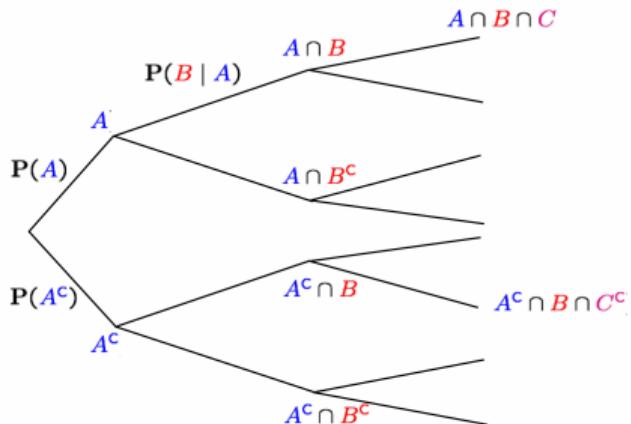
A rearrangement of the definition of conditional probability:

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$$

This can be generalized for a sequence of events:

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \dots P(A_n|A_1 \cap \dots \cap A_{n-1})$$

Example with three events:  $P(A^c \cap B \cap C^c) = P(A^c)P(B|A^c)P(C^c|A^c \cap B)$ .



# Total Probability Theorem

Let  $A_1, A_2, \dots, A_n$  be a partition of the sample space  $\Omega$ .

The probability of an event  $B$  can be found by summing over the partition:

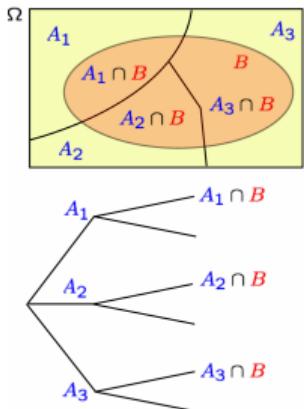
$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \dots + P(B \cap A_n)$$

Using the multiplication rule on each term:

## Total Probability Theorem

$$P(B) = \sum_{i=1}^n P(A_i)P(B|A_i)$$

This is a weighted average of the conditional probabilities  $P(B|A_i)$ , with the weights being the probabilities  $P(A_i)$ .



# Bayes' Rule

Bayes' rule allows us to update our beliefs about a hypothesis  $A_i$  after observing evidence  $B$ .

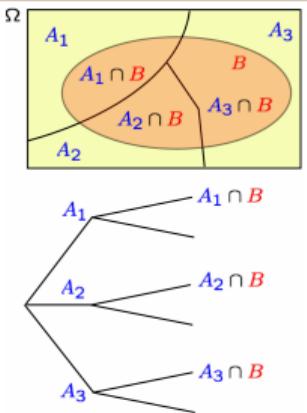
- We have initial beliefs (priors):  $P(A_i)$ .
- We have a model (likelihoods):  $P(B|A_i)$ .
- We want to find the revised beliefs (posterior):  $P(A_i|B)$ .

## Bayes' Rule

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{P(B)}$$

Using the Total Probability Theorem for the denominator:

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^n P(A_j)P(B|A_j)}$$



# Bayes' Rule and Inference

- Attributed to Thomas Bayes (c. 1701-1761), providing a systematic approach for incorporating new evidence.
- At the core of Bayesian inference.

## The Process of Inference

We start with a model of the world and use it to draw conclusions about the causes of an observed event.

$$\text{Cause } A_i \xrightarrow{\text{Model: } P(B|A_i)} \text{Evidence } B$$

$$\text{Evidence } B \xrightarrow{\text{Inference: } P(A_i|B)} \text{Cause } A_i$$

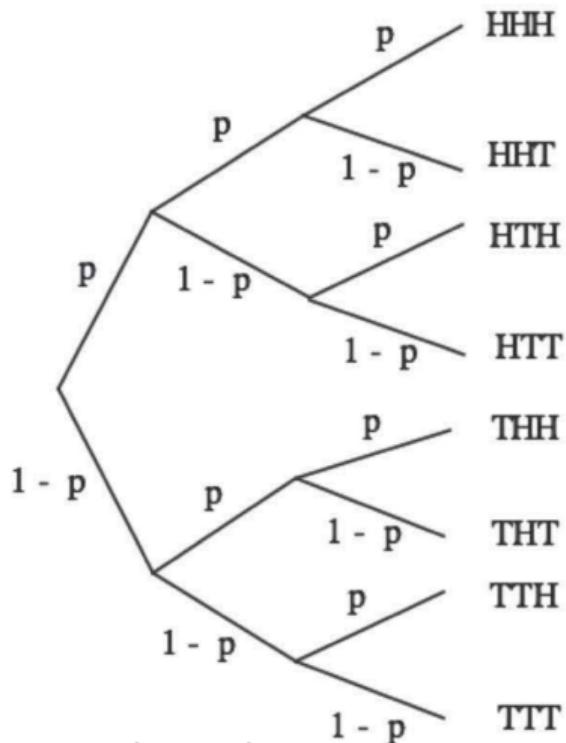
## Lecture 3: Independence

---



# A Model Based on Conditional Probabilities

3 tosses of a biased coin:  $P(H) = p$ ,  $P(T) = 1 - p$



# A Model Based on Conditional Probabilities

- **Multiplication rule:** The probability of a sequence is the product of conditional probabilities along the path. For example:

$$P(THT) = P(T \text{ on 1st}) \cdot P(H|T \text{ on 1st}) \cdot P(T|TH \text{ on 1st, 2nd})$$

this simplifies to:

$$P(THT) =$$

- **Total probability:** To find the probability of exactly one head, we sum the probabilities of all such outcomes:

$$P(1 \text{ head}) = P(HTT) + P(THT) + P(TTH)$$

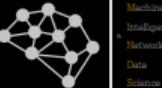
# A Model Based on Conditional Probabilities

- Bayes' rule:

$$P(\text{first toss is H} \mid 1 \text{ head}) = \frac{P(\text{first is H AND 1 head})}{P(1 \text{ head})}$$

The event in the numerator is just the outcome HTT.

$$= \frac{P(HTT)}{P(1 \text{ head})} =$$



# Independence of Two Events

## Intuitive “Definition”

The occurrence of event A provides no new information about the probability of event B.

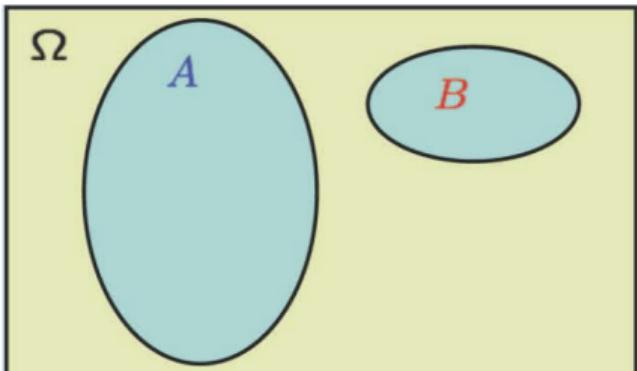
$$P(B|A) = P(B)$$

## Formal Definition

Events A and B are independent if:

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

This definition is symmetric and more general, as it applies even if  $P(A)$  or  $P(B)$  is zero.





# Independence of Event Complements

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

## Intuitive Argument

If learning that A occurred does not change my belief in B, then it should also not change my belief in B not occurring.

## Formal Proof

# Independence of Event Complements

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

## Formal Proof

We want to show that  $P(A \cap B^c) = P(A)P(B^c)$ . We know that the event A can be partitioned into two disjoint parts:  $A = (A \cap B) \cup (A \cap B^c)$ . By additivity:

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Rearranging gives:

$$P(A \cap B^c) = P(A) - P(A \cap B)$$

Since A and B are independent, we can substitute  $P(A \cap B) = P(A)P(B)$ :

$$P(A \cap B^c) = P(A) - P(A)P(B)$$

$$= P(A)(1 - P(B))$$

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

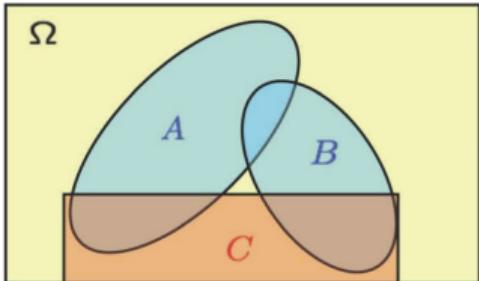
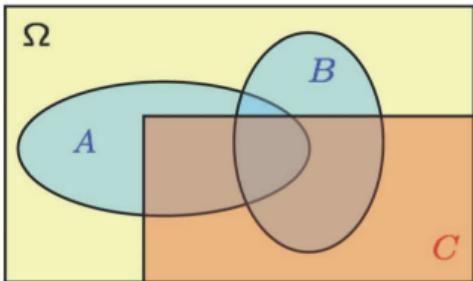


# Conditional Independence

## Definition

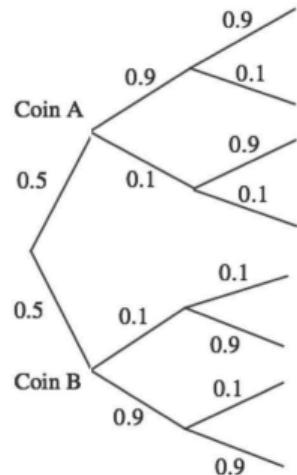
Events A and B are **conditionally independent given C** if they are independent under the probability law  $P(\cdot|C)$ . That is:  $P(A \cap B|C) = P(A|C)P(B|C)$

**Important:** Independence does not imply conditional independence, and conditional independence does not imply independence.



# Conditioning May Affect Independence

We have two unfair coins, A and B, with  $P(H|\text{coin A}) = 0.9$  and  $P(H|\text{coin B}) = 0.1$ . We choose one of the coins with equal probability (0.5) and then toss it repeatedly.



Are the outcomes of the coin tosses independent? Let's compare:

- $P(\text{toss 11} = \text{H}):$
- $P(\text{toss 11} = \text{H} \mid \text{first 10 tosses are heads}):$

# Independence of a Collection of Events

## Definition

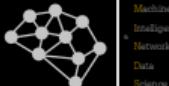
Events  $A_1, A_2, \dots, A_n$  are called **independent** if for any subset of indices  $\{i, j, \dots, m\}$ , the following holds:

$$P(A_i \cap A_j \cap \dots \cap A_m) = P(A_i)P(A_j) \cdots P(A_m)$$

This means the probability of the intersection of any sub-collection of events is the product of their individual probabilities.

For  $n = 3$ , this requires checking 4 conditions:

- $P(A_1 \cap A_2) = P(A_1)P(A_2)$  (Pairwise Independence)
- $P(A_1 \cap A_3) = P(A_1)P(A_3)$  (Pairwise Independence)
- $P(A_2 \cap A_3) = P(A_2)P(A_3)$  (Pairwise Independence)
- $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$



# Independence vs. Pairwise Independence

Consider two independent fair coin tosses. The sample space is  $\{HH, HT, TH, TT\}$ , with each outcome having probability  $1/4$ . Let's define three events:

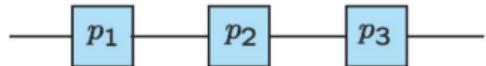
- $H_1$ : First toss is H.  $P(H_1)$
- $H_2$ : Second toss is H.  $P(H_2)$
- $C$ : The two tosses had the same result (HH or TT).  $P(C)$

# Reliability

Independence is a common assumption in reliability analysis of systems with multiple components.

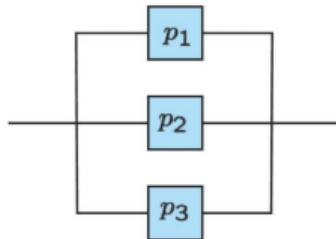
- A system is composed of multiple components (units).
- Let  $p_i$  be the probability that unit  $i$  is “up” (working).
- Assume the units fail independently.

## Series System



The system is up only if all units are up.

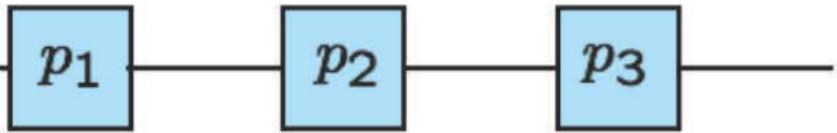
## Parallel System



The system is up if at least one unit is up.

# Reliability

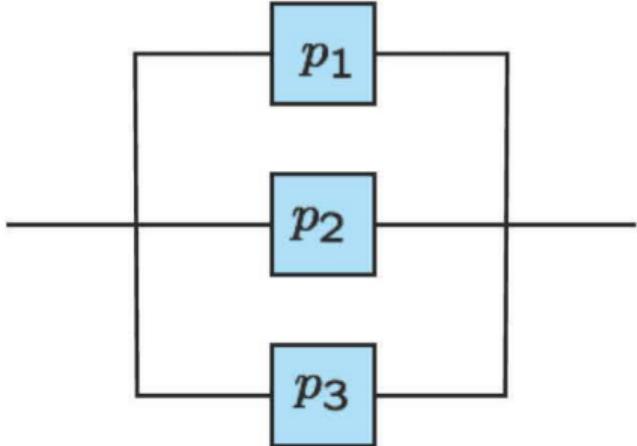
## Series System



The system is up only if all units are up.

$$P(\text{System Up}) =$$

# Reliability



The system is up if at least one unit is up. It is easier to calculate the probability that the system is down (all units fail).

$$P(\text{System Up}) =$$



# The King's Sibling Puzzle

## The Puzzle

The king comes from a family of two children. What is the probability that his sibling is female?

(This is a classic conditional probability problem where defining the sample space correctly is the key.)

## Lecture 4: Counting

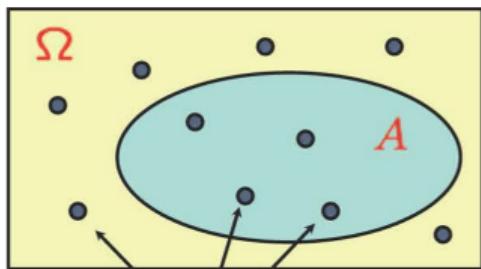
---

# Outline: Counting and the Discrete Uniform Law

Many problems in probability involve a finite sample space where all outcomes are equally likely. In such cases, the probability of an event A is given by the **discrete uniform law**:

$$P(A) = \frac{\text{Number of elements in } A}{\text{Total number of elements in } \Omega} = \frac{k}{n}$$

This reduces probability problems to counting problems.



$$\text{prob} = \frac{1}{n}$$



# Basic Counting Principle

This principle is the foundation for solving most counting problems.

## The Principle

Consider a process that consists of  $r$  sequential stages. If there are:

- $n_1$  choices for the first stage,
- $n_2$  choices for the second stage (for each choice in the first),
- ...
- $n_r$  choices for the  $r$ -th stage (for each combination of choices in previous stages),

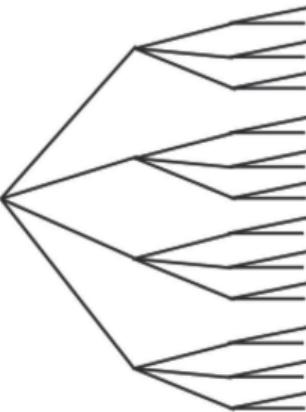
then the total number of possible outcomes is the product  $n_1 \cdot n_2 \cdots n_r$ .

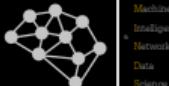


# Basic Counting Principle

## Example

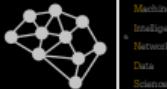
You have 4 shirts, 3 ties, and 2 jackets. How many different attires can you form?





## Basic Counting Principle: Examples

- **License Plates:** How many license plates can be formed with 2 letters followed by 3 digits?
  - If repetition is allowed:
  - If repetition is prohibited:
- **Permutations:** How many ways are there to order  $n$  distinct elements?
  - This is a sequential process of picking elements without replacement or with replacement?
  - Number of orderings =
- **Number of Subsets:** How many subsets can be formed from a set of  $n$  elements,  $\{1, \dots, n\}$ ?
  - For each element, we make a choice:
  - Since there are  $n$  elements, the total number of subsets is



# Example: Probability Calculation

## Problem

Find the probability that six rolls of a fair six-sided die all result in different numbers. (Assume all outcomes are equally likely).

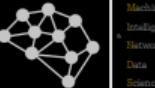
- **Total number of outcomes:**
- **Number of favorable outcomes (A):** For the outcomes to be different, we must choose without repetition or with repetition?

- 
- 
- ...
- 

So,

- **Probability:**

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



# Combinations

## Definition

A combination is a selection of items from a set where the order of selection does not matter.

The notation  $\binom{n}{k}$ , read “n choose k”, represents the number of  $k$ -element subsets of a given  $n$ -element set.

## Derivation of the Formula

We can find the formula for  $\binom{n}{k}$  by counting the number of *ordered* sequences of  $k$  distinct items in two different ways.

1. **Method 1: Choose one item at a time.** The number of ways is

$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}.$$

2. **Method 2: Choose the  $k$  items first, then order them.** First, choose a  $k$ -element subset (there are  $\binom{n}{k}$  ways). Then, order those  $k$  elements (there are  $k!$  ways). The total number of ways is  $\binom{n}{k} \cdot k!$ .

Equating the two expressions:  $\binom{n}{k} \cdot k! = \frac{n!}{(n-k)!}$ . which implies  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

# Properties of the Binomial Coefficient

The binomial coefficient  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$  has several important properties:

- The number of ways to choose all  $n$  items is 1:

$$\binom{n}{n} = \frac{n!}{n!(n-n)!} = \frac{n!}{n!0!} = 1$$

- The number of ways to choose 0 items is 1 (the empty set):

$$\binom{n}{0} = \frac{n!}{0!(n-0)!} = \frac{n!}{0!n!} = 1$$

- The total number of subsets of a set of size  $n$  is the sum of the number of subsets of each possible size:

$$\sum_{k=0}^n \binom{n}{k} = 2^n$$

# Binomial Probabilities

The binomial coefficient is the key to calculating probabilities in experiments involving a fixed number of independent trials.

## Scenario

Consider  $n$  independent coin tosses, where the probability of heads is  $P(H) = p$ .

- The probability of any *particular* sequence of outcomes is found using independence. For example:

$$P(HTTHHH) = P(H)P(T)P(T)P(H)P(H)P(H) = p(1-p)(1-p)p^3 = p^4(1-p)^2$$

- The probability of any *particular* sequence that contains exactly  $k$  heads and  $n - k$  tails is  $p^k(1-p)^{n-k}$ .
- To find the total probability of getting exactly  $k$  heads, we multiply the probability of one such sequence by the number of such sequences. The number of ways to arrange  $k$  heads in  $n$  positions is  $\binom{n}{k}$ .

# Binomial Probabilities

- The probability of any *particular* sequence of outcomes is found using independence. For example:

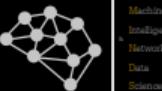
$$P(HTTHHH) = P(H)P(T)P(T)P(H)P(H)P(H) = p(1-p)(1-p)ppp = p^4(1-p)^2$$

- The probability of any *particular* sequence that contains exactly  $k$  heads and  $n - k$  tails is  $p^k(1-p)^{n-k}$ .
- To find the total probability of getting exactly  $k$  heads, we multiply the probability of one such sequence by the number of such sequences. The number of ways to arrange  $k$  heads in  $n$  positions is  $\binom{n}{k}$ .

## Binomial Probability Formula

The probability of getting exactly  $k$  heads in  $n$  independent tosses is:

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



# A Coin Tossing Problem

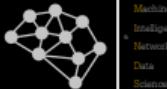
## Problem

Given that there were exactly 3 heads in 10 independent tosses of a coin with  $P(H) = p$ , what is the probability that the first two tosses were heads?

Let A be the event that the first two tosses are heads. Let B be the event that there are exactly 3 heads in 10 tosses. We want to find  $P(A|B)$ .

### First Solution: Using the Definition

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



# A Coin Tossing Problem (Continued)

## Problem

Given that there were exactly 3 heads in 10 independent tosses, what is the probability that the first two tosses were heads?

### Second Solution: Using a Reduced Sample Space

The conditioning event B tells us we are in a world where exactly 3 heads occurred. All sequences with 3 heads are equally likely, regardless of the value of  $p$  (since the term  $p^3(1 - p)^7$  is common to all of them). So, we can use a new uniform probability law on the set of outcomes in B.

# Partitions

We now consider the problem of dividing a set of  $n$  distinct items among  $r$  different people, giving  $n_i$  items to person  $i$ , where  $n_1 + n_2 + \dots + n_r = n$ .

## Derivation

Imagine ordering all  $n$  items in a line ( $n!$  ways).

- Give the first  $n_1$  items to person 1.
- Give the next  $n_2$  items to person 2.
- ... and so on.

This process overcounts the number of unique partitions. For person  $i$ , the  $n_i!$  different orderings of the items they receive all result in the same partition. To correct for this, we divide by the number of permutations within each group.

## Multinomial Coefficient

The number of ways to partition  $n$  distinct items into  $r$  groups of specified sizes  $n_1, \dots, n_r$  is given by the multinomial coefficient:

$$\binom{n}{n_1, n_2, \dots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}$$

# Example: Card Dealing

## Problem

A 52-card deck is dealt fairly to four players (13 cards each). Find the probability that each player gets exactly one ace.

## Solution using Partitions

- **Total outcomes:** The number of ways to partition 52 cards into four hands of 13 is the multinomial coefficient:

$$|\Omega| = \binom{52}{13, 13, 13, 13} = \frac{52!}{13!13!13!13!}$$

- **Favorable outcomes (A):** We construct a favorable outcome in stages.

1. Distribute the 4 aces: 4 choices for Player 1, 3 for P2, 2 for P3, 1 for P4. Total ways:  $4!$ .
2. Distribute the remaining 48 non-ace cards: Each player needs 12 more cards. The number of ways to partition the 48 cards into four hands of 12 is:  $\binom{48}{12, 12, 12, 12} = \frac{48!}{12!12!12!12!}$ .

By the counting principle,  $|A| = 4! \cdot \frac{48!}{12!12!12!12!}$ .

- **Probability:**

$$P(A) = \frac{|A|}{|\Omega|} = \frac{4! \cdot \frac{48!}{(12!)^4}}{\frac{52!}{(13!)^4}} = \frac{24 \cdot 48! \cdot (13!)^4}{52! \cdot (12!)^4} = \frac{24 \cdot 13^4}{52 \cdot 51 \cdot 50 \cdot 49} \approx 0.1055$$

# Example: Card Dealing (A Smarter Solution?)

## Problem

Find the probability that each of four players gets an ace.

## Alternative Approach

Focus only on the locations of the four aces within the 52 slots of the deck. By symmetry, any set of four positions for the aces is equally likely.

Let's consider the second ace. Where can it go?

- There are 51 available slots for the second ace.
- For it to go to a different player than the first ace, it must land in one of the 39 slots belonging to the other three players (13 slots per player).
- The probability that the second ace goes to a different player is  $\frac{39}{51}$ .

## Example: Card Dealing (A Smarter Solution?)

Now consider the third ace.

- There are 50 available slots.
- For it to go to a different player than the first two, it must land in one of the 26 slots belonging to the remaining two players.
- The probability is  $\frac{26}{50}$ .

Finally, the fourth ace.

- There are 49 available slots.
- It must go to the one remaining player, who has 13 slots.
- The probability is  $\frac{13}{49}$ .

The total probability is the product:

$$P(\text{each player gets an ace}) = \frac{39}{51} \cdot \frac{26}{50} \cdot \frac{13}{49} \approx 0.1055$$

# Partitions: The Multinomial Distribution

The binomial distribution describes the probability of  $k$  successes in  $n$  trials. The multinomial distribution generalizes this to scenarios with more than two possible outcomes.

## Example: Building a Circuit

A large bin contains thousands of circuit components. The proportions are:

- 50% Resistors (R), so  $p_R = 0.5$
- 30% Capacitors (C), so  $p_C = 0.3$
- 20% Inductors (L), so  $p_L = 0.2$

We draw  $n = 10$  components independently. What is the probability of drawing exactly 5 Resistors, 3 Capacitors, and 2 Inductors?

# Partitions: The Multinomial Distribution

## Step 1: Probability of One Sequence

The probability of any *one specific sequence*, e.g., RRRRRCCCLL, is found by multiplying the probabilities due to independence:

$$\begin{aligned} P(\text{RRRRRCCCLL}) &= p_R^5 p_C^3 p_L^2 \\ &= (0.5)^5 (0.3)^3 (0.2)^2 \end{aligned}$$

## Step 2: Count All Sequences

We need to count how many distinct sequences contain 5 R's, 3 C's, and 2 L's. This is a partition problem, solved with the multinomial coefficient:

$$\binom{10}{5, 3, 2} = \frac{10!}{5!3!2!} = 2520$$

## The Multinomial Distribution

The probability of getting  $n_1$  of type 1,  $n_2$  of type 2, ...,  $n_k$  of type k in  $n$  trials is:

$$P(N_1 = n_1, \dots, N_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

# **Lecture 5: Discrete Random Variables Part I Probability Mass Functions**

---



# Random variables: the idea

# Random Variables: The Formalism

## Definition

A **random variable** (r.v.) is a function that associates a numerical value to every possible outcome in the sample space  $\Omega$ .

$$X : \Omega \rightarrow \mathbb{R}$$

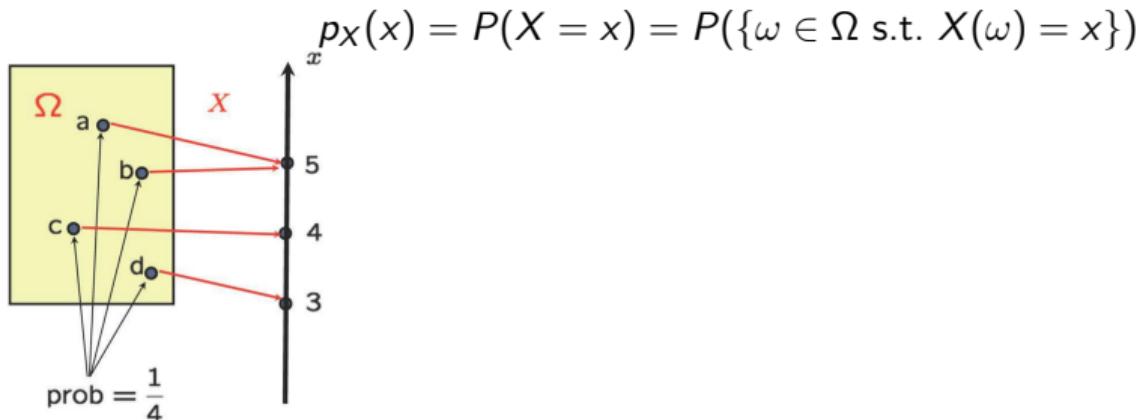
- A random variable can take on discrete or continuous values.
- We use uppercase letters (e.g.,  $X$ ) to denote a random variable, and lowercase letters (e.g.,  $x$ ) to denote a specific numerical value it can take.
- We can define multiple random variables on the same sample space.
- A function of one or several random variables is also a random variable. For example, if  $X$  and  $Y$  are random variables, then  $Z = X + Y$  is a new random variable where for any outcome  $\omega \in \Omega$ , its value is  $Z(\omega) = X(\omega) + Y(\omega)$ .

# Probability Mass Function (PMF)

The PMF describes the probability distribution of a discrete random variable.

## Definition

The **probability mass function (PMF)** of a discrete random variable  $X$  is a function  $p_X(x)$  that gives the probability that  $X$  is equal to a specific value  $x$ .



## Properties of a PMF

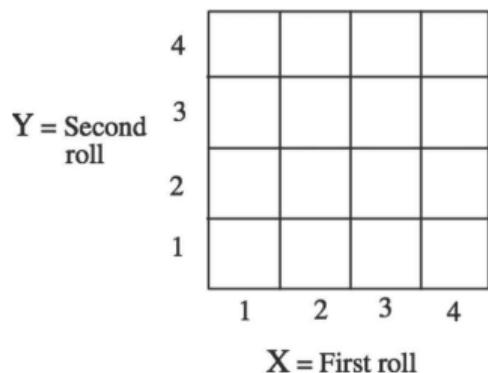
- $p_X(x) \geq 0$  for all  $x$ .
- $\sum_x p_X(x) = 1$ , where the sum is over all possible values  $x$  that  $X$  can take.

## Two Rolls of a Tetrahedral Die

The sample space has 16 equally likely outcomes, each with probability 1/16. Let the random variable be the sum of the two rolls,  $Z = X + Y$ .

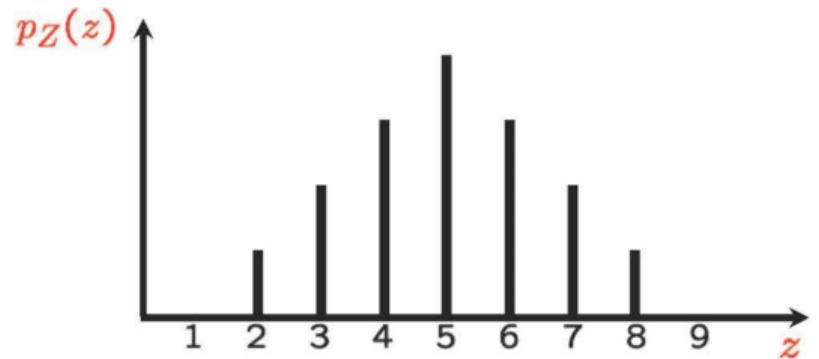
To find the PMF of  $Z$ ,  $p_Z(z)$ , we:

1. Identify all possible values  $z$  for the sum (from 2 to 8).
2. For each value  $z$ , find all outcomes  $(X, Y)$  such that  $X + Y = z$ .
3. Add the probabilities of these outcomes.



# PMF Calculation Example

- $p_Z(2) = P(Z = 2) = P(\{(1, 1)\}) = 1/16.$
- $p_Z(3) = P(Z = 3) = P(\{(1, 2), (2, 1)\}) = 2/16.$
- $p_Z(4) = P(Z = 4) = P(\{(1, 3), (2, 2), (3, 1)\}) = 3/16.$
- $p_Z(5) = P(Z = 5) = P(\{(1, 4), (2, 3), (3, 2), (4, 1)\}) = 4/16.$
- $p_Z(6) = P(Z = 6) = P(\{(2, 4), (3, 3), (4, 2)\}) = 3/16.$
- $p_Z(7) = P(Z = 7) = P(\{(3, 4), (4, 3)\}) = 2/16.$
- $p_Z(8) = P(Z = 8) = P(\{(4, 4)\}) = 1/16.$



# Common Discrete Random Variables: Bernoulli

## Bernoulli Random Variable

Models a single trial with two outcomes: success or failure.

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

An important special case is the **indicator random variable** for an event A, denoted  $I_A$ , which is 1 if A occurs and 0 otherwise.  $P(I_A = 1) = P(A)$ .

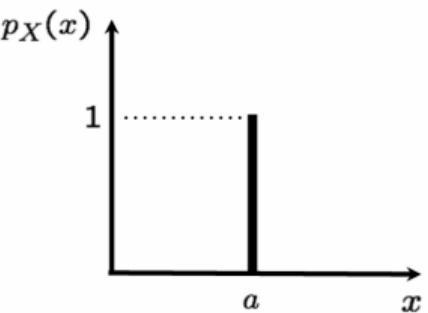
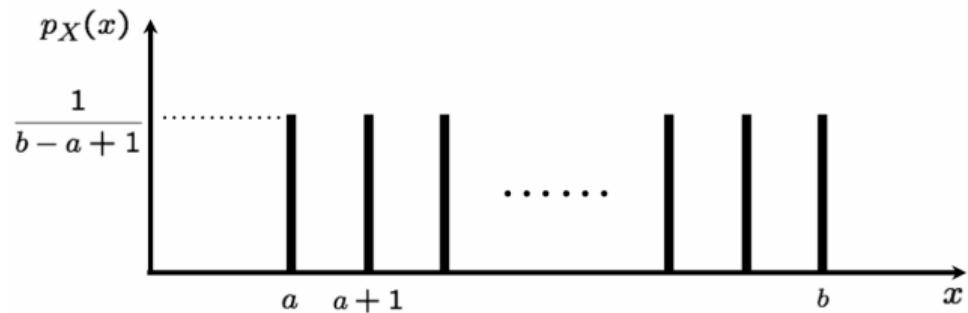
# Common Discrete Random Variables: Discrete Uniform

## Discrete Uniform Random Variable

Models a situation where one of a set of integers  $\{a, a+1, \dots, b\}$  is chosen, with all choices being equally likely.

$$p_X(k) = \frac{1}{b-a+1}, \quad \text{for } k = a, a+1, \dots, b$$

If  $a = b$ , this becomes a deterministic (constant) random variable.

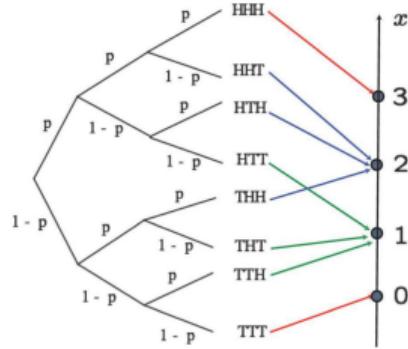


# Binomial Random Variable

- Parameters: a positive integer  $n$  and a probability  $p \in [0, 1]$ .
- Experiment:  $n$  independent tosses of a coin with  $P(\text{Heads}) = p$ .
- Random Variable  $X$ : The number of Heads observed.

The PMF is given by the binomial probability formula:

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for } k = 0, 1, \dots, n$$

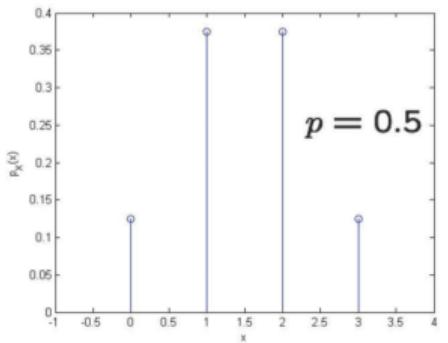


The binomial distribution models # successes in a fixed number of independent trials.

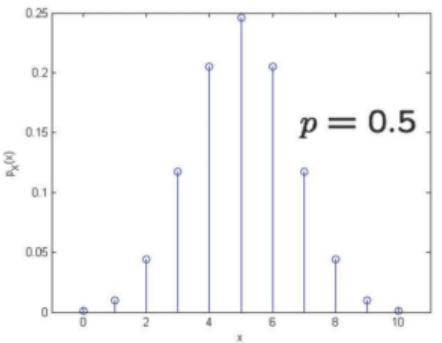


# Binomial Random Variable

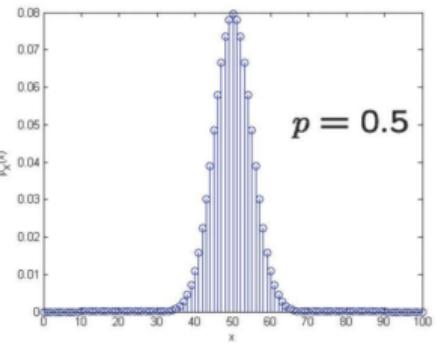
$n = 3$



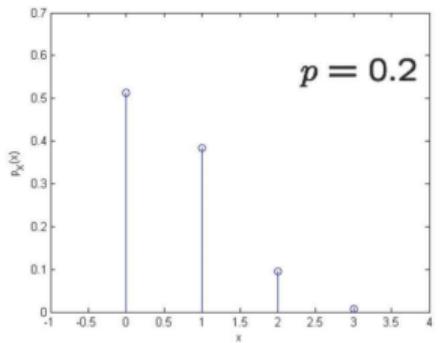
$n = 10$



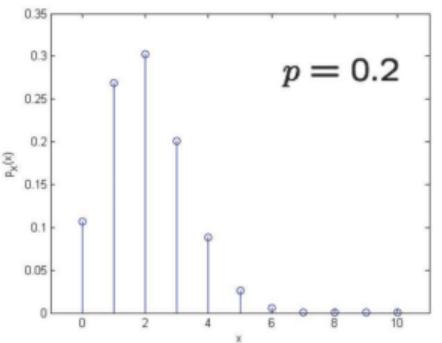
$n = 100$



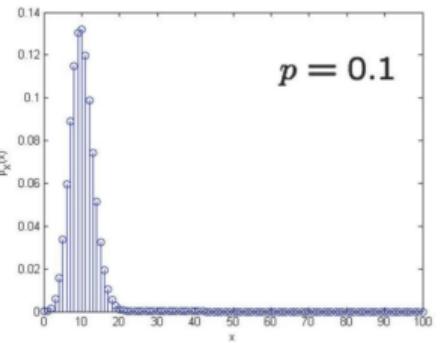
$p = 0.2$



$p = 0.2$



$p = 0.1$



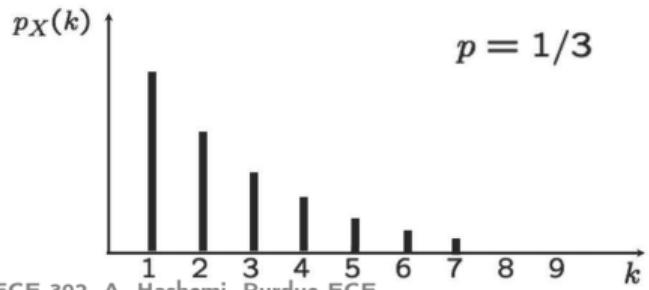
# Geometric Random Variable

- Parameter: a probability  $p$  with  $0 < p \leq 1$ .
- Experiment: A sequence of independent tosses of a coin with  $P(\text{Heads}) = p$ .
- Random Variable  $X$ : The number of tosses until the first Head is observed.

The event  $\{X = k\}$  corresponds to the sequence of  $k - 1$  tails followed by one head:  
 $TT \dots TH$ . The PMF is:

$$p_X(k) = (1 - p)^{k-1} p, \quad \text{for } k = 1, 2, 3, \dots$$

The geometric distribution is a model for waiting times or the number of trials until the first success.



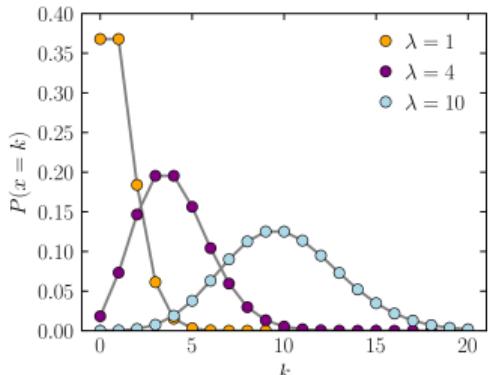
# Poisson Random Variable

- Parameter:  $\lambda > 0$ , which is the average number of events in a given interval of time or space.
- Random Variable  $X$ : The number of events that occur in the interval.

The Probability Mass Function (PMF) is given by:

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

Poisson distribution models the number of occurrences of a rare event within a fixed period.  
 It's often used as an approximation for the Binomial distribution when  $n$  is large and  $p$  is small.



# **Lecture 6: Discrete Random Variables Part II**

## **Expectation; Variance; Conditioning**

---

# Expectation (Mean) of a Random Variable

The expectation is a weighted average of the possible values of a random variable, where the weights are the probabilities.

## Definition

The **expected value** (or expectation, or mean) of a discrete random variable  $X$  is:

$$E[X] = \sum_x x \cdot p_X(x)$$

The sum is taken over all possible values  $x$  of  $X$ .

- **Interpretation:** The expectation is the long-run average value of the random variable over many independent repetitions of the experiment.
- **Caution:** For the expectation to be well-defined, the sum must converge absolutely:  
 $\sum_x |x| p_X(x) < \infty$ .

# Calculating Expectations

- **Bernoulli( $p$ ):**

$$E[X] =$$

If  $X$  is an indicator variable for event A,  $X = I_A$ , then  $E[I_A] = P(A)$ .

# Calculating Expectations

Discrete Uniform on  $\{0, 1, \dots, n\}$ :

$$E[X] =$$





# Elementary properties of Expectation

# The Expected Value Rule

How do we calculate the expectation of a function of a random variable, say  $Y = g(X)$ ?

## Method 1: Using the PMF of Y

Find the PMF  $p_Y(y)$  and then use the definition:  $E[Y] = \sum_y y \cdot p_Y(y)$ . This can be tedious.

## Method 2: The Expected Value Rule

It is much easier to average over the values of  $X$ :

$$E[Y] = E[g(X)] = \sum_x g(x)p_X(x)$$

**Caution:** In general,  $E[g(X)] \neq g(E[X])$ .

### Example

Let  $X$  have PMF  $p_X(2) = 0.1, p_X(3) = 0.2, p_X(4) = 0.3, p_X(5) = 0.4$ . Find  $E[X^2]$ .

# Linearity of Expectation

A fundamental property of expectation is linearity.

## Property

For any random variable  $X$  and any constants  $a$  and  $b$ :

$$E[aX + b] = aE[X] + b$$

## Derivation

We use the expected value rule with  $g(X) = aX + b$ .

# Variance

While the expectation summarizes the center of a PMF, the **variance** measures its spread or dispersion.

## Definition

For a random variable  $X$  with mean  $\mu = E[X]$ , the variance is the expected value of the squared distance from the mean:

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

- Using the expected value rule, we can write this as:

$$\text{var}(X) = \sum_x (x - \mu)^2 p_X(x)$$

- The **standard deviation**,  $\sigma_X$ , is the square root of the variance. It has the same units as  $X$ .

$$\sigma_X = \sqrt{\text{var}(X)}$$

# Properties of Variance

- **Shifting:** Adding a constant  $b$  to  $X$  shifts its mean but does not change its spread.

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

- **Scaling:** Multiplying by a constant  $a$  scales the variance by  $a^2$ .

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

## General Property

$$\text{var}(aX + b) = a^2 \text{var}(X)$$

### A Useful Computational Formula

A more convenient way to calculate variance is  $\text{var}(X) = E[X^2] - (E[X])^2$ .

# Variance of a Bernoulli Random Variable

Let  $X$  be a Bernoulli random variable with parameter  $p$ .

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

We know  $E[X] = p$ . To find the variance, we first find  $E[X^2]$ .

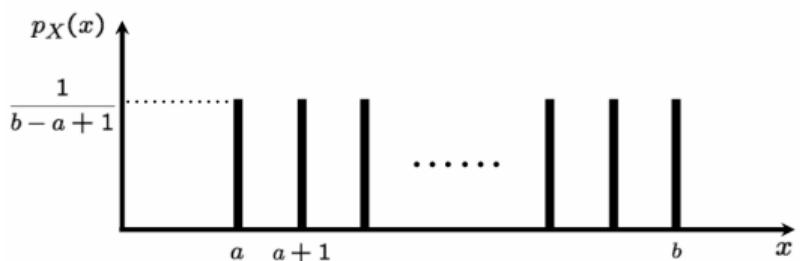
$$E[X^2] =$$

Now, using the computational formula:

$$\text{var}(X) =$$



# Variance of a Uniform Random Variable



# Conditioning a Random Variable on an Event

Just as we can have conditional probabilities, we can also define a conditional PMF and conditional expectation for a random variable, given that an event A has occurred.

## Unconditional

$$\text{PMF: } p_X(x) = P(X = x)$$

$$\sum_x p_X(x) = 1$$

$$\text{Expectation: } E[X] = \sum_x x p_X(x)$$

$$E[g(X)] = \sum_x g(x) p_X(x)$$

## Conditional on A

$$\text{PMF: } p_{X|A}(x) = P(X = x|A)$$

$$\sum_x p_{X|A}(x) = 1$$

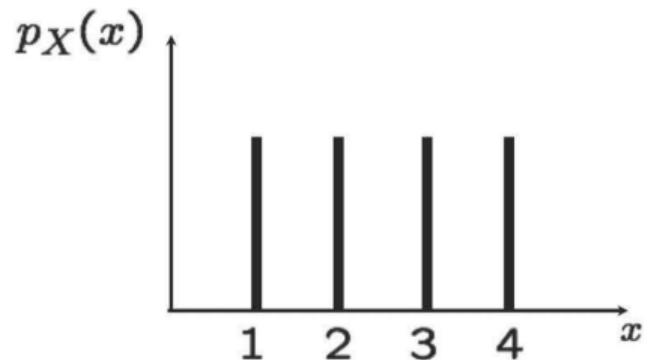
$$\text{Expectation: } E[X|A] = \sum_x x p_{X|A}(x)$$

$$E[g(X)|A] = \sum_x g(x) p_{X|A}(x)$$

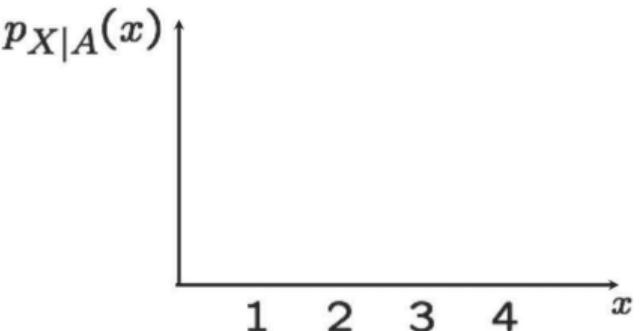
# Example of Conditioning

Let  $X$  be a discrete uniform random variable on  $\{1, 2, 3, 4\}$ . So,  $p_X(x) = 1/4$  for  $x \in \{1, 2, 3, 4\}$ . Let the conditioning event be  $A = \{X \geq 2\}$ .

**Unconditional PMF**



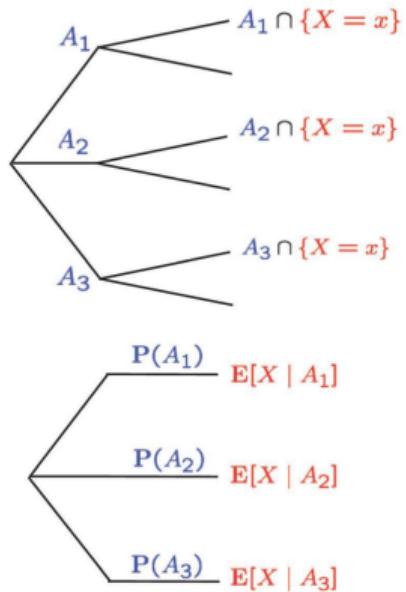
**Conditional PMF given A**



# The Total Expectation Theorem

Similar to the total probability theorem, this theorem provides a way to find the overall expectation by averaging conditional expectations over a partition of the sample space.

Let  $A_1, \dots, A_n$  be a partition of  $\Omega$ .



# The Total Expectation Theorem

Similar to the total probability theorem, this theorem provides a way to find the overall expectation by averaging conditional expectations over a partition of the sample space.

Let  $A_1, \dots, A_n$  be a partition of  $\Omega$ .

## Derivations

- Total Probability for a PMF:

$$p_X(x) = \sum_{i=1}^n P(A_i)p_{X|A_i}(x)$$

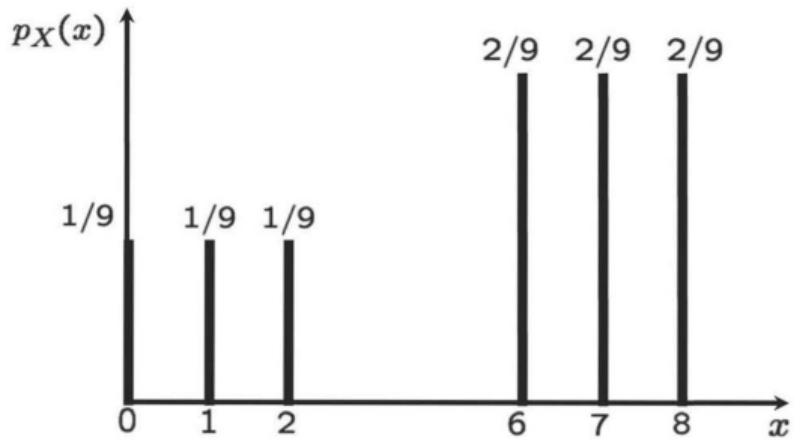
- Total Expectation Theorem:

$$E[X] = \sum_x x p_X(x) = \sum_x x \sum_{i=1}^n P(A_i)p_{X|A_i}(x) = \sum_{i=1}^n P(A_i) \left( \sum_x x p_{X|A_i}(x) \right) = \sum_{i=1}^n P(A_i) E[X|A_i]$$

## Total Expectation Theorem

$$E[X] = P(A_1)E[X|A_1] + \cdots + P(A_n)E[X|A_n]$$

# The Total Expectation Example



# The Geometric PMF: Memorylessness

Let  $X$  be a geometric random variable with parameter  $p$ . It represents the number of independent trials until the first success.

$$p_X(k) = (1 - p)^{k-1} p, \quad k = 1, 2, \dots$$

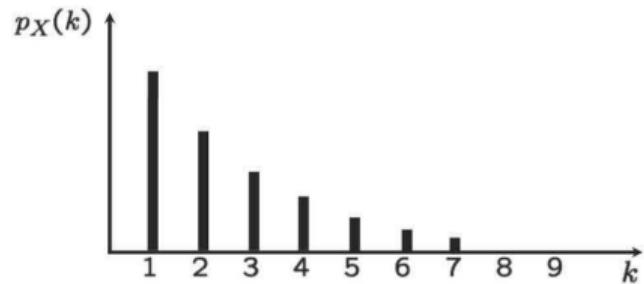
## Memorylessness Property

Given that the first trial was a failure (event  $A = \{X > 1\}$ ), the number of *additional* trials needed,  $X - 1$ , follows the same geometric distribution as  $X$ .

$$P(X - 1 = k | X > 1) = P(X = k) \quad \text{for } k = 1, 2, \dots$$

This means the process “forgets” the past failures and resets. In general, for any  $n$ :

$$P(X - n = k | X > n) = P(X = k)$$



# The Mean of the Geometric PMF

We can find the mean of the geometric distribution,  $E[X]$ , using the total expectation theorem and memorylessness.

Let's partition the sample space by the outcome of the first toss:  $A_1 = \{\text{First toss is H}\}$ ,  $A_2 = \{\text{First toss is T}\}$ .  $P(A_1) = p$ ,  $P(A_2) = 1 - p$ .

# Lecture 7: Discrete Random Variables Part III

## Multiple Random Variables; Conditioning on a Random Variable; Independence of r.v.'s

---

# Multiple Random Variables and Joint PMFs

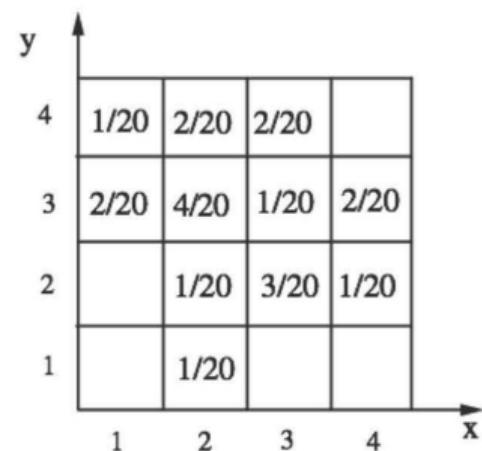
The **joint PMF** of two discrete random variables X and Y describes the probability of them taking on specific values simultaneously.

$$p_{X,Y}(x,y) = P(X = x, Y = y)$$

From the joint PMF, we can recover the individual PMFs, called **marginal PMFs**, by summing over the other variable.

$$p_X(x) = \sum_y p_{X,Y}(x,y) \quad (\text{Summing down columns})$$

$$p_Y(y) = \sum_x p_{X,Y}(x,y) \quad (\text{Summing across rows})$$



# Functions of Multiple Random Variables

Let  $Z = g(X, Y)$  be a function of two random variables.

## Expected Value Rule for Multiple Variables

The expectation of  $Z$  can be computed using the joint PMF of  $X$  and  $Y$ , without needing to find the PMF of  $Z$  first.

$$E[Z] = E[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y)$$

## Linearity of Expectations

The most important consequence is the linearity of expectations, which holds for any random variables  $X$  and  $Y$ , regardless of whether they are independent.

$$E[X + Y] = E[X] + E[Y]$$

This generalizes to any number of random variables and any linear combination:

$$E[a_1 X_1 + \cdots + a_n X_n] = a_1 E[X_1] + \cdots + a_n E[X_n]$$

# The Mean of the Binomial Distribution

Finding the mean of a binomial r.v.  $X \sim \text{Binomial}(n, p)$  using the definition is algebraically intensive:

$$E[X] = \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k}$$

## A Simpler Way using Linearity

A binomial r.v.  $X$  can be seen as the sum of  $n$  independent Bernoulli random variables, where  $X_i$  is an indicator for success on the  $i$ -th trial.

$$X = X_1 + X_2 + \cdots + X_n$$

Using the linearity of expectations:

$$E[X] = E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n]$$

Since each  $X_i$  is a Bernoulli trial with parameter  $p$ , we know  $E[X_i] = p$ .

$$E[X] = p + p + \cdots + p \quad (n \text{ times}) = np$$

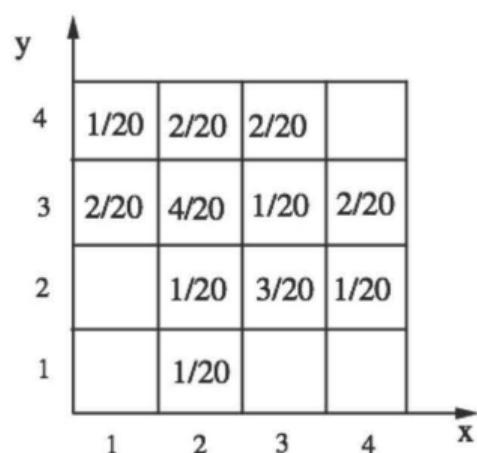
# Conditional PMFs

The **conditional PMF** of  $X$  given  $Y = y$  is:

$$p_{X|Y}(x|y) = P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

This is defined for all  $y$  such that  $p_Y(y) > 0$ . For a fixed  $y$ , the conditional PMF is a valid PMF for  $X$ , meaning  $\sum_x p_{X|Y}(x|y) = 1$ . This leads to a version of the multiplication rule for PMFs:

$$p_{X,Y}(x,y) = p_Y(y)p_{X|Y}(x|y) = p_X(x)p_{Y|X}(y|x)$$



# Conditional PMFs with Multiple Variables

The notation and concepts extend naturally to more than two random variables.

- $p_{X|Y,Z}(x|y,z) = P(X = x | Y = y, Z = z)$
- $p_{X,Y|Z}(x,y|z) = P(X = x, Y = y | Z = z)$

## Multiplication Rule for Random Variables

Similar to the rule for events, we can chain conditional PMFs to find a joint PMF:

$$p_{X,Y,Z}(x,y,z) = p_X(x)p_{Y|X}(y|x)p_{Z|X,Y}(z|x,y)$$

# Conditional Expectation

The concept of expectation can also be conditioned on an event or on the value of another random variable.

## Definition

The **conditional expectation** of  $X$  given  $Y = y$  is the expectation computed using the conditional PMF:

$$E[X|Y = y] = \sum_x x \cdot p_{X|Y}(x|y)$$

The expected value rule also has a conditional version:

$$E[g(X)|Y = y] = \sum_x g(x)p_{X|Y}(x|y)$$

# Total Probability and Expectation Theorems

The total probability and total expectation theorems can be stated in terms of random variables.

Let the values  $y$  of a random variable  $Y$  form a partition of the sample space.

## Total Probability Theorem for PMFs

$$p_X(x) = \sum_y P(Y = y)P(X = x|Y = y) = \sum_y p_Y(y)p_{X|Y}(x|y)$$

## Total Expectation Theorem

The unconditional expectation is the weighted average of the conditional expectations.

$$E[X] = \sum_y p_Y(y)E[X|Y = y]$$

This is also written as  $E[X] = E[E[X|Y]]$ .

# Independence of Random Variables

## Definition

Two random variables  $X$  and  $Y$  are **independent** if their joint PMF is the product of their marginal PMFs for all pairs of values  $(x, y)$ .

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \quad \text{for all } x, y$$

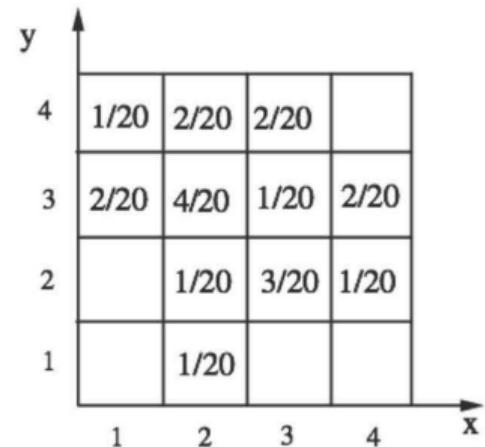
This is equivalent to the condition  $P(X = x, Y = y) = P(X = x)P(Y = y)$  for all  $x, y$ .

For a collection of random variables  $X, Y, Z, \dots$  to be independent, their joint PMF must factor into the product of all their marginal PMFs.

$$p_{X,Y,Z}(x,y,z) = p_X(x)p_Y(y)p_Z(z), \quad \text{for all } x, y, z$$



## Example: Independence



Are  $X$  and  $Y$  independent? how about conditioned on  $X \leq 2$  and  $Y \geq 3$ ?

# Independence and Expectations

While  $E[X + Y] = E[X] + E[Y]$  is always true, the expectation of a product is not always the product of expectations.

## Expectation of a Product

If  $X$  and  $Y$  are **independent** random variables, then:

$$E[XY] = E[X]E[Y]$$

More generally, for any functions  $g$  and  $h$ , if  $X$  and  $Y$  are independent, then  $g(X)$  and  $h(Y)$  are also independent, and:

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

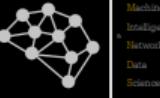
# Independence and Variances

The variance of a sum is not always the sum of the variances.

In general:

$$\begin{aligned}\text{var}(X + Y) &= E[(X + Y)^2] - (E[X + Y])^2 \\&= E[X^2 + 2XY + Y^2] - (E[X] + E[Y])^2 \\&= E[X^2] + 2E[XY] + E[Y^2] - (E[X]^2 + 2E[X]E[Y] + E[Y]^2) \\&= (\text{var}(X) + E[X]^2) + 2E[XY] + (\text{var}(Y) + E[Y]^2) - (E[X]^2 + 2E[X]E[Y] + E[Y]^2) \\&= \text{var}(X) + \text{var}(Y) + 2(E[XY] - E[X]E[Y])\end{aligned}$$

The term  $E[XY] - E[X]E[Y]$  is the covariance, which is zero for independent variables.



# Variance of a Sum

If  $X$  and  $Y$  are **independent** random variables, then:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$$

If  $X$  and  $Y$  are independent, then

$$\text{var}(X - Y) = \text{var}(X) + \text{var}(-Y) = \text{var}(X) + (-1)^2 \text{var}(Y) = \text{var}(X) + \text{var}(Y).$$

# Variance of the Binomial

Let  $X \sim \text{Binomial}(n, p)$ . We can find its variance using a simple trick.

We represent  $X$  as a sum of  $n$  independent Bernoulli indicator variables:

$$X = X_1 + X_2 + \cdots + X_n$$

where  $X_i = 1$  if the  $i$ -th trial is a success.



# The Hat Problem: Mean and Variance

$n$  people throw their hats in a box and pick one at random. Let  $X$  be the number of people who get their own hat back. Find  $E[X]$  and  $\text{var}(X)$ .

Let  $X_i$  be an indicator variable for the  $i$ -th person getting their own hat back.

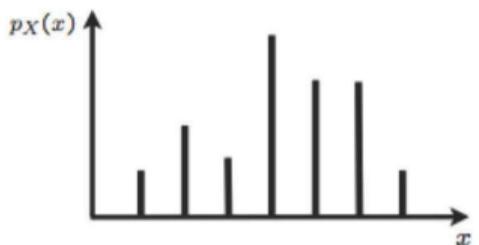
$$X = X_1 + X_2 + \cdots + X_n$$

# **Lecture 8: Continuous Random Variables Part I Probability Density Functions**

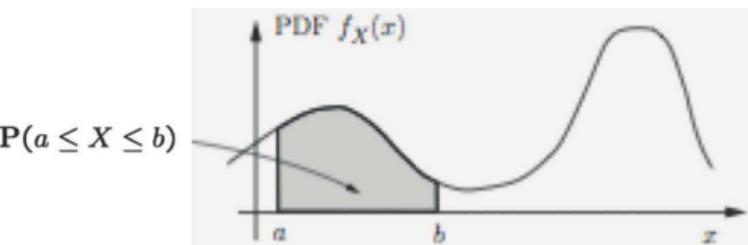
---

# Probability Density Functions (PDFs)

## Discrete (PMF)



## Continuous (PDF)



- $p_X(x) \geq 0$
- $\sum_x p_X(x) = 1$
- $P(a \leq X \leq b) = \sum_{x:a \leq x \leq b} p_X(x)$

- $f_X(x) \geq 0$
- $\int_{-\infty}^{\infty} f_X(x) dx = 1$
- $P(a \leq X \leq b) = \int_a^b f_X(x) dx$



# Interpreting the PDF

Unlike a PMF, the value of a PDF at  $x$ ,  $f_X(x)$ , is **not** a probability. It is a probability *density*.

- The probability of a continuous random variable falling within a very small interval  $[a, a + \delta]$  is approximately the area of a rectangle with height  $f_X(a)$  and width  $\delta$ .

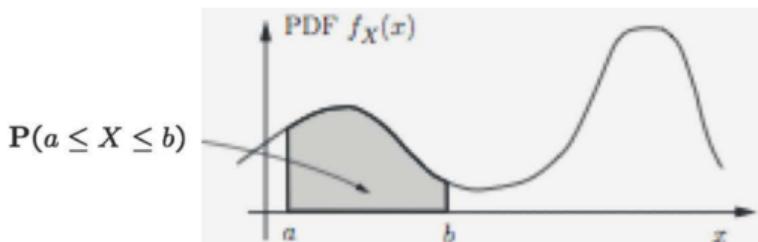
$$P(a \leq X \leq a + \delta) \approx f_X(a) \cdot \delta$$

- The probability of a continuous random variable taking on a single specific value is zero.

$$P(X = a) = \int_a^a f_X(x) dx = 0$$

This implies that for continuous random variables,

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b).$$

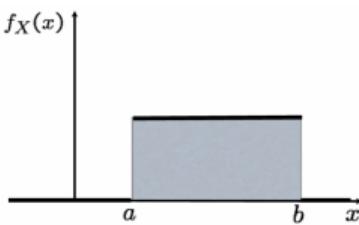
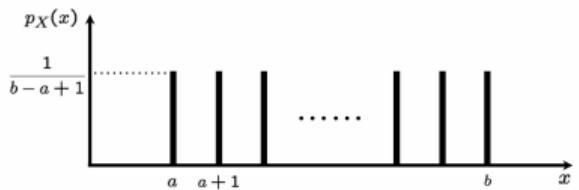


## Example: Continuous Uniform PDF

The continuous uniform random variable models a situation where a number is chosen from an interval  $[a, b]$ , and any sub-interval of a given length has the same probability.

### Uniform PDF

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$



A generalization is a piecewise constant PDF, where the density is constant over different intervals.

# Expectation of a Continuous Random Variable

The definition of expectation is analogous to the discrete case, with the sum replaced by an integral.

**Discrete (PMF)**

$$E[X] = \sum_x x p_X(x)$$

**Continuous (PDF)**

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

The properties of expectations carry over from the discrete world:

- **Expected Value Rule:**  $E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$
- **Linearity:**  $E[aX + b] = aE[X] + b$
- If  $a \leq X \leq b$ , then  $a \leq E[X] \leq b$ .

# Variance of a Continuous Random Variable

The definition and properties of variance are also analogous to the discrete case.

## Definition

Let  $\mu = E[X]$ . The variance of  $X$  is:

$$\text{var}(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx$$

- **Computational Formula:**  $\text{var}(X) = E[X^2] - (E[X])^2$
- **Linear Transformation:**  $\text{var}(aX + b) = a^2\text{var}(X)$

# Mean and Variance of the Continuous Uniform

Let  $X$  be uniform on  $[a, b]$ .

- **Mean:**

$$E[X] = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a} \left[ \frac{x^2}{2} \right]_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}$$

- **Variance:** First, find  $E[X^2]$ .

$$E[X^2] = \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{b-a} \left[ \frac{x^3}{3} \right]_a^b = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}$$

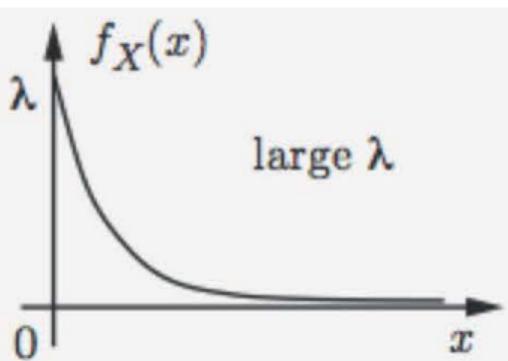
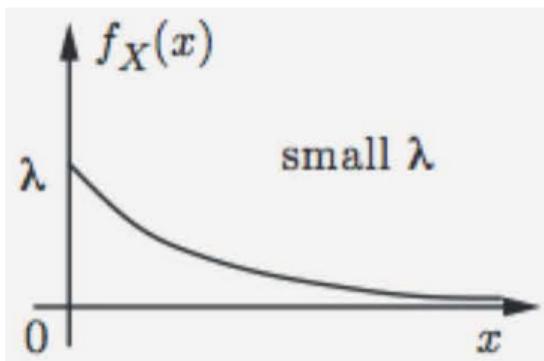
$$\text{var}(X) = E[X^2] - (E[X])^2 = \frac{a^2 + ab + b^2}{3} - \left( \frac{a+b}{2} \right)^2 = \frac{(b-a)^2}{12}$$

# The Exponential Random Variable

The exponential random variable is a continuous model often used for waiting times. It is the continuous analogue of the geometric distribution.

The exponential PDF with parameter  $\lambda > 0$  is:

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



Using integration by parts, one can show  $E[X] = \frac{1}{\lambda}$  and  $\text{var}(X) = \frac{1}{\lambda^2}$

# Cumulative Distribution Function (CDF)

The CDF can describe any random variable, discrete or continuous.

## Definition

The **Cumulative Distribution Function (CDF)** of a random variable  $X$  is the function  $F_X(x)$  defined as:

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

For **discrete** random variables:

$$F_X(x) = \sum_{k \leq x} p_X(k)$$

For **continuous** random variables:

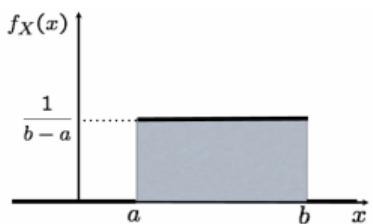
$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$

The PDF is the derivative of the CDF:  
 $f_X(x) = \frac{dF_X(x)}{dx}$ .

# CDF Examples

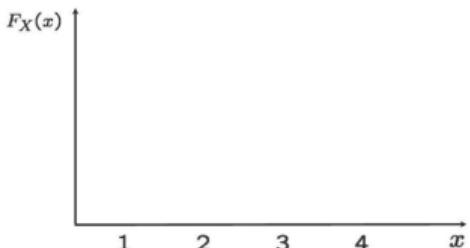
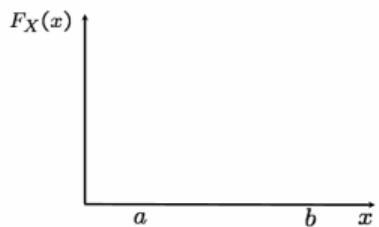
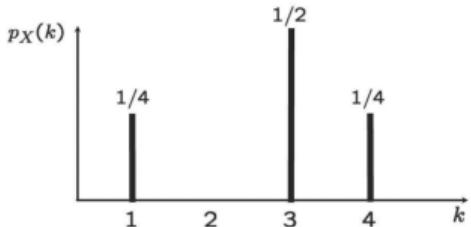
For **discrete** random variables:

$$F_X(x) = \sum_{k \leq x} p_X(k)$$



For **continuous** random variables:

$$F_X(x) = \int_{-\infty}^x f_X(t)dt$$





# General CDF Properties

Any valid CDF,  $F_X(x)$ , must have the following properties:

- It is a non-decreasing function of  $x$ .
- $F_X(x) \rightarrow 0$  as  $x \rightarrow -\infty$ .
- $F_X(x) \rightarrow 1$  as  $x \rightarrow \infty$ .

# Normal (Gaussian) Random Variables

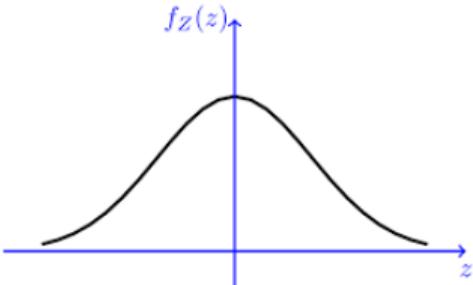
The normal distribution is arguably the most important in probability and statistics, due to its convenient analytical properties and its appearance in the Central Limit Theorem.

## Standard Normal: $N(0, 1)$

The PDF of a standard normal random variable is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

For the standard normal,  $E[X] = 0$  and  $\text{var}(X) = 1$ .



## General Normal: $N(\mu, \sigma^2)$

A general normal random variable has two parameters: mean  $\mu$  and variance  $\sigma^2$ .

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

For this distribution,  $E[X] = \mu$  and  $\text{var}(X) = \sigma^2$ .

# Properties of Normal Random Variables

A key property of normal random variables is that they are “closed under linear transformations.”

## Linear Functions of a Normal

If  $X \sim N(\mu, \sigma^2)$  and we define a new random variable  $Y = aX + b$ , then  $Y$  is also a normal random variable. We can find its mean and variance:

- $E[Y] = E[aX + b] = aE[X] + b = a\mu + b.$
- $\text{var}(Y) = \text{var}(aX + b) = a^2\text{var}(X) = a^2\sigma^2.$

Therefore,  $Y \sim N(a\mu + b, a^2\sigma^2).$

**Standardizing a Random Variable:** A common transformation is to standardize a random variable  $X$  (with mean  $\mu$  and variance  $\sigma^2$ ) to create a new variable  $Y$  with mean 0 and variance 1:

$$Y = \frac{X - \mu}{\sigma}$$

If  $X$  is normal, then  $Y$  is a standard normal random variable,  $Y \sim N(0, 1).$

# Calculating Normal Probabilities

The CDF of a normal random variable does not have a closed-form expression. We rely on pre-computed tables for the standard normal CDF, usually denoted by  $\Phi(y)$ .

$$\Phi(y) = P(Y \leq y), \quad \text{where } Y \sim N(0, 1)$$

To calculate a probability for a general normal  $X \sim N(\mu, \sigma^2)$ , we first standardize it:

$$P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P\left(Y \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

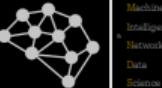
The table gives values of  $\Phi(y)$  for  $y \geq 0$ . For negative values, we use the symmetry of the PDF:

$$\Phi(-y) = P(Y \leq -y) = P(Y > y) = 1 - P(Y \leq y) = 1 - \Phi(y)$$

# **Lecture 9: Continuous Random Variables Part II**

## **Conditioning on an Event; Multiple Continuous r.v.'s**

---



# Conditional PDF, Given an Event

## Discrete

- $p_{X|A}(x) = P(X = x|A)$
- $\sum_x p_{X|A}(x) = 1$
- $P(X \in B|A) = \sum_{x \in B} p_{X|A}(x)$

## Continuous

- $f_{X|A}(x) \cdot \delta \approx P(x \leq X \leq x + \delta|A)$
- $\int_{-\infty}^{\infty} f_{X|A}(x) dx = 1$
- $P(X \in B|A) = \int_B f_{X|A}(x) dx$

# Conditional PDF of X, given that $X \in A$

The conditional PDF,  $f_{X|X \in A}(x)$ , must be zero outside of A. Inside A, the original PDF  $f_X(x)$  is rescaled so that the new total area is 1.

## Formula

Let A be an event with  $P(A) > 0$ . The conditional PDF of X given that  $X \in A$  is:

$$f_{X|X \in A}(x) = \begin{cases} \frac{f_X(x)}{P(A)}, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases}$$

where  $P(A) = \int_A f_X(t)dt$ .

# Conditional Expectation of X, Given an Event

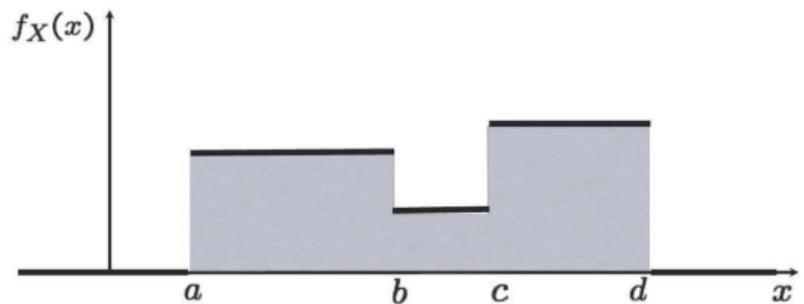
## Discrete

- $E[X|A] = \sum_x x p_{X|A}(x)$
- $E[g(X)|A] = \sum_x g(x) p_{X|A}(x)$

## Continuous

- $E[X|A] = \int x f_{X|A}(x) dx$
- $E[g(X)|A] = \int g(x) f_{X|A}(x) dx$

Example: Let  $A$  be the event that  $\frac{a+b}{2} \leq X \leq b$



# Memorylessness of the Exponential PDF

Let  $T$  be an exponential random variable with parameter  $\lambda$ , representing the lifetime of a light bulb.

$$f_T(t) = \lambda e^{-\lambda t}, \quad \text{for } t \geq 0$$

The probability that the bulb lasts longer than time  $x$  is  $P(T > x) = \int_x^{\infty} \lambda e^{-\lambda t} dt = e^{-\lambda x}$ .

## Memorylessness Property

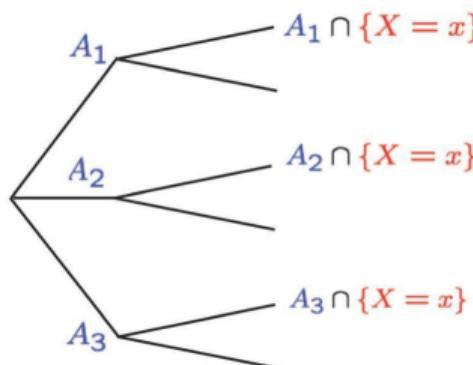
Suppose we know the bulb is still working at time  $t$  (event  $A = \{T > t\}$ ). Let  $X = T - t$  be the remaining lifetime.

$$\begin{aligned} P(X > x | T > t) &= P(T - t > x | T > t) = P(T > t + x | T > t) \\ &= \frac{P(T > t + x \text{ and } T > t)}{P(T > t)} = \frac{P(T > t + x)}{P(T > t)} \\ &= \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = e^{-\lambda x} = P(T > x) \end{aligned}$$

The remaining lifetime has the same exponential distribution as a brand new bulb.  
 Probabilistically, a used exponential bulb is as good as new!

# Total Probability and Expectation Theorems

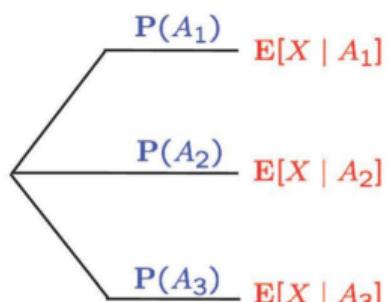
These theorems extend to continuous random variables. Let  $A_1, \dots, A_n$  be a partition of the sample space.

 Total Probability Theorem for PDFs

$$f_X(x) = P(A_1)f_{X|A_1}(x) + \cdots + P(A_n)f_{X|A_n}(x)$$

Total Expectation Theorem

$$E[X] = P(A_1)E[X|A_1] + \cdots + P(A_n)E[X|A_n]$$





## Example

Bill goes to the supermarket shortly, with probability  $1/3$ , at a time uniformly distributed between 0 and 2 hours from now; or with probability  $2/3$ , later in the day at a time uniformly distributed between 6 and 8 hours from now

# Mixed Distributions

A random variable is called **mixed** if it is neither purely discrete nor purely continuous. This often happens when combining different scenarios.

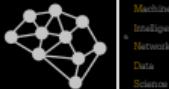
## Scenario

Let event  $A_1$  be “the value is drawn from a uniform distribution on  $[0,2]$ ”, with  $P(A_1) = 1/2$ .

Let event  $A_2$  be “the value is fixed at 1”, with  $P(A_2) = 1/2$ . Let  $X$  be the resulting random variable.

The expectation can be found using the total expectation theorem:

$$E[X] = \frac{1}{2}E[X|A_1] + \frac{1}{2}E[X|A_2] = \frac{1}{2}(1) + \frac{1}{2}(1) = 1.$$

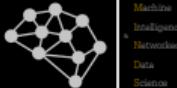


# Mixed Distributions

The CDF can be found using the law of total probability:

$$F_X(x) = P(A_1)F_{X|A_1}(x) + P(A_2)F_{X|A_2}(x)$$

# Jointly Continuous Random Variables



The concept of a joint PMF extends to a **joint PDF** for continuous random variables.

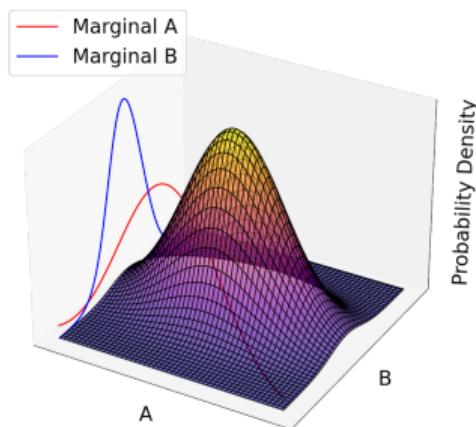
## Discrete

- $p_{X,Y}(x,y) = P(X = x, Y = y)$
- $P((X, Y) \in B) = \sum_{(x,y) \in B} p_{X,Y}(x,y)$

## Continuous

- $f_{X,Y}(x,y)$  is the probability density.
- $P((X, Y) \in B) = \iint_B f_{X,Y}(x,y) dx dy$

For a valid joint PDF, we must have  $f_{X,Y}(x,y) \geq 0$  and  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$ .





# From Joint to Marginal PDFs

We can obtain the marginal PDF of one variable by “integrating out” the other variable from the joint PDF.

**Discrete**

$$p_X(x) = \sum_y p_{X,Y}(x,y)$$

**Continuous**

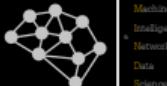
$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy$$

## Example: Uniform Joint PDF on a Set S

Let  $(X, Y)$  be chosen uniformly from a set  $S$ .



$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\text{Area of } S}, & \text{if } (x,y) \in S \\ 0, & \text{otherwise} \end{cases}$$



# Multiple Continuous Variables

## Discrete

- $p_{X,Y,Z}(x,y,z) = P(X = x, Y = y, Z = z)$
- $P((X, Y, Z) \in B) = \sum_{(x,y,z) \in B} p_{X,Y,Z}(x,y,z)$
- $p_{X,Y}(x) = \sum_z p_{X,Y,Z}(x,y,z)$

## Continuous

- $f_{X,Y,Z}(x,y,z)$  is the probability density.
- $P((X, Y, Z) \in B) = \iiint_B f_{X,Y,Z}(x,y,z) dx dy dz$
- $f_{X,Y}(x,y) = \int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z) dz$

### Expected Value Rule

For a function  $Z = g(X, Y)$ :

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy$$

### Linearity of Expectations

This property is universal and holds for continuous variables just as it does for discrete variables.

$$E[X + Y] = E[X] + E[Y], \quad E[aX + bY + c] = aE[X] + bE[Y] + c$$

$$E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n]$$

# The Joint CDF

## Definition

The joint CDF of random variables X and Y is:

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$$

For jointly continuous variables, the CDF is the integral of the PDF:

$$F_{X,Y}(x,y) = \int_{-\infty}^y \int_{-\infty}^x f_{X,Y}(u,v) dudv$$

Conversely, the joint PDF can be recovered from the joint CDF using partial derivatives:

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x,y)$$

# Lecture 10: Continuous Random Variables Part III Conditioning on a Random Variable; Independence; Bayes' Rule

---

# Conditional PDFs, Given Another Random Variable

The concept of a conditional PMF extends directly to a conditional PDF.

## Definition

The **conditional PDF** of  $X$  given  $Y = y$  is defined as:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

This is defined for any  $y$  such that the marginal PDF  $f_Y(y)$  is positive.

The probability of  $X$  being in a set  $A$ , given  $Y = y$ , is found by integrating the conditional PDF:

$$P(X \in A | Y = y) = \int_A f_{X|Y}(x|y) dx$$

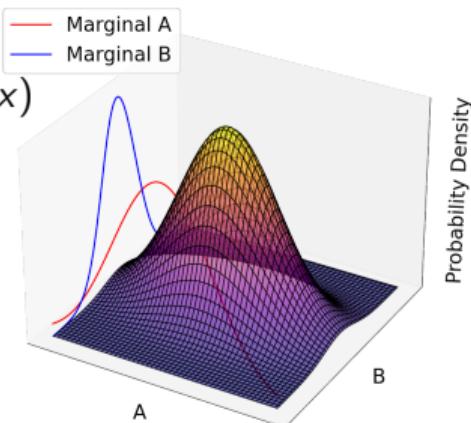
# Comments on Conditional PDFs

Let's interpret the conditional PDF,  $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ .

- For a fixed value of  $y$ ,  $f_{X|Y}(x|y)$  is a valid PDF for  $X$ .
  - $f_{X|Y}(x|y) \geq 0$ .
  - $\int_{-\infty}^{\infty} f_{X|Y}(x|y) dx = \frac{\int_{-\infty}^{\infty} f_{X,Y}(x,y) dx}{f_Y(y)} = \frac{f_Y(y)}{f_Y(y)} = 1$ .
- The shape of the conditional PDF  $f_{X|Y}(\cdot|y)$  can be visualized as a "slice" of the joint PDF surface at that particular  $y$ , rescaled to have a total area of 1.

## Multiplication Rule

$$f_{X,Y}(x,y) = f_Y(y)f_{X|Y}(x|y) = f_X(x)f_{Y|X}(y|x)$$



# Total Probability and Expectation Theorems

## Discrete Case

- $p_X(x) = \sum_y p_Y(y)p_{X|Y}(x|y)$
- $E[X] = \sum_y p_Y(y)E[X|Y = y]$

## Continuous Case

- $f_X(x) = \int_{-\infty}^{\infty} f_Y(y)f_{X|Y}(x|y)dy$
- $E[X] = \int_{-\infty}^{\infty} f_Y(y)E[X|Y = y]dy$

The conditional expectation is defined as:

$$E[X|Y = y] = \int_{-\infty}^{\infty} xf_{X|Y}(x|y)dx$$

# Independence

## Definition

Random variables  $X$  and  $Y$  are **independent** if their joint PDF is the product of their marginal PDFs for all  $x$  and  $y$ .

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

This is equivalent to the condition that the conditional PDF equals the marginal PDF:

$$f_{X|Y}(x|y) = f_X(x), \quad \text{for all } y \text{ with } f_Y(y) > 0$$

## Properties for Independent Variables

If  $X$  and  $Y$  are independent:

- $E[XY] = E[X]E[Y]$
- $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$
- $g(X)$  and  $h(Y)$  are also independent.

# Stick-Breaking Example

## Problem Setup

We break a stick of length  $\ell$  twice.

- The first break is at position  $X$ , which is uniform in  $[0, \ell]$ .
- The second break is at position  $Y$ , which is uniform in  $[0, X]$ .

# Stick-Breaking Example: Calculations

The joint PDF is  $f_{X,Y}(x,y) = \frac{1}{\ell x}$  for  $0 \leq y \leq x \leq \ell$ .

- **Marginal PDF of Y:**

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx = \int_y^{\ell} \frac{1}{\ell x} dx = \frac{1}{\ell} [\ln(x)]_y^{\ell} = \frac{1}{\ell} (\ln \ell - \ln y)$$

for  $0 < y \leq \ell$ .

- **Expectation of Y:** We can use the total expectation theorem, which is often simpler.

$$E[Y|X=x] = \frac{x}{2} \quad (\text{mean of a uniform on } [0,x])$$

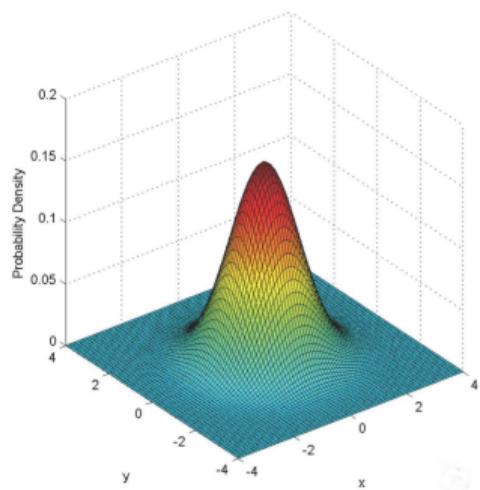
$$\begin{aligned} E[Y] &= E[E[Y|X]] = \int_0^{\ell} E[Y|X=x] f_X(x) dx = \int_0^{\ell} \frac{x}{2} \cdot \frac{1}{\ell} dx \\ &= \frac{1}{2\ell} \left[ \frac{x^2}{2} \right]_0^{\ell} = \frac{1}{2\ell} \frac{\ell^2}{2} = \frac{\ell}{4} \end{aligned}$$

# Independent Normal Random Variables

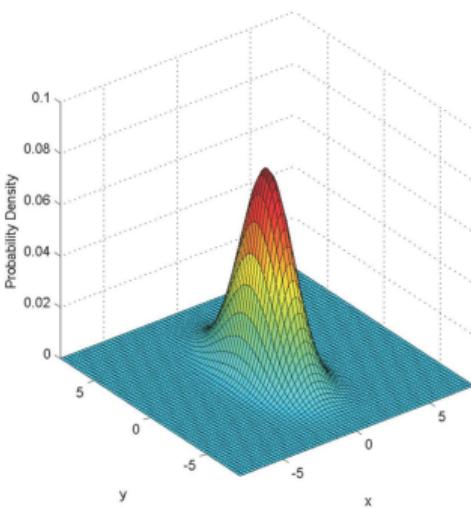
Let  $X \sim N(\mu_x, \sigma_x^2)$  and  $Y \sim N(\mu_y, \sigma_y^2)$  be independent. Their joint PDF is the product of their marginals.

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp \left\{ -\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2} \right\}$$

**Standard Normals** ( $\sigma_x^2 = \sigma_y^2 = 1$ )



**Different Variances** ( $\sigma_x^2 = 1, \sigma_y^2 = 4$ )



# The Bayes' Rule Theme with Variations

Bayes' rule is a general inference framework for updating beliefs about an unobserved variable  $X$  after observing a related variable  $Y$ .

The core relationship is:

$$\text{Posterior} = \frac{\text{Prior} \times \text{Likelihood}}{\text{Evidence}}$$

## Discrete Case

$$p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{p_Y(y)}$$

where  $p_Y(y) = \sum_{x'} p_X(x')p_{Y|X}(y|x')$ .

## Continuous Case

$$f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)}$$

where  $f_Y(y) = \int f_X(x')f_{Y|X}(y|x')dx'$ .

# Bayes' Rule: Mixed Cases

Bayes' rule also applies when one variable is discrete and the other is continuous.

## Case 1: Discrete Unknown (K), Continuous Measurement (Y)

$$p_{K|Y}(k|y) = \frac{p_K(k)f_{Y|K}(y|k)}{f_Y(y)}$$

where  $f_Y(y) = \sum_{k'} p_K(k')f_{Y|K}(y|k')$ .

## Case 2: Continuous Unknown (Y), Discrete Measurement (K)

$$f_{Y|K}(y|k) = \frac{f_Y(y)p_{K|Y}(k|y)}{p_K(k)}$$

where  $p_K(k) = \int f_Y(y')P(K = k|Y = y')dy'$ .

**Example:** Send a signal  $K$ , which is  $-1$  or  $+1$  with equal probability. Receive  $Y = K + W$ , where  $W \sim N(0, 1)$ . Given  $Y = y$ , what is  $P(K = 1|Y = y)$ ?

# Example Solution

- $P(K = 1) = 1/2$ .
- The model is  $f_{Y|K}(y|1) \sim N(1, 1)$  and  $f_{Y|K}(y|-1) \sim N(-1, 1)$ .
- The evidence term is  $f_Y(y) = \frac{1}{2}f_{Y|K}(y|1) + \frac{1}{2}f_{Y|K}(y|-1)$ .
- $p_{K|Y}(1|y) = \frac{\frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-(y-1)^2/2}}{\frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-(y-1)^2/2} + \frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-(y+1)^2/2}} = \frac{e^y}{e^y + e^{-y}}$ .

## **Lecture 11: Derived Distributions**

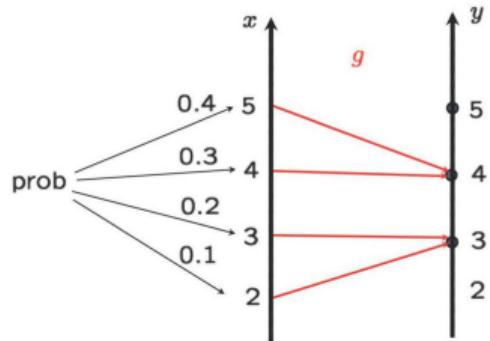
---

# Derived Distributions: The Discrete Case

If  $X$  is a discrete random variable and  $Y = g(X)$ , we can find the PMF of  $Y$  by summing the probabilities of all the  $x$  values that map to a given  $y$  value.

## Formula

$$p_Y(y) = P(Y = y) = P(g(X) = y) = \sum_{x:g(x)=y} p_X(x)$$

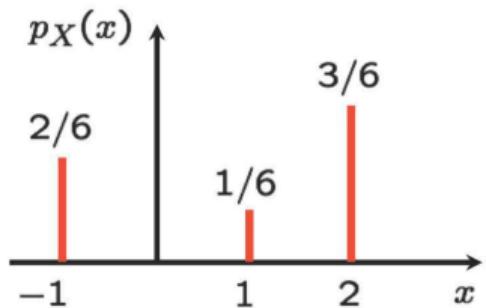


For linear transformation

$$p_Y(y) = P(aX + b \leq y) = P\left(X \leq \frac{y - b}{a}\right) = p_X\left(\frac{y - b}{a}\right)$$

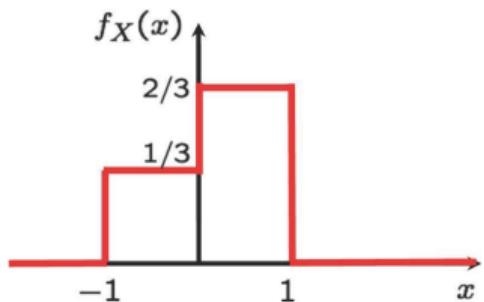
## The Linear Case: $Y = aX + b$ : Discrete

Find the PMF of  $Z = 2X$  and  $Y = 2X + 3$



## The Linear Case: $Y = aX + b$ : Continuous

Find the PDF of  $Z = 2X$  and  $Y = 2X + 3$



# Derived Distributions: The Continuous Case

For a continuous random variable  $X$  and a function  $Y = g(X)$ , finding the PDF of  $Y$  is more involved.

## General Two-Step Procedure

1. First, find the Cumulative Distribution Function (CDF) of  $Y$ .

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$$

To calculate this, we must find the set of  $x$  values for which  $g(x) \leq y$  and integrate the PDF of  $X$  over that set.

2. Then, differentiate the CDF to find the PDF of  $Y$ .

$$f_Y(y) = \frac{dF_Y}{dy}(y)$$

## The Linear Case: $Y = aX + b$

Let  $X$  be a continuous random variable and  $Y = aX + b$  with  $a \neq 0$ . We can find a direct formula for the PDF of  $Y$ .

Using the CDF method, assuming  $a > 0$ :

$$F_Y(y) = P(aX + b \leq y) = P\left(X \leq \frac{y - b}{a}\right) = F_X\left(\frac{y - b}{a}\right)$$

Differentiating with respect to  $y$  using the chain rule:

$$f_Y(y) = \frac{dF_Y}{dy}(y) = f_X\left(\frac{y - b}{a}\right) \cdot \frac{1}{a}$$

### General Formula for Linear Transformation ( $a \neq 0$ )

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y - b}{a}\right)$$

This shows that the PDF of  $Y$  is a scaled and shifted version of the PDF of  $X$ .

# Linear Function of a Normal RV is Normal

We can use the linear transformation rule to prove that a linear function of a normal random variable is also normal.

Let  $X \sim \mathcal{N}(\mu, \sigma^2)$  and  $Y = aX + b$  with  $a \neq 0$ .

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}$$

Using the formula  $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$ :

$$f_Y(y) = \frac{1}{|a|} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left(\frac{y-b}{a} - \mu\right)^2}{2\sigma^2}\right\} = \frac{1}{|a|\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y - (a\mu + b))^2}{2a^2\sigma^2}\right\}$$

This is the PDF of a normal random variable with mean  $a\mu + b$  and variance  $a^2\sigma^2$ .

Example:  $Y = X^3$  when  $X$  uniform on  $[0,2]$

## Example

You go to the gym and set the speed  $X$  of the treadmill to a number between 5 and 10 km/hr Uniformly. Find the PDF of the time it takes to run 10km.

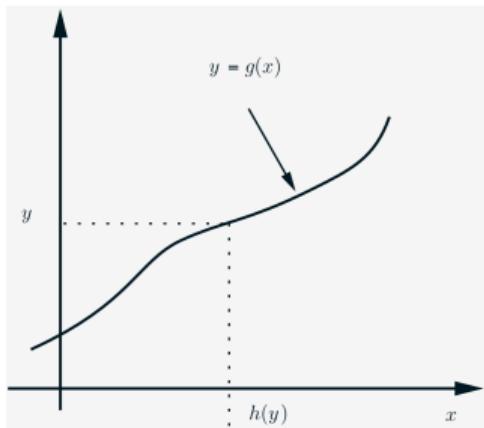
# General Formula for Monotonic Functions

When  $g(X)$  is a strictly monotonic and differentiable function, we can derive a direct formula for the PDF of  $Y = g(X)$ .

Let  $h$  be the inverse function of  $g$ , so that if  $y = g(x)$ , then  $x = h(y)$ .

## PDF Formula for Monotonic g

$$f_Y(y) = f_X(h(y)) \left| \frac{dh}{dy}(y) \right|$$

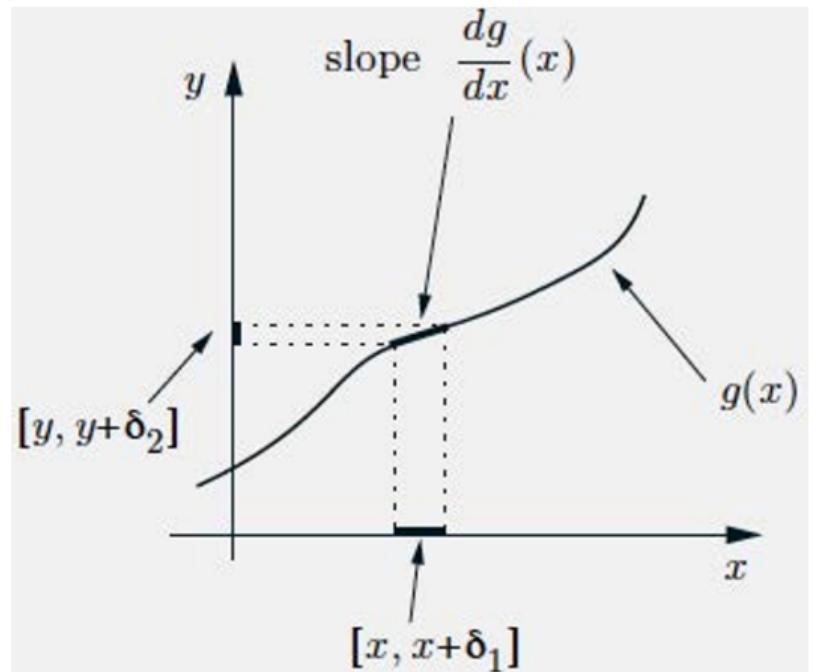


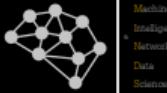
## Example

**Example:** Let  $X \sim U[0, 1]$  and  $Y = X^2$ . Here,  $g(x) = x^2$  is increasing on  $[0, 1]$ . The inverse is  $h(y) = \sqrt{y}$ .  $\frac{dh}{dy} = \frac{1}{2\sqrt{y}}$ .  $f_X(x) = 1$  for  $x \in [0, 1]$ .

$$f_Y(y) = f_X(\sqrt{y}) \left| \frac{1}{2\sqrt{y}} \right| = 1 \cdot \frac{1}{2\sqrt{y}}, \quad \text{for } y \in [0, 1]$$

# Intuition for Monotone Functions





## Non-Monotonic Example: $Y = X^2$

When  $g(x)$  is not monotonic, we must return to the CDF method.

# Functions of Multiple Random Variables: $Z = g(X, Y)$

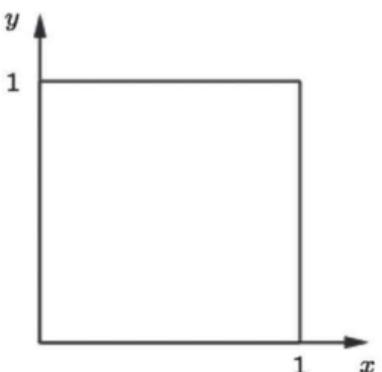
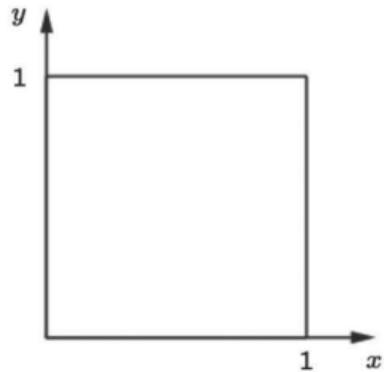
The same two-step CDF methodology applies when finding the distribution of a function of multiple random variables.

## Procedure for $Z = g(X, Y)$

1. Find the CDF of  $Z$ :  $F_Z(z) = P(Z \leq z) = P(g(X, Y) \leq z)$ . This involves integrating the joint PDF  $f_{X,Y}(x, y)$  over the 2D region where  $g(x, y) \leq z$ .
2. Differentiate the CDF to find the PDF:  $f_Z(z) = \frac{dF_Z}{dz}(z)$ .

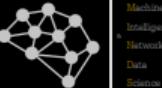
# Functions of Multiple Random Variables: $Z = g(X, Y)$

**Example:** Let  $X, Y$  be independent and uniform on  $[0, 1]$ . Find the PDF of  $Z = Y/X$ . For  $z > 0$ ,  $F_Z(z) = P(Y/X \leq z) = P(Y \leq zX)$ . We must integrate the joint PDF ( $f_{X,Y}(x,y) = 1$ ) over the region  $\{0 \leq x \leq 1, 0 \leq y \leq 1, y \leq zx\}$ .



# **Lecture 12: Sums of Independent Random Variables; Covariance and Correlation**

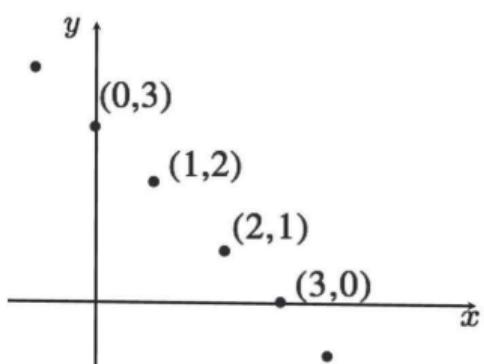
---



# The Distribution of $Z = X + Y$ : The Discrete Case

Let  $X$  and  $Y$  be independent discrete random variables with known PMFs. We want to find the PMF of their sum,  $Z = X + Y$ .

To find  $p_Z(z)$ , we sum the probabilities of all pairs  $(x, y)$  such that  $x + y = z$ .

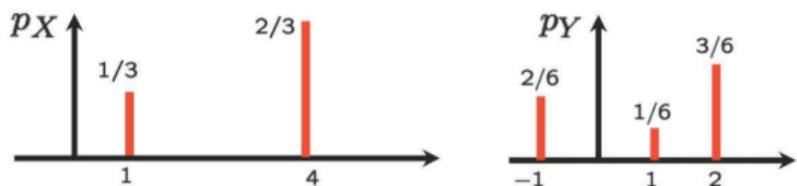


$$\begin{aligned} p_Z(z) &= P(X + Y = z) = \sum_x P(X = x, Y = z - x) \\ &= \sum_x P(X = x)P(Y = z - x) \quad (\text{by independence}) \\ &= \sum_x p_X(x)p_Y(z - x) \quad \text{Discrete Convolution} \end{aligned}$$

# Discrete Convolution Mechanics

$p_Z(z) = \sum_x p_X(x)p_Y(z - x)$  can be visualized as a “flip, shift, multiply, and sum” operation.

1. Take the PMF of  $Y$ ,  $p_Y(k)$ , and flip it horizontally around the vertical axis to get  $p_Y(-k)$ .
2. Shift the flipped PMF to the right by  $z$  to get  $p_Y(z - k)$ .
3. Place this shifted, flipped PMF under the PMF of  $X$ .
4. Multiply the overlapping values point-by-point and sum the results.



# The Distribution of $Z = X + Y$ : The Continuous Case

The logic for continuous random variables is analogous.

We start with the CDF of  $Z = X + Y$ :

$$\begin{aligned} F_Z(z) &= P(X + Y \leq z) = \iint_{x+y \leq z} f_X(x)f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} f_X(x) \left( \int_{-\infty}^{z-x} f_Y(y) dy \right) dx = \int_{-\infty}^{\infty} f_X(x) F_Y(z - x) dx \end{aligned}$$

Differentiating with respect to  $z$  gives the PDF of  $Z$ :

$$f_Z(z) = \frac{dF_Z}{dz}(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z - x)dx \quad \text{Continuous Convolution Formula}$$

The mechanics are the same as in the discrete case, but with integration instead of summation.

# The Sum of Independent Normal RVs is Normal

A key result in probability theory is that the sum of independent normal random variables is also a normal random variable.

## Theorem

If  $X \sim N(\mu_x, \sigma_x^2)$  and  $Y \sim N(\mu_y, \sigma_y^2)$  are independent, then their sum  $Z = X + Y$  is a normal random variable.

We already know the mean and variance of the sum:

- $E[Z] = E[X] + E[Y] = \mu_x + \mu_y$
- $\text{var}(Z) = \text{var}(X) + \text{var}(Y) = \sigma_x^2 + \sigma_y^2$

Therefore, if  $X$  and  $Y$  are independent normal random variables:

$$X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

This can be proven by carrying out the convolution integral of the two normal PDFs (“completing the square” inside an exponential).

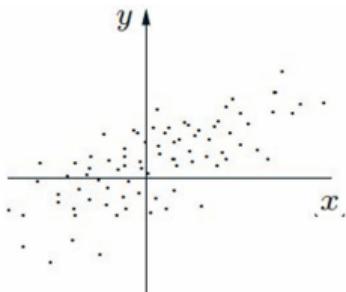
# Covariance

## Definition

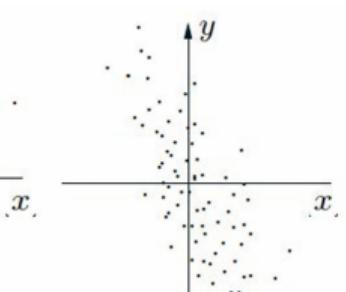
The **covariance** of two random variables  $X$  and  $Y$  is:

$$\text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

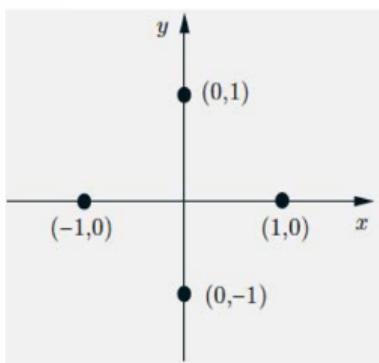
- If the variables tend to be on the same side of their respective means (e.g.,  $X$  high when  $Y$  is high), the covariance is positive.
- If they tend to be on opposite sides, the covariance is negative.
- If  $X$  and  $Y$  are independent, then  $\text{cov}(X, Y) = E[X - E[X]]E[Y - E[Y]] = 0 \cdot 0 = 0$ .
- The converse is not true: zero covariance does not imply independence.



$E[XY]$



$E[XY]$



# Covariance Properties

- $\text{cov}(X, X) = E[(X - E[X])^2] = \text{var}(X)$
- A more convenient computational formula:

$$\text{cov}(X, Y) = E[XY - XE[Y] - YE[X] + E[X]E[Y]]$$

$$= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] = E[XY] - E[X]E[Y]$$

Thus

$$\text{cov}(X, Y) = E[XY] - E[X]E[Y] = E[(X - E[X])(Y - E[Y])]$$

- Covariance is bilinear:

$$\text{cov}(aX+b, Y) = a \cdot \text{cov}(X, Y)$$

$$\text{cov}(X, Y+Z) = \text{cov}(X, Y) + \text{cov}(X, Z)$$

# The Variance of a Sum

Using the properties of covariance, we can find a general formula for the variance of a sum of random variables.

$$\begin{aligned}\text{var}(X_1 + X_2) &= \text{cov}(X_1 + X_2, X_1 + X_2) \\ &= \text{cov}(X_1, X_1) + \text{cov}(X_1, X_2) + \text{cov}(X_2, X_1) + \text{cov}(X_2, X_2) \\ &= \text{var}(X_1) + \text{var}(X_2) + 2\text{cov}(X_1, X_2)\end{aligned}$$

## General Formula for Variance of a Sum

$$\text{var} \left( \sum_{i=1}^n X_i \right) = \sum_{i=1}^n \text{var}(X_i) + \sum_{i \neq j} \text{cov}(X_i, X_j)$$

This shows that the variance of the sum is the sum of the variances only if the random variables are pairwise uncorrelated ( $\text{cov}(X_i, X_j) = 0$  for all  $i \neq j$ ).

# The Correlation Coefficient

Covariance depends on the units of the random variables. The correlation coefficient is a normalized, dimensionless version.

## Definition

The **correlation coefficient**  $\rho(X, Y)$  is:

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

- It is a measure of the linear relationship between  $X$  and  $Y$ .
- It is always between -1 and 1:  $-1 \leq \rho(X, Y) \leq 1$ .
- If  $X, Y$  are independent, they are uncorrelated and  $\rho = 0$ .
- $|\rho| = 1$  if and only if  $Y - E[Y]$  is a linear function of  $X - E[X]$ .
- $\rho(aX + b, Y) = \text{sgn}(a)\rho(X, Y)$ , where  $\text{sgn}(a)$  is the sign of  $a$ .

## Proof of Correlation Properties

To show that  $-1 \leq \rho \leq 1$ , we can assume without loss of generality that  $E[X] = E[Y] = 0$  and  $\text{var}(X) = \text{var}(Y) = 1$ . In this case,  $\rho = E[XY]$ .

Consider the non-negative quantity  $E[(X - \rho Y)^2]$ :

$$\begin{aligned} 0 \leq E[(X - \rho Y)^2] &= E[X^2 - 2\rho XY + \rho^2 Y^2] \\ &= E[X^2] - 2\rho E[XY] + \rho^2 E[Y^2] \\ &= \text{var}(X) - 2\rho \cdot \rho + \rho^2 \text{var}(Y) \\ &= 1 - 2\rho^2 + \rho^2 = 1 - \rho^2 \end{aligned}$$

So,  $1 - \rho^2 \geq 0$ , which implies  $\rho^2 \leq 1$ , or  $-1 \leq \rho \leq 1$ .

If  $|\rho| = 1$ , then  $\rho^2 = 1$ , which means  $E[(X - \rho Y)^2] = 0$ . Since  $(X - \rho Y)^2$  is a non-negative random variable, its expectation is zero if and only if the random variable is always zero. Thus,  $X - \rho Y = 0$  with probability 1, meaning  $X$  and  $Y$  are linearly related.

# Interpreting Correlation

Correlation does not imply causation!

A strong correlation between two variables (e.g., math aptitude and musical ability) may not mean that one influences the other. Instead, it often reflects an underlying common, or "hidden," factor (e.g., general intelligence, dedication).

## Example

Let  $Z, V, W$  be independent random variables with zero mean and unit variance. Let  $X = Z + V$  and  $Y = Z + W$ .

$$E[X] = E[Z] + E[V] = 0. \quad E[Y] = E[Z] + E[W] = 0. \quad \text{var}(X) = \text{var}(Z) + \text{var}(V) = 2. \\ \text{var}(Y) = \text{var}(Z) + \text{var}(W) = 2.$$

$$\text{cov}(X, Y) = \text{cov}(Z + V, Z + W) = \text{cov}(Z, Z) + \text{cov}(Z, W) + \text{cov}(V, Z) + \text{cov}(V, W) \quad \text{Since} \\ \text{they are independent, all cross-covariances are zero.} \quad \text{cov}(X, Y) = \text{var}(Z) = 1.$$

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{1}{\sqrt{2}\sqrt{2}} = \frac{1}{2}.$$

$X$  and  $Y$  are correlated because they share the common underlying factor  $Z$ .

Ignoring correlation can lead to a massive underestimation of risk.

## Example

A company invests \$10M in 10 states. The return in each state,  $X_i$ , has mean \$1M and standard deviation \$1.3M. Total return is  $S = X_1 + \dots + X_{10}$ .

$$\text{var}(S) = \sum_{i=1}^{10} \text{var}(X_i) + \sum_{i \neq j} \text{cov}(X_i, X_j)$$

- **Case 1: Uncorrelated returns ( $\rho = 0$ )** All covariances are 0.

$\text{var}(S) = 10 \cdot \text{var}(X_i) = 10 \cdot (1.3)^2 = 16.9$ . The standard deviation of the total return is  $\sigma_S = \sqrt{16.9} \approx \$4.11M$ .

- **Case 2: Highly correlated returns ( $\rho = 0.9$ )**  $\text{cov}(X_i, X_j) = \rho \sigma_i \sigma_j = 0.9 \cdot 1.3 \cdot 1.3 = 1.521$ .

There are  $10 \times 9 = 90$  such covariance terms.

$\text{var}(S) = 16.9 + 90 \cdot (1.521) = 16.9 + 136.89 = 153.79$ . The standard deviation is  $\sigma_S = \sqrt{153.79} \approx \$12.4M$ .

High correlation dramatically increases the overall risk (variance).

## **Lecture 13: Conditional expectation and variance revisited; Sum of a random number of independent r.v.'s**

---

# Conditional Expectation as a Random Variable

We have previously defined the conditional expectation  $E[X|Y = y]$  as a number that depends on the specific value  $y$ .

Let's define a function  $g(y) = E[X|Y = y]$ .

Now, consider the quantity  $g(Y)$ . This is a function of the random variable  $Y$ , which makes it a random variable itself.

**Definition:**  $E[X|Y]$

We define the conditional expectation  $E[X|Y]$  as the random variable  $g(Y)$ . It is the random variable whose value is  $E[X|Y = y]$  when the outcome of the experiment is such that  $Y = y$ .

- $E[X|Y]$  is a function of the random variable  $Y$ .
- $E[X|Y]$  is itself a random variable.
- As a random variable, it has a distribution, a mean, a variance, etc.

# The Mean of $E[X|Y]$ : Law of Iterated Expectations

What is the expected value of the random variable  $E[X|Y]$ ?

Let  $g(Y) = E[X|Y]$ . By the expected value rule:

$$E[g(Y)] = \sum_y g(y)p_Y(y) = \sum_y E[X|Y=y]p_Y(y)$$

This is precisely the formula for the total expectation theorem.

## Law of Iterated Expectations

The expectation of the conditional expectation of X given Y is simply the expectation of X.

$$E[E[X|Y]] = E[X]$$

This is a more abstract, and often more powerful, way of stating the total expectation theorem.

# Stick-Breaking Example Revisited

## Problem

A stick of length  $\ell$  is broken at a uniformly chosen point  $Y$ . The left part, of length  $Y$ , is then broken again at a uniformly chosen point  $X$ .

- $Y \sim U[0, \ell]$ , so  $f_Y(y) = 1/\ell$  for  $y \in [0, \ell]$ .
- Given  $Y = y$ ,  $X$  is uniform on  $[0, y]$ , so  $f_{X|Y}(x|y) = 1/y$  for  $x \in [0, y]$ .

The conditional expectation of  $X$  given  $Y = y$  is the mean of a  $U[0, y]$  distribution:

$$E[X|Y = y] = \frac{y}{2}$$

The conditional expectation as a random variable,  $E[X|Y]$ , is therefore the random variable whose value is  $y/2$  when  $Y = y$ . We can write this simply as:

$$E[X|Y] = \frac{Y}{2}$$

Using the law of iterated expectations to find the overall mean of  $X$ :

$$E[X] = E[E[X|Y]] = E\left[\frac{Y}{2}\right] = \frac{1}{2}E[Y] = \frac{1}{2} \cdot \frac{\ell}{2} = \frac{\ell}{4}$$

# Forecast Revisions

The law of iterated expectations has a nice interpretation in the context of forecasting.

- Let  $X$  be an unknown quantity we want to predict (e.g., February sales).
- Our initial forecast, with no extra information, is the unconditional mean  $E[X]$ .
- Suppose at the end of January, we observe some related data  $Y$  (e.g., January sales).
- Our new, revised forecast for  $X$  is the conditional expectation  $E[X|Y]$ .

The law of iterated expectations,  $E[E[X|Y]] = E[X]$ , means that the average of all possible revised forecasts, weighted by the likelihood of observing the information that leads to them, is equal to our original, uninformed forecast. Your forecast might go up or down depending on the new information, but on average, it stays the same.

# The Conditional Variance as a Random Variable

We can define the conditional variance in a similar way.

## Definition

The conditional variance of  $X$  given  $Y = y$  is the variance of  $X$  under the conditional distribution:

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

We then define  $\text{var}(X|Y)$  as the random variable that takes the value  $\text{var}(X|Y = y)$  when  $Y = y$ .

**Example:** If  $X$  is uniform on  $[0, Y]$ , then  $\text{var}(X|Y = y) = \frac{(y-0)^2}{12} = \frac{y^2}{12}$ . The random variable is therefore  $\text{var}(X|Y) = \frac{Y^2}{12}$ .

## Law of Total Variance

$$\text{var}(X) = E[\text{var}(X|Y)] + \text{var}(E[X|Y])$$

# Derivation of the Law of Total Variance

The formula is  $\text{var}(X) = E[\text{var}(X|Y)] + \text{var}(E[X|Y])$ .

- The term  $E[\text{var}(X|Y)]$  is the “expected value of the conditional variance”. It represents the average variability of  $X$  around its conditional mean,  $E[X|Y]$ .

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

- The term  $\text{var}(E[X|Y])$  is the “variance of the conditional expectation”. It represents the variability in the conditional mean itself as  $Y$  changes.

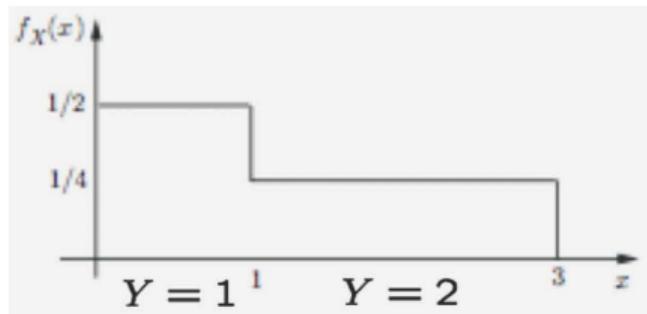
$$\text{var}(E[X|Y]) = E[(E[X|Y])^2] - (E[E[X|Y]])^2 = E[(E[X|Y])^2] - (E[X])^2$$

Adding the two terms together:

$$\begin{aligned} E[\text{var}(X|Y)] + \text{var}(E[X|Y]) &= (E[X^2] - E[(E[X|Y])^2]) + (E[(E[X|Y])^2] - (E[X])^2) \\ &= E[X^2] - (E[X])^2 = \text{var}(X) \end{aligned}$$

## A Simple Example

The formula is  $\text{var}(X) = E[\text{var}(X|Y)] + \text{var}(E[X|Y])$ .





## Another Example



## Another Example (Cont'd)

## Sum of a Random Number of Independent RVs: Mean

Consider a sum  $Y = X_1 + \dots + X_N$ , where  $N$  is a random variable representing the number of terms in the sum. Assume the  $X_i$  are i.i.d. and are independent of  $N$ . Let  $E[X_i] = E[X]$  and  $\text{var}(X_i) = \text{var}(X)$ .

We use the law of iterated expectations, conditioning on  $N$ .

$$E[Y] = E[E[Y|N]]$$

First, we find the inner conditional expectation,  $E[Y|N = n] = E[X_1 + \dots + X_n|N = n]$ .

Since the  $X_i$  are independent of  $N$ , this is simply

$$E[X_1 + \dots + X_n] = E[X_1] + \dots + E[X_n] = nE[X].$$

So, the random variable  $E[Y|N]$  is equal to the random variable  $N \cdot E[X]$ .

Now, we take the outer expectation:

$$E[Y] = E[N \cdot E[X]] = E[X]E[N] \quad (\text{Since } E[X] \text{ is a constant})$$

# Sum of a Random Number of Independent RVs: Variance

We find the variance of  $Y = X_1 + \dots + X_N$  using the law of total variance:

$$\text{var}(Y) = E[\text{var}(Y|N)] + \text{var}(E[Y|N])$$

- **First term:**  $E[\text{var}(Y|N)]$ . We first find  $\text{var}(Y|N = n)$ . Since the  $X_i$  are independent, the variance of their sum is the sum of their variances:

$$\text{var}(Y|N = n) = \text{var}(X_1 + \dots + X_n) = n \cdot \text{var}(X)$$

This means the random variable  $\text{var}(Y|N)$  is equal to  $N \cdot \text{var}(X)$ . Taking the expectation:

$$E[\text{var}(Y|N)] = E[N \cdot \text{var}(X)] = E[N]\text{var}(X)$$

- **Second term:**  $\text{var}(E[Y|N])$ . Recall, we know the random variable  $E[Y|N]$  is  $N \cdot E[X]$ .

$$\text{var}(E[Y|N]) = \text{var}(N \cdot E[X]) = (E[X])^2\text{var}(N) \quad (\text{Since } E[X] \text{ is a constant})$$

## **Lecture 14: Bi-variate and Multivariate Normal**

---

# Definition of Jointly Normal Random Variables

The bivariate normal distribution is a fundamental model for the joint behavior of two continuous random variables.

## Definition

Two random variables  $X$  and  $Y$  are said to be “jointly normal” if they can be expressed as linear combinations of two independent normal random variables,  $U$  and  $V$ .

$$X = aU + bV$$

$$Y = cU + dV$$

where  $a, b, c, d$  are scalars.

- Both  $X$  and  $Y$  are individually normal.
- Any linear combination  $Z = s_1X + s_2Y$  is also normal.

# Zero Correlation Implies Independence

For general random variables, zero correlation does not imply independence. However, for jointly normal random variables, it does.

## Key Property

If two random variables  $X$  and  $Y$  are jointly normal and are uncorrelated (i.e.,  $\text{cov}(X, Y) = 0$ ), then they are independent.

This property is a cornerstone of the theory and simplifies many analyses involving normal random variables.

# The Conditional Distribution of $X$ Given $Y$

We can decompose  $X$  into two parts: one that is perfectly predictable from  $Y$ , and an error term that is independent of  $Y$ .

Let's define the linear least squares estimator of  $X$  given  $Y$  (something we will learn more about later):

$$\hat{X} = E[X] + \rho \frac{\sigma_X}{\sigma_Y} (Y - E[Y])$$

And the estimation error (something we will learn more about later):

$$\tilde{X} = X - \hat{X}$$

It can be shown that  $\tilde{X}$  and  $Y$  are jointly normal and uncorrelated, and therefore independent.



# The Conditional Distribution of X Given Y

# Conditional Expectation of X Given Y

Using the decomposition  $X = \hat{X} + \tilde{X}$ , we can find the conditional expectation.

$$\begin{aligned}E[X|Y] &= E[\hat{X} + \tilde{X}|Y] \\&= E[\hat{X}|Y] + E[\tilde{X}|Y]\end{aligned}$$

Since  $\hat{X}$  is a function of  $Y$ ,  $E[\hat{X}|Y] = \hat{X}$ . Since  $\tilde{X}$  is independent of  $Y$ ,  $E[\tilde{X}|Y] = E[\tilde{X}] = 0$ .

## Result

The conditional expectation of  $X$  given  $Y$  is a linear function of  $Y$ :

$$E[X|Y] = \hat{X} = E[X] + \rho \frac{\sigma_X}{\sigma_Y} (Y - E[Y])$$

## Conditional Variance of X Given Y

The conditional distribution of  $X$  given  $Y = y$  is the distribution of  $\tilde{X}$  shifted by the constant  $\hat{X}(y)$ .

Therefore, the conditional variance of  $X$  given  $Y$  is simply the variance of the error term  $\tilde{X}$ .

The variance of the error term can be calculated as:

$$\sigma_{\tilde{X}}^2 = E[(X - \hat{X})^2] = (1 - \rho^2)\sigma_X^2$$

### Summary: Conditional Distribution

If  $X$  and  $Y$  are jointly normal, the conditional distribution of  $X$  given  $Y = y$  is normal with:

- Mean:  $E[X|Y = y] = E[X] + \rho \frac{\sigma_X}{\sigma_Y} (y - E[Y])$
- Variance:  $\text{var}(X|Y = y) = (1 - \rho^2)\sigma_X^2$



# Conditional Variance of X Given Y

# The Form of the Bivariate Normal PDF

Using the multiplication rule  $f_{X,Y}(x,y) = f_Y(y)f_{X|Y}(x|y)$  and the results for the conditional distribution, we can derive the full joint PDF. For simplicity, assume zero means ( $E[X] = E[Y] = 0$ ).

The joint PDF is of the form  $f_{X,Y}(x,y) = c \cdot e^{-q(x,y)}$ , where  $c$  is a normalizing constant and  $q(x,y)$  is a quadratic exponent term.

## Bivariate Normal PDF (Zero Mean)

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left( \frac{x^2}{\sigma_X^2} - \frac{2\rho xy}{\sigma_X\sigma_Y} + \frac{y^2}{\sigma_Y^2} \right) \right\}$$

The PDF is completely determined by the two means, two variances, and the correlation coefficient.



# The Form of the Bivariate Normal PDF

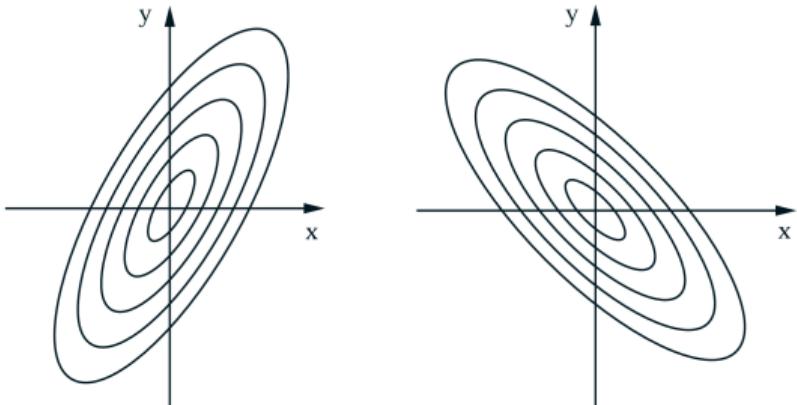
# Contours of the Bivariate Normal PDF

The sets of points where the PDF is constant, called contours or level sets, are described by the equation  $q(x, y) = \text{constant}$ .

$$\frac{x^2}{\sigma_X^2} - \frac{2\rho xy}{\sigma_X \sigma_Y} + \frac{y^2}{\sigma_Y^2} = \text{constant}$$

This is the equation of an ellipse centered at the mean  $(\mu_X, \mu_Y)$ .

- If  $\rho = 0$ , the axes of the ellipse are horizontal and vertical.
- If  $\rho \neq 0$ , the axes are tilted. The tilt direction depends on the sign of  $\rho$ .



## Example: Part 1

### Problem

Let  $X$  and  $Z$  be zero-mean jointly normal random variables with  $\sigma_X^2 = 4$ ,  $\sigma_Z^2 = 17/9$ , and  $E[XZ] = 2$ . Define  $Y = 2X - 3Z$ . Find the PDF of  $Y$ .

**Solution:** Since  $Y$  is a linear combination of jointly normal variables,  $Y$  is normal.

- Mean:  $E[Y] = E[2X - 3Z] = 2E[X] - 3E[Z] = 0$ .
- Variance:

$$\begin{aligned}\sigma_Y^2 &= E[Y^2] = E[(2X - 3Z)^2] = E[4X^2 - 12XZ + 9Z^2] \\&= 4E[X^2] - 12E[XZ] + 9E[Z^2] \\&= 4\sigma_X^2 - 12E[XZ] + 9\sigma_Z^2 \\&= 4(4) - 12(2) + 9(17/9) = 16 - 24 + 17 = 9.\end{aligned}$$

So,  $Y \sim N(0, 9)$ , and its PDF is  $f_Y(y) = \frac{1}{3\sqrt{2\pi}} e^{-y^2/18}$ .

## Example: Part 2

### Problem

Continuing with the previous example, find the conditional PDF of  $X$  given  $Y$ .

**Solution:** We need the parameters for the conditional normal distribution.

- We know  $\sigma_X^2 = 4$  and  $\sigma_Y^2 = 9$ , so  $\sigma_X = 2$  and  $\sigma_Y = 3$ .
- We need the correlation coefficient  $\rho$ . First, find the covariance:

$$\begin{aligned}\text{cov}(X, Y) &= E[XY] = E[X(2X - 3Z)] = 2E[X^2] - 3E[XZ] \\ &= 2(4) - 3(2) = 8 - 6 = 2.\end{aligned}$$

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{2}{2 \cdot 3} = \frac{1}{3}$$

- The conditional mean is  $E[X|Y = y] = \rho \frac{\sigma_X}{\sigma_Y} y = \frac{1}{3} \frac{2}{3} y = \frac{2}{9} y$ .
- The conditional variance is  $\sigma_{\tilde{X}}^2 = (1 - \rho^2)\sigma_X^2 = (1 - (1/3)^2) \cdot 4 = (1 - 1/9) \cdot 4 = \frac{8}{9} \cdot 4 = \frac{32}{9}$ .

The conditional distribution of  $X$  given  $Y = y$  is  $N\left(\frac{2y}{9}, \frac{32}{9}\right)$ .

# A Cautionary Note

- If  $X$  and  $Y$  are jointly normal, then  $X$  is normal and  $Y$  is normal.
- The converse is **not** true.

## Counterexample

Let  $X \sim N(0, 1)$ . Let  $Z$  be an independent random variable with  $P(Z = 1) = P(Z = -1) = 1/2$ . Define  $Y = ZX$ .

- The marginal PDF of  $Y$  is normal  $N(0, 1)$ .
- $X$  and  $Y$  are uncorrelated:  $E[XY] = E[X(ZX)] = E[ZE[X^2]] = E[Z]E[X^2] = 0 \cdot 1 = 0$ .
- However,  $X$  and  $Y$  are clearly dependent. If we know  $X = 2$ , then  $Y$  must be either 2 or -2.
- Since they are dependent but uncorrelated, they cannot be jointly normal.

# The Multivariate Normal PDF

The concepts generalize to more than two random variables.

## Definition

Random variables  $X_1, \dots, X_n$  are “jointly normal” if they are all linear functions of a set of independent normal random variables  $U_1, \dots, U_n$ .

- Zero correlation still implies independence.
- Conditional expectations are linear functions of the conditioning variables.
- The joint PDF has the form  $f(x) = c \cdot e^{-q(x)}$ , where  $q(x)$  is a quadratic function of the variables  $x_1, \dots, x_n$ .

## **Lecture 15: Transforms and Moment Generating Functions (MGFs)**

---

# What is a Transform?

The transform provides an alternative representation of a probability law. It is not always intuitive, but it is a powerful mathematical tool.

## Definition

The transform associated with a random variable  $X$ , also known as the “moment generating function” (MGF), is a function  $M_X(s)$  of a scalar parameter  $s$ , defined by:

$$M_X(s) = E[e^{sX}]$$

- **Discrete Case:**

$$M_X(s) = \sum_x e^{sx} p_X(x)$$

- **Continuous Case:**

$$M_X(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx$$

The transform is only defined for values of  $s$  for which the expectation is finite.

## Example: Transform of a Poisson RV

Let  $X$  be a Poisson random variable with parameter  $\lambda$ . Its PMF is  $p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}$  for  $k = 0, 1, 2, \dots$ .

The corresponding transform is:

$$\begin{aligned}
 M_X(s) &= E[e^{sX}] = \sum_{k=0}^{\infty} e^{sk} \cdot e^{-\lambda} \frac{\lambda^k}{k!} \\
 &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^s \lambda)^k}{k!} \\
 &= e^{-\lambda} e^{\lambda e^s} \quad (\text{using the series expansion for } e^z) \\
 &= e^{\lambda(e^s - 1)}
 \end{aligned}$$

### Result

The transform for a  $\text{Poisson}(\lambda)$  random variable is  $M_X(s) = e^{\lambda(e^s - 1)}$ .

## Example: Transform of an Exponential RV

Recall PDF is  $f_X(x) = \lambda e^{-\lambda x}$  for  $x \geq 0$ .

$$\begin{aligned} M_X(s) &= E[e^{sx}] = \int_0^{\infty} e^{sx} \lambda e^{-\lambda x} dx \\ &= \lambda \int_0^{\infty} e^{(s-\lambda)x} dx \\ &= \lambda \left[ \frac{e^{(s-\lambda)x}}{s-\lambda} \right]_0^{\infty} \end{aligned}$$

This integral converges only if  $s - \lambda < 0$ , i.e.,  $s < \lambda$ . In that case:

$$M_X(s) = \lambda \left( 0 - \frac{1}{s-\lambda} \right) = \frac{\lambda}{\lambda-s}$$

### Result

The transform for an  $\text{Exponential}(\lambda)$  is  $M_X(s) = \frac{\lambda}{\lambda-s}$ , defined for  $s < \lambda$ .

## Example: Transform of a Geometric RV

Recall PMF is  $p_X(k) = p(1 - p)^{k-1}$  for  $k \geq 1$ .

$$M_X(s) = E[e^{sX}] =$$

# From Transforms to Moments

The name “moment generating function” comes from the fact that the moments of  $X$  can be easily derived from its transform.

By differentiating the transform definition with respect to  $s$ :

$$\frac{d}{ds} M_X(s) = \frac{d}{ds} E[e^{sX}] = E \left[ \frac{d}{ds} e^{sX} \right] = E[Xe^{sX}]$$

Evaluating at  $s = 0$ :

$$\frac{dM_X(s)}{ds} \Big|_{s=0} = E[Xe^0] = E[X]$$

## Moment Generating Property

The  $n$ -th moment of  $X$  is the  $n$ -th derivative of the transform, evaluated at  $s = 0$ .

$$E[X^n] = \frac{d^n M_X(s)}{ds^n} \Big|_{s=0}$$

## Example: Moments of the Exponential

For  $X \sim \text{Exponential}(\lambda)$ , we have  $M_X(s) = \lambda(\lambda - s)^{-1}$ .

- **First Moment (Mean):**

$$\frac{dM_X(s)}{ds} = \lambda(-1)(\lambda - s)^{-2}(-1) = \frac{\lambda}{(\lambda - s)^2}$$

$$E[X] = \left. \frac{\lambda}{(\lambda - s)^2} \right|_{s=0} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

- **Second Moment:**

$$\frac{d^2M_X(s)}{ds^2} = \frac{d}{ds} \left( \frac{\lambda}{(\lambda - s)^2} \right) = \lambda(-2)(\lambda - s)^{-3}(-1) = \frac{2\lambda}{(\lambda - s)^3}$$

$$E[X^2] = \left. \frac{2\lambda}{(\lambda - s)^3} \right|_{s=0} = \frac{2\lambda}{\lambda^3} = \frac{2}{\lambda^2}$$

The variance is  $\text{var}(X) = E[X^2] - (E[X])^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$ .



## Example: MGF and Moments of the Bernoulli

# Inversion of Transforms

A crucial property of transforms is that they uniquely determine the distribution.

## Inversion Property

The transform  $M_X(s)$  associated with a random variable  $X$  uniquely determines the CDF of  $X$  (assuming  $M_X(s)$  is finite in an interval around  $s = 0$ ).

This means if two random variables have the same transform, they must have the same distribution.

In practice, we don't use complex inversion formulas. We find a transform and then look it up in a table of known transform-distribution pairs to identify the distribution. This is a "pattern matching" approach.

## Example: Inversion by Pattern Matching

The transform of a random variable  $Y$  is  $M_Y(s) = \frac{2}{3} \cdot \frac{6}{6-s} + \frac{1}{3} \cdot \frac{4}{4-s}$ . What is the PDF of  $Y$ ?

**Solution:** We recognize the terms in the sum.

- The term  $\frac{6}{6-s}$  is the transform of an exponential random variable with parameter  $\lambda_1 = 6$ .
- The term  $\frac{4}{4-s}$  is the transform of an exponential random variable with parameter  $\lambda_2 = 4$ .

The overall transform is a weighted sum (a “mixture”) of these two transforms. This implies that the PDF of  $Y$  is a mixture of the corresponding PDFs.

$Y$  is generated as follows: with probability 2/3, its value is drawn from an Exponential(6) distribution, and with probability 1/3, its value is drawn from an Exponential(4) distribution.

The PDF is:

$$f_Y(y) = \frac{2}{3}(6e^{-6y}) + \frac{1}{3}(4e^{-4y}) = 4e^{-6y} + \frac{4}{3}e^{-4y}, \quad \text{for } y \geq 0$$

# Sums of Independent Random Variables

One of the most powerful applications of transforms is in analyzing sums of independent random variables.

Let  $X$  and  $Y$  be independent, and let  $Z = X + Y$ .

$$\begin{aligned}M_Z(s) &= E[e^{sZ}] = E[e^{s(X+Y)}] = E[e^{sX}e^{sY}] \\&= E[e^{sX}]E[e^{sY}] \quad (\text{since } X, Y \text{ are independent}) \\&= M_X(s)M_Y(s)\end{aligned}$$

## Key Property

The transform of a sum of independent random variables is the product of their individual transforms.

$$M_{X_1+\dots+X_n}(s) = M_{X_1}(s) \cdots M_{X_n}(s)$$

## Example: Sum of Independent Poissons

Let  $X \sim \text{Poisson}(\lambda)$  and  $Y \sim \text{Poisson}(\mu)$  be independent. Let  $Z = X + Y$ . What is the distribution of  $Z$ ?

We use transforms:

$$M_X(s) = e^{\lambda(e^s - 1)} \quad \text{and} \quad M_Y(s) = e^{\mu(e^s - 1)}$$

The transform of the sum is the product:

$$M_Z(s) = M_X(s)M_Y(s) = e^{\lambda(e^s - 1)} \cdot e^{\mu(e^s - 1)} = e^{(\lambda + \mu)(e^s - 1)}$$

We recognize this as the transform of a Poisson random variable with parameter  $\lambda + \mu$ . By the uniqueness of transforms, we conclude that  $Z$  must be a Poisson random variable.

### Result

The sum of independent Poisson random variables is a Poisson random variable whose parameter is the sum of the individual parameters.

# Sum of a Random Number of Independent RVs

Transforms are also useful for sums where the number of terms is itself a random variable. Let  $Y = X_1 + \dots + X_N$ , where the  $X_i$  are i.i.d. and  $N$  is a random variable independent of the  $X_i$ .

Using the law of iterated expectations:

$$M_Y(s) = E[e^{sY}] = E[E[e^{sY}|N]]$$

Given  $N = n$ ,  $Y = X_1 + \dots + X_n$ , so  $E[e^{sY}|N = n] = (M_X(s))^n$ . The random variable  $E[e^{sY}|N]$  is therefore  $(M_X(s))^N$ .

$$\text{So, } M_Y(s) = E[(M_X(s))^N] = \sum_{n=0}^{\infty} (M_X(s))^n p_N(n).$$

This is the PMF of  $N$ , but with  $e^s$  replaced by  $M_X(s)$ .

## Formula for Random Sums

$$M_Y(s) = M_N(\log(M_X(s)))$$

## Example: Sum of Geometric Number of Exponentials

Let  $N \sim \text{Geometric}(p)$ , and let each  $X_i \sim \text{Exponential}(\lambda)$ . PDF of  $Y = \sum_{i=1}^N X_i$ ?

**Solution:** We have the individual transforms:

$$M_N(s) = \frac{pe^s}{1 - (1 - p)e^s} \quad \text{and} \quad M_X(s) = \frac{\lambda}{\lambda - s}$$

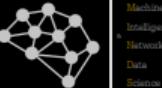
We find the transform of  $Y$  by starting with  $M_N(s)$  and replacing  $e^s$  with  $M_X(s)$ :

$$M_Y(s) = \frac{p \cdot M_X(s)}{1 - (1 - p)M_X(s)} = \frac{p \cdot \frac{\lambda}{\lambda - s}}{1 - (1 - p)\frac{\lambda}{\lambda - s}}$$

Multiplying numerator and denominator by  $(\lambda - s)$ :

$$= \frac{p\lambda}{(\lambda - s) - (1 - p)\lambda} = \frac{p\lambda}{\lambda - s - \lambda + p\lambda} = \frac{p\lambda}{p\lambda - s}$$

Transform of an exponential R.V with parameter  $p\lambda$ .



# Multivariate Transforms

The concept of a transform can be extended to multiple random variables to capture their joint distribution.

## Definition

The multivariate transform of random variables  $X_1, \dots, X_n$  is a function of  $n$  parameters  $s_1, \dots, s_n$ :

$$M_{X_1, \dots, X_n}(s_1, \dots, s_n) = E[e^{s_1 X_1 + \dots + s_n X_n}]$$

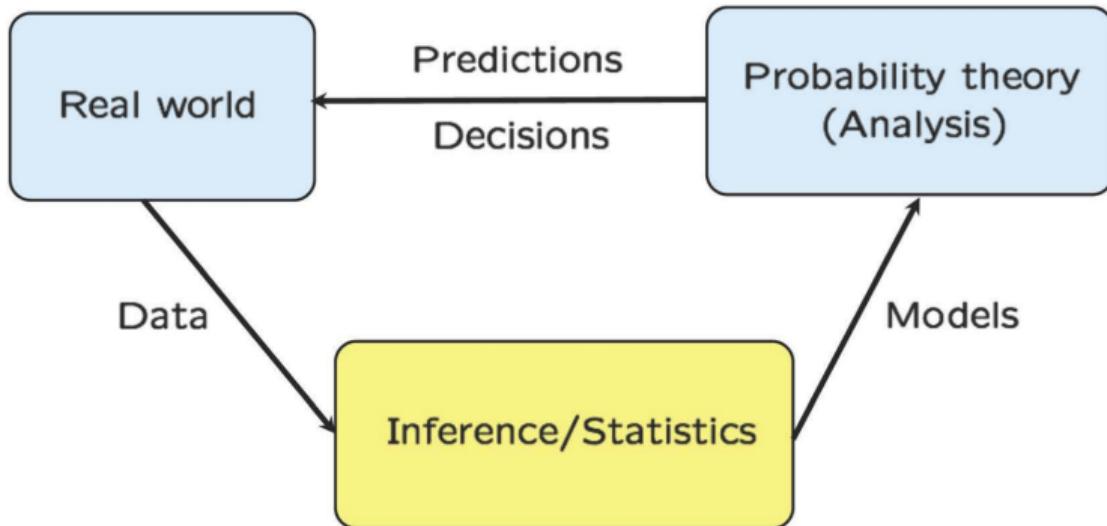
- The multivariate transform uniquely determines the joint distribution.
- It can be used to find moments and cross-moments (like  $E[X_1 X_2]$ ) through partial differentiation.
- It is a key tool for proving properties of jointly normal random variables.

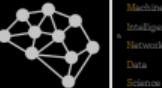
## **Lecture 16: Introduction to Bayesian Inference**

---

# Inference: The Big Picture

Inference is the process of building probabilistic models from real-world data. These models are then used within the framework of probability theory to make predictions and decisions about the real world.





# Inference Then and Now

## Then

Small data sets led to simple conclusions. For example:

- “10 patients were treated: 3 died”
- “10 patients were not treated: 5 died”
- “Therefore...”

## Now

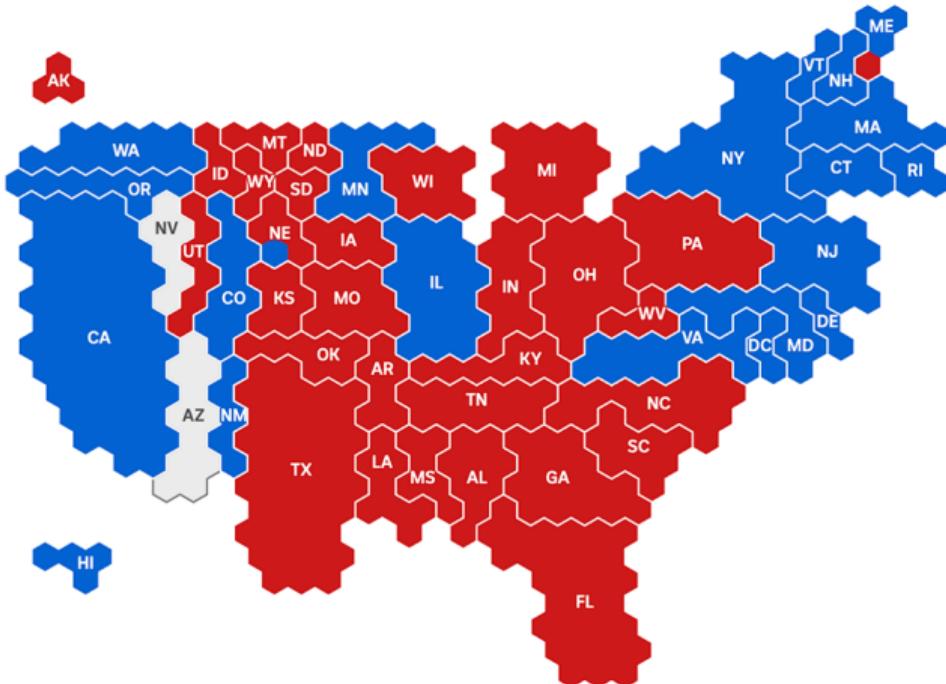
The modern era of inference is characterized by:

- Big data
- Big models
- Big computers

This allows for much more sophisticated and powerful analysis.

# A Sample of Application Domains

Inference is used in a vast range of fields to design and interpret experiments. One prominent example is political polling and election forecasting.



# A Sample of Application Domains

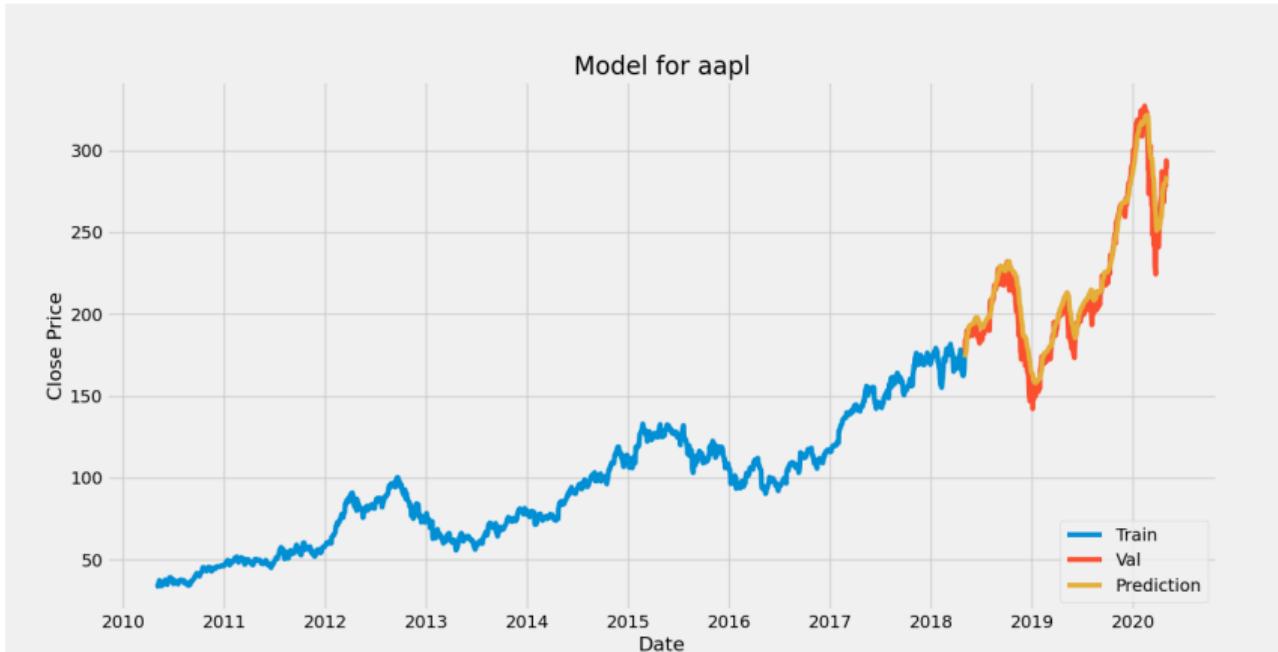
Bayesian inference methods are at the core of many modern technologies, including:

- Marketing and advertising
- Recommendation systems (e.g., the Netflix competition)

$$\begin{array}{c}
 \text{Item} \\
 \begin{array}{cccc} W & X & Y & Z \end{array} \\
 \begin{array}{c} A \\ B \\ C \\ D \end{array} \quad \left( \begin{array}{|c|c|c|c|} \hline & 4.5 & 2.0 & \\ \hline 4.0 & & 3.5 & \\ \hline & 5.0 & & 2.0 \\ \hline & 3.5 & 4.0 & 1.0 \\ \hline \end{array} \right) = \begin{array}{c} A \\ B \\ C \\ D \end{array} \times \begin{array}{c} W \\ X \\ Y \\ Z \end{array} \\
 \begin{array}{c} \text{Rating Matrix} \\ \text{User Matrix} \\ \text{Item Matrix} \end{array}
 \end{array}$$

# A Sample of Application Domains

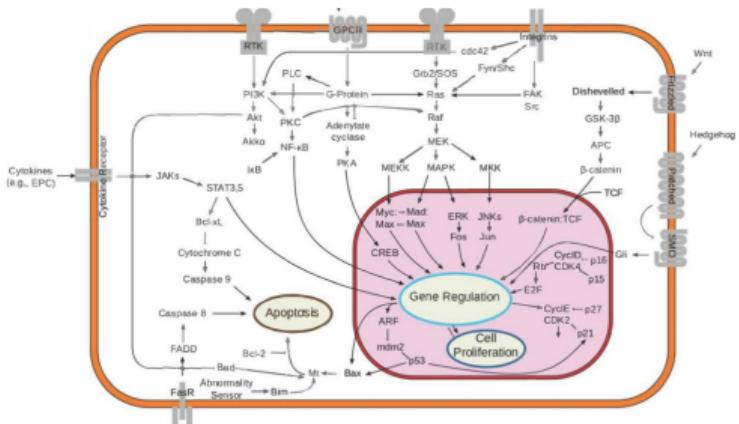
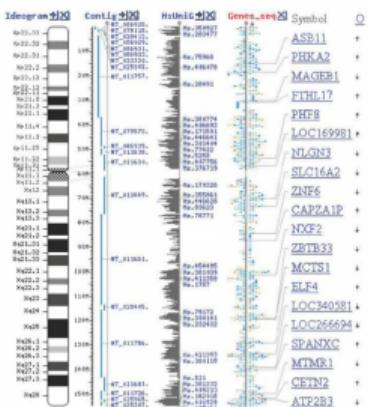
Inference is critical in finance for modeling asset prices, volatility, and risk.

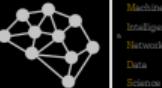


# A Sample of Application Domains

The life sciences heavily rely on statistical inference for areas such as:

- Genomics
- Systems biology
- Neuroscience

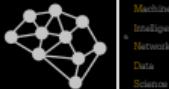




# A Sample of Application Domains

Inference is fundamental to a wide range of scientific and engineering disciplines:

- Modeling and monitoring the oceans
- Modeling and monitoring global climate
- Modeling and monitoring pollution
- Interpreting data from physics experiments
- Interpreting astronomy data



# A Sample of Application Domains

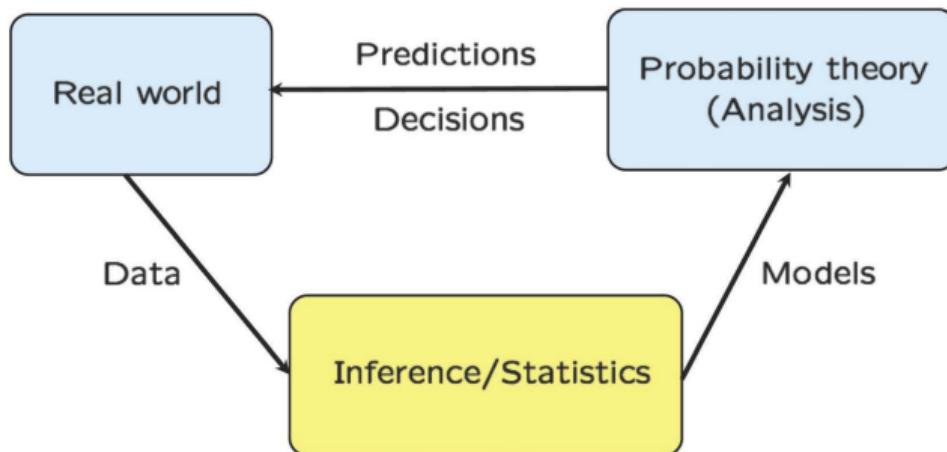
Signal processing is a field built on the principles of inference.

- Communication systems (dealing with noise)
- Speech processing and understanding
- Image processing and understanding
- Tracking of objects
- Positioning systems (e.g., GPS)
- Detection of abnormal events

# Model Building vs. Inferring Unobserved Variables

Inference problems can often be categorized into two types, using a simple signal model  $X = aS + W$  (Observation = scaling factor  $\times$  Signal + Noise).

- **Model Building:** We know the transmitted signal  $S$  and observe the received signal  $X$ . The goal is to infer the properties of the channel, represented by the unknown parameter  $a$ .
- **Variable Estimation:** We know the channel parameter  $a$  and observe the received signal  $X$ . The goal is to infer the value of the original, unobserved signal  $S$ .



# Hypothesis Testing vs. Estimation

## Hypothesis Testing

- The unknown quantity takes one of a few possible values.
- Example: Is an object in a radar image an airplane or a bird?
- The goal is to make a decision that minimizes the probability of being incorrect.

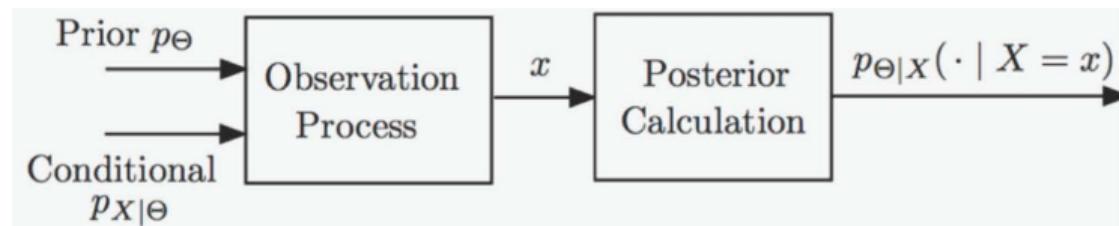
## Estimation

- The unknown is a numerical parameter or set of parameters.
- Example: What is the exact location of the airplane?
- The goal is to produce an estimate that is “close” to the true but unknown value.

# The Bayesian Inference Framework

The Bayesian approach treats the unknown quantity  $\Theta$  as a random variable.

- We start with a **prior distribution**,  $p_\Theta(\theta)$  or  $f_\Theta(\theta)$ , which represents our belief about  $\Theta$  before seeing any data.
- We have an **observation model**,  $p_{X|\Theta}(x|\theta)$  or  $f_{X|\Theta}(x|\theta)$ , which tells us the probability of observing data  $X$  for a given value of  $\Theta$ .
- After observing  $X = x$ , we use Bayes' rule to compute the **posterior distribution**,  $p_{\Theta|X}(\theta|x)$  or  $f_{\Theta|X}(\theta|x)$ . This represents our updated belief about  $\Theta$ .

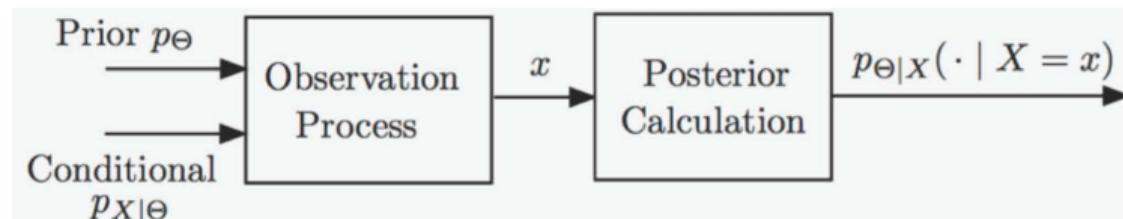


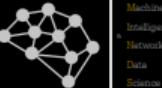
# The Output of Bayesian Inference

The primary output of a Bayesian inference problem is the full posterior distribution. It encapsulates all available information about the unknown quantity.

From this posterior distribution, we can derive simpler summaries:

- **Point Estimates:** A single “best guess” for the value of  $\Theta$ .
- **Error Analysis:** Measures of how confident we are in our estimates.





# Point Estimates in Bayesian Inference

An **estimator**,  $\hat{\Theta} = g(X)$ , is a rule (a function) for generating a guess based on the observation  $X$ . The resulting guess,  $\hat{\theta} = g(x)$ , is the **estimate**.

Two common types of Bayesian point estimates are:

## Maximum a Posteriori Probability (MAP)

The MAP estimate is the value  $\theta^*$  that maximizes the posterior distribution.

$$\hat{\theta}_{MAP} = \arg \max_{\theta} p_{\Theta|X}(\theta|x) \quad \text{or} \quad \arg \max_{\theta} f_{\Theta|X}(\theta|x)$$

## Conditional Expectation (LMS)

The Least Mean Squares (LMS) estimate is the expected value of the posterior distribution.

$$\hat{\theta}_{LMS} = E[\Theta|X = x]$$

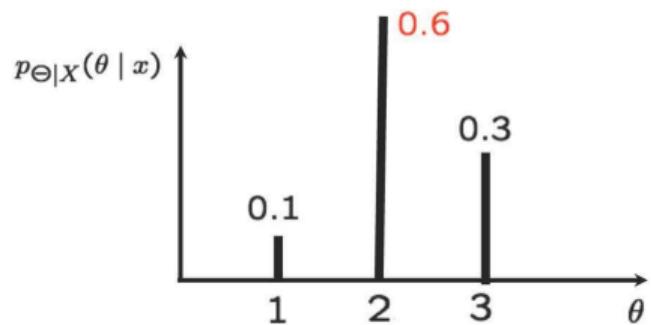
# Case 1: Discrete $\Theta$ , Discrete $X$

This is the classic hypothesis testing scenario.

## Bayes' Rule

$$p_{\Theta|X}(\theta|x) = \frac{p_{\Theta}(\theta)p_{X|\Theta}(x|\theta)}{p_X(x)} \quad \text{where} \quad p_X(x) = \sum_{\theta'} p_{\Theta}(\theta')p_{X|\Theta}(x|\theta')$$

- **MAP rule:** Choose the hypothesis  $\theta$  that maximizes the posterior  $p_{\Theta|X}(\theta|x)$ .
- **Conditional Probability of Error:** Given  $X = x$ , the probability of error for the MAP estimate  $\hat{\theta}$  is  $P(\hat{\theta} \neq \Theta|X = x) = 1 - p_{\Theta|X}(\hat{\theta}|x)$ . The MAP rule minimizes this for every  $x$ .
- **Overall Probability of Error:**  $P(\hat{\Theta} \neq \Theta) = \sum_x P(\hat{\Theta} \neq \Theta|X = x)p_X(x)$ .



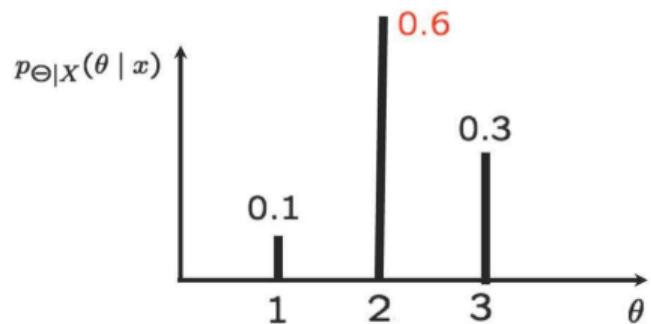
## Case 2: Discrete $\Theta$ , Continuous $X$

This case is common in signal detection, e.g., identifying a discrete transmitted signal  $\Theta$  from a noisy continuous observation  $X = \Theta + W$ .

### Bayes' Rule

$$p_{\Theta|X}(\theta|x) = \frac{p_{\Theta}(\theta)f_{X|\Theta}(x|\theta)}{f_X(x)} \quad \text{where} \quad f_X(x) = \sum_{\theta'} p_{\Theta}(\theta')f_{X|\Theta}(x|\theta')$$

The MAP rule and error probability calculations are analogous to the discrete-discrete case.  
 The MAP rule still minimizes the probability of error.



## Case 3: Continuous $\Theta$ , Continuous $X$

This is the classic parameter estimation scenario, such as estimating an unknown signal amplitude  $\Theta$  from a noisy measurement  $X = \Theta + W$ .

### Bayes' Rule

$$f_{\Theta|X}(\theta|x) = \frac{f_{\Theta}(\theta)f_{X|\Theta}(x|\theta)}{f_X(x)} \quad \text{where} \quad f_X(x) = \int f_{\Theta}(\theta')f_{X|\Theta}(x|\theta')d\theta'$$

Since the probability of any single point estimate being exactly correct is zero, we evaluate estimators using a different metric. The most common is the **mean squared error**:

- Conditional MSE:  $E[(\hat{\Theta} - \Theta)^2|X = x]$
- Overall MSE:  $E[(\hat{\Theta} - \Theta)^2]$

The LMS estimate  $E[\Theta|X = x]$  is the one that minimizes the MSE.

# Infering the Unknown Bias of a Coin

This is a key example in Bayesian inference. Let  $\Theta$  be the unknown bias of a coin, with a prior PDF  $f_\Theta(\theta)$ . We toss the coin  $n$  times and observe  $K = k$  heads.

## Bayes' Rule

$$f_{\Theta|K}(\theta|k) = \frac{f_\Theta(\theta)p_{K|\Theta}(k|\theta)}{p_K(k)} \quad \text{where} \quad p_K(k) = \int_0^1 f_\Theta(\theta')p_{K|\Theta}(k|\theta')d\theta'$$

and  $p_{K|\Theta}(k|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$ .

If we assume a uniform prior  $f_\Theta(\theta) = 1$  for  $\theta \in [0, 1]$ , the posterior is:

$$f_{\Theta|K}(\theta|k) = c \cdot \theta^k (1-\theta)^{n-k}$$

This is known as a **Beta distribution** with parameters  $(k+1, n-k+1)$ . If the prior is itself a Beta distribution, the posterior is also a Beta distribution (this is called a “conjugate prior”).

# Infering Coin Bias: Point Estimates

- **MAP Estimate:** We find the value of  $\theta$  that maximizes the posterior  $f_{\Theta|K}(\theta|k) \propto \theta^k(1-\theta)^{n-k}$ .  
 By taking the derivative with respect to  $\theta$  and setting to zero, we find:

$$\hat{\theta}_{MAP} = \frac{k}{n}$$

This is the intuitive sample mean.

- **LMS Estimate:** We need to calculate the mean of the Beta distribution. Using the formula  $\int_0^1 \theta^\alpha (1-\theta)^\beta d\theta = \frac{\alpha!\beta!}{(\alpha+\beta+1)!}$ :

$$E[\Theta|K=k] = \int_0^1 \theta \cdot f_{\Theta|K}(\theta|k) d\theta = \frac{\int_0^1 \theta^{k+1}(1-\theta)^{n-k} d\theta}{\int_0^1 \theta^k(1-\theta)^{n-k} d\theta} = \frac{(k+1)!(n-k)!/(n+2)!}{k!(n-k)!/(n+1)!} = \frac{k+1}{n+2}$$

$$\hat{\theta}_{LMS} = \frac{k+1}{n+2}$$

The LMS estimate is slightly different from the MAP, “pulling” the estimate away from 0 and 1.

# Summary

- **Problem Data:** A prior distribution  $p_\Theta$  or  $f_\Theta$ , and an observation model  $p_{X|\Theta}$  or  $f_{X|\Theta}$ .
- **Goal:** Given an observation  $X = x$ , use Bayes' rule to find the posterior distribution  $p_{\Theta|X}$  or  $f_{\Theta|X}$ .
- **Estimator vs. Estimate:** An estimator  $\hat{\Theta} = g(X)$  is a random variable; an estimate  $\hat{\theta} = g(x)$  is a number.
- **Common Point Estimates:**
  - MAP: Maximizes the posterior distribution.
  - LMS: The conditional expectation,  $E[\Theta|X = x]$ . Minimizes mean squared error.
- **Performance Evaluation:**
  - For hypothesis testing: Probability of error,  $P(\hat{\Theta} \neq \Theta)$ .
  - For estimation: Mean Squared Error,  $E[(\hat{\Theta} - \Theta)^2]$ .

## **Lecture 17: Linear Models With Normal Noise**

---

# Recognizing normal PDFs

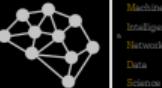
The normal  $X \sim N(\mu, \sigma^2)$  with PDF

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

$$c \cdot e^{-8(x-3)^2}$$

$$f_X(x) = c \cdot e^{-(\alpha x^2 + \beta x + \gamma)} \quad \alpha > 0$$

- Normal with mean  $-\beta/2\alpha$  and variance  $1/2\alpha$



# Estimating a normal random variable

In the presence of additive normal noise

$$X = \Theta + W, \quad \Theta, W \sim N(0, 1), \text{ independent}$$

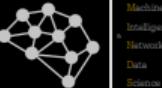
$$f_{X|\Theta}(x|\theta) :$$

$$f_{\Theta|X}(\theta|x) = \frac{f_{\Theta}(\theta)f_{X|\Theta}(x|\theta)}{f_X(x)}$$

$$f_X(x) = \int f_{\Theta}(\theta)f_{X|\Theta}(x|\theta)d\theta$$

$$\hat{\theta}_{MAP} = \hat{\theta}_{LMS} = E[\Theta|X = x] =$$

$$\hat{\Theta}_{MAP} = E[\Theta|X] =$$



# Estimating a normal random variable

in the presence of additive normal noise

$$X = \Theta + W, \quad \Theta, W \sim N(0, 1), \text{ independent}$$

$$f_{\Theta|X}(\theta|x) = \frac{f_{\Theta}(\theta)f_{X|\Theta}(x|\theta)}{f_X(x)}$$

$$f_X(x) = \int f_{\Theta}(\theta)f_{X|\Theta}(x|\theta)d\theta$$

$$\hat{\Theta}_{MAP} = \hat{\Theta}_{LMS} = \mathbb{E}[\Theta|X] = \frac{x}{2}$$

Even with general means and variances:

- posterior is normal
- LMS and MAP estimators coincide
- these estimators are “linear,” of the form  $\hat{\Theta} = aX + b$

# The case of multiple observations

$$X_1 = \Theta + W_1$$

⋮

$$X_n = \Theta + W_n$$

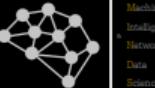
$$\Theta \sim N(x_0, \sigma_0^2) \quad W_i \sim N(0, \sigma_i^2)$$

$\Theta, W_1, \dots, W_n$  independent

$$f_{\Theta|X}(\theta|x) = \frac{f_{\Theta}(\theta)f_{X|\Theta}(x|\theta)}{f_X(x)} \quad f_X(x) = \int f_{\Theta}(\theta)f_{X|\Theta}(x|\theta)d\theta$$

$$f_{X_i|\Theta}(x_i|\theta) =$$

$$f_{X|\Theta}(x|\theta) =$$



# The case of multiple observations

$$f_{\Theta|X}(\theta|x) = c \cdot \exp\{-quad(\theta)\}$$

$$quad(\theta) = \frac{(\theta - x_0)^2}{2\sigma_0^2} + \frac{(\theta - x_1)^2}{2\sigma_1^2} + \dots + \frac{(\theta - x_n)^2}{2\sigma_n^2}$$

$$\hat{\theta}_{MAP} = \hat{\theta}_{LMS} = E[\Theta|X = x] = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

# The case of multiple observations

- Key conclusions:
  - posterior is normal
  - LMS and MAP estimates coincide
  - these estimates are “linear,” of the form  $\hat{\theta} = a_0 + a_1x_1 + \cdots + a_nx_n$
- Interpretations:
  - estimate  $\hat{\theta}$ : weighted average of  $x_0$  (prior mean) and  $x_i$  (observations)
  - weights determined by variances

$$\hat{\theta}_{MAP} = \hat{\theta}_{LMS} = E[\Theta|X = x] = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

# The mean squared error

$$f_{\Theta|X}(\theta|x) = c \cdot \exp\{-quad(\theta)\}$$

$$quad(\theta) = \frac{(\theta - x_0)^2}{2\sigma_0^2} + \frac{(\theta - x_1)^2}{2\sigma_1^2} + \dots + \frac{(\theta - x_n)^2}{2\sigma_n^2}$$

$$\hat{\theta} = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

- Performance measures:

- $\mathbb{E}[(\Theta - \hat{\Theta})^2 | X = x] = E[(\Theta - \hat{\theta})^2 | X = x] = var(\Theta | X = x) = 1 / \sum_{i=0}^n \frac{1}{\sigma_i^2}$
- $E[(\Theta - \hat{\Theta})^2] =$

$$f_X(x) = c \cdot e^{-(\alpha x^2 + \beta x + \gamma)} \quad \alpha > 0 \text{ Normal with mean } -\beta/2\alpha \text{ and variance } 1/2\alpha$$

# The mean squared error

$$E[(\Theta - \hat{\Theta})^2 | X = x] = E[(\Theta - \hat{\Theta})^2] = 1 / \sum_{i=0}^n \frac{1}{\sigma_i^2}$$

- conditional mean squared error same for all  $x$

- Example:  $\sigma_0^2 = \sigma_1^2 = \dots = \sigma_n^2 = \sigma^2$

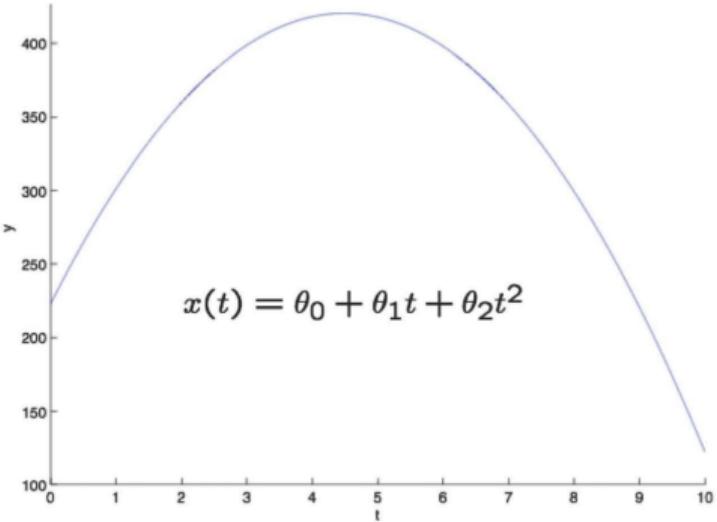
- Example:  $X = \Theta + W$
- $\Theta \sim N(0, 1)$ ,  $W \sim N(0, 1)$  independent
- $\hat{\Theta} = X/2$
- $E[(\Theta - \hat{\Theta})^2 | X = x] =$

$$\hat{\theta} = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

# The case of multiple parameters: trajectory estimation

$$x(t) = \theta_0 + \theta_1 t + \theta_2 t^2$$

- Random variables  $\Theta_0, \Theta_1, \Theta_2$ 
  - independent; priors  $f_{\Theta_j}$
- Measurements at times  $t_1, \dots, t_n$ 
  - $x_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$
- noise model:  $f_{W_i}$ 
  - independent  $W_i$ ; independent from  $\Theta_j$



# A model with normality assumptions

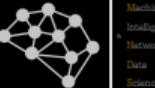
- $x_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i \quad i = 1, \dots, n$
- assume  $\Theta_j \sim N(0, \sigma_j^2)$ ,  $W_i \sim N(0, \sigma^2)$ ; independent

Given  $\Theta = \theta = (\theta_0, \theta_1, \theta_2)$ ,  $X_i$  is:

$$f_{X_i|\Theta}(x_i|\theta) = c \cdot \exp\{-(x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2 / 2\sigma^2\}$$

- $f_{\Theta|X}(\theta|x) = \frac{f_{\Theta}(\theta)f_{X|\Theta}(x|\theta)}{f_X(x)}$
- posterior:  $f_{\Theta|X}(\theta|x) =$

$$c(x) \exp\left\{-\frac{1}{2}\left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2\right\}$$



# A model with normality assumptions

$$f_{\Theta|x}(\theta|x) = c(x) \exp\left\{-\frac{1}{2}\left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2\right\}$$

- MAP estimate: maximize over  $(\theta_0, \theta_1, \theta_2)$ ;
  - (minimize quadratic function)

# Linear normal models

$\Theta_j$  and  $X_i$  and are linear functions of independent normal random variables

- $f_{\Theta|X}(\theta|x) = c(x) \exp\{-quadratic(\theta_1, \dots, \theta_m)\}$
- MAP estimate: maximize over  $(\theta_1, \dots, \theta_m)$ ;
  - (minimize quadratic function)
- $\hat{\Theta}_{MAP,j}$  linear function of  $X = (X_1, \dots, X_n)$

Facts:

- $\hat{\Theta}_{MAP,j} = E[\Theta_j|X]$
- marginal posterior PDF of  $\Theta_j$ :  $f_{\Theta_j|X}(\theta_j|x)$ , is normal
- MAP estimate based on the joint posterior PDF:
  - same as MAP estimate based on the marginal posterior PDF
- $E[(\hat{\Theta}_{i,MAP} - \Theta_i)^2|X = x]$  : same for all  $x$

# An illustration

Estimating the trajectory of a free-falling object

$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2),$$

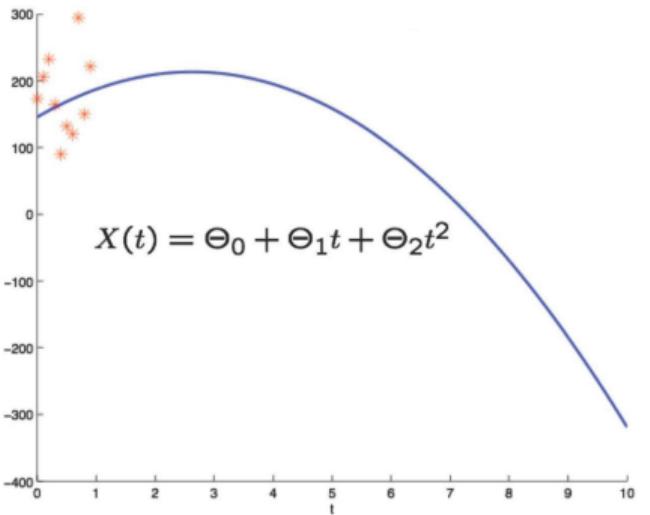
$$\Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$X(t) = \Theta_0 + \Theta_1 t + \Theta_2 t^2$$

$$x_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

minimize over  $\theta_0, \theta_1, \theta_2$

$$\frac{1}{2} \left( \frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2} \right) + \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2$$



# An illustration

Estimating the trajectory of a free-falling object

$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2),$$

$$\Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

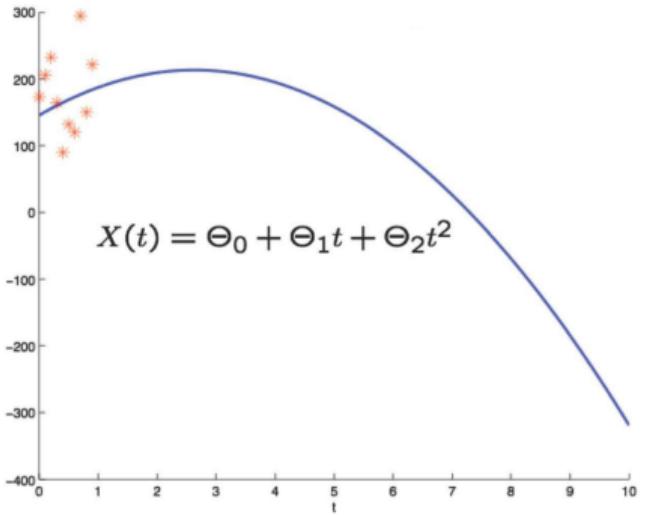
$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

minimize over  $\theta_0, \theta_1$

$$(\theta_0 - 200)^2 + (\theta_1 - 50)^2$$

$$+ \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$

$$X(t) = \Theta_0 + \Theta_1 t + \Theta_2 t^2$$



# An illustration

Estimating the trajectory of a free-falling object

$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2),$$

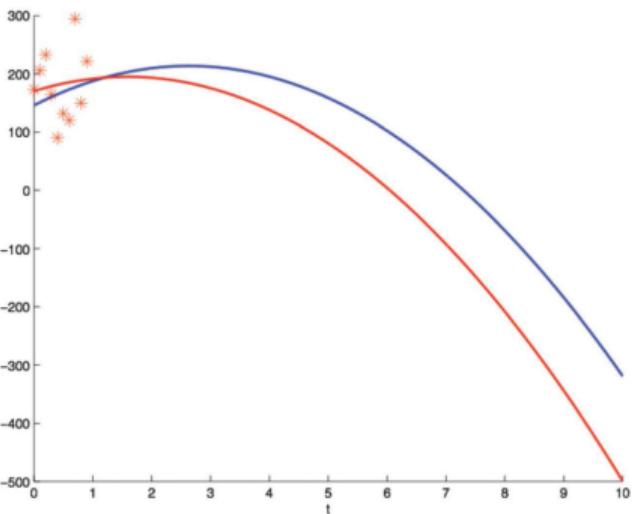
$$\Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$x_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

minimize over  $\theta_0, \theta_1$

$$(\theta_0 - 200)^2 + (\theta_1 - 50)^2$$

$$+ \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$



# An illustration

Estimating the trajectory of a free-falling object

$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2),$$

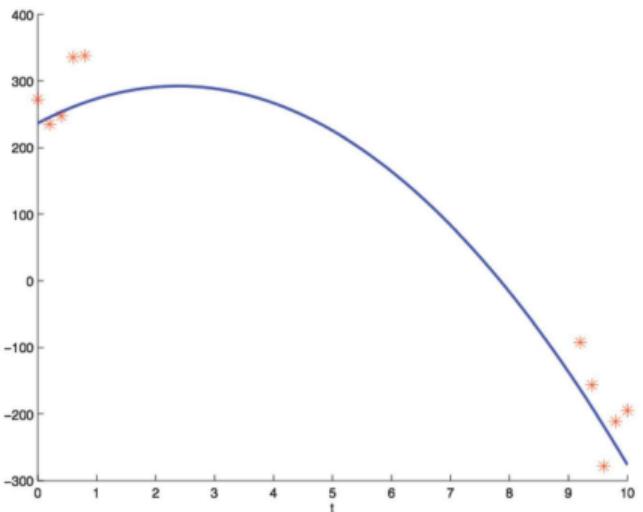
$$\Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$x_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

minimize over  $\theta_0, \theta_1$

$$(\theta_0 - 200)^2 + (\theta_1 - 50)^2$$

$$+ \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$



# An illustration

Estimating the trajectory of a free-falling object

$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2),$$

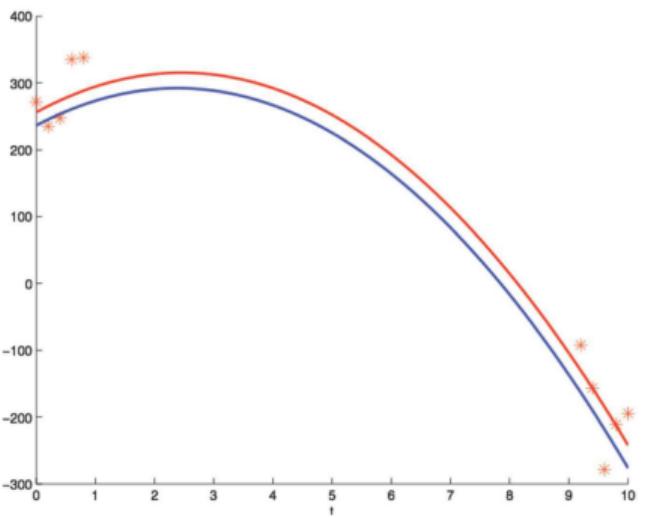
$$\Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$x_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

minimize over  $\theta_0, \theta_1$

$$(\theta_0 - 200)^2 + (\theta_1 - 50)^2$$

$$+ \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$



# An illustration

Estimating the trajectory of a free-falling object

$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2),$$

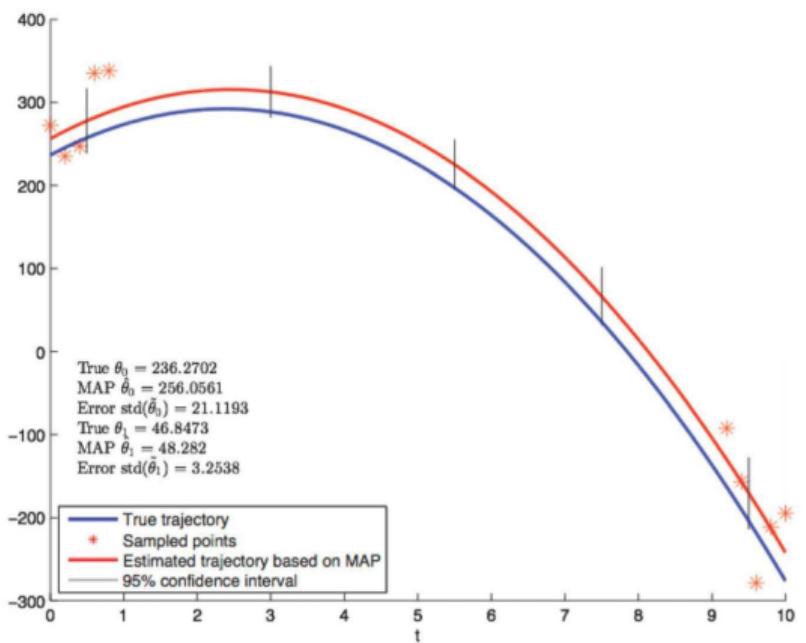
$$\Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$x_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

minimize over  $\theta_0, \theta_1$

$$(\theta_0 - 200)^2 + (\theta_1 - 50)^2$$

$$+ \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$



## Lecture 18: Least Mean Squares (LMS) Estimation

---

# LMS Estimation Without Observations

Consider estimating an unknown parameter  $\Theta$  when we only have its prior distribution  $f_{\Theta}(\theta)$  (or  $p_{\Theta}(\theta)$  for discrete cases) and no specific observations.

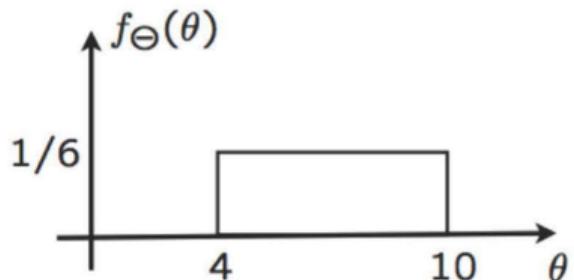
We are interested in finding a single point estimate  $\hat{\theta}$  that best represents  $\Theta$ .

Methods for choosing  $\hat{\theta}$ :

- MAP rule: Choose  $\hat{\theta}$  to maximize the prior  $f_{\Theta}(\theta)$ .
- Expectation: Choose  $\hat{\theta} = E[\Theta]$ .

LMS Criterion: Minimize the Mean Squared Error (MSE),  $E[(\Theta - \hat{\theta})^2]$ .

We want to find the value  $\hat{\theta}$  that minimizes this expectation.



# LMS Estimation Without Observations: Solution

- **Least Mean Squares Formulation:**

- We seek to find the constant  $\hat{\theta}$  that minimizes the MSE:  $\min_{\hat{\theta}} E[(\Theta - \hat{\theta})^2]$ .
- Let  $g(\hat{\theta}) = E[(\Theta - \hat{\theta})^2]$ . To minimize, we set the derivative to zero:

$$\frac{d}{d\hat{\theta}} E[(\Theta - \hat{\theta})^2] = E \left[ \frac{d}{d\hat{\theta}} (\Theta - \hat{\theta})^2 \right] = E[-2(\Theta - \hat{\theta})] = -2(E[\Theta] - \hat{\theta})$$

- Setting the derivative to zero gives  $-2(E[\Theta] - \hat{\theta}) = 0$ , which implies  $\hat{\theta} = E[\Theta]$ .
- The LMS estimate in the absence of observations is the prior mean.

- **Optimal Mean Squared Error:**

- The minimum possible MSE is achieved when  $\hat{\theta} = E[\Theta]$ .
- The minimum MSE value is  $E[(\Theta - E[\Theta])^2]$ , which is the definition of the variance of  $\Theta$ .
- Optimal MSE =  $\text{var}(\Theta)$ .

# LMS Estimation Based on Observation X

Now, suppose we have an observation  $X$  related to the unknown parameter  $\Theta$ .

- We have the prior  $f_\Theta(\theta)$ .
- We have a model for the observation, the likelihood  $f_{X|\Theta}(x|\theta)$  (or  $p_{X|\Theta}(x|\theta)$ ).
- We observe a specific value  $X = x$ .
- Minimizing overall MSE:  $E[(\Theta - \hat{\theta})^2]$  without using  $x$  leads to  $\hat{\theta} = E[\Theta]$ .
- Minimizing *conditional* MSE:  $E[(\Theta - \hat{\theta})^2|X = x]$  after observing  $x$ .

**LMS Estimate Definition:** The estimate  $\hat{\theta}$  that minimizes the conditional MSE  $E[(\Theta - \hat{\theta})^2|X = x]$  is the conditional expectation:

$$\hat{\theta}_{\text{LMS}} = E[\Theta|X = x]$$

The **LMS Estimator** is the function of the random variable  $X$  that gives the estimate:

$$\hat{\Theta}_{\text{LMS}} = E[\Theta|X]$$

# Optimality Properties of LMS Estimation

Recap of minimization results:

- $E[\Theta]$  minimizes the overall MSE  $E[(\Theta - c)^2]$  over all possible constant estimates  $c$ .
- $E[\Theta|X = x]$  minimizes the conditional MSE  $E[(\Theta - c)^2|X = x]$  over all possible constant estimates  $c$ , given the specific observation  $x$ .

A more general optimality property:

- The LMS estimator  $\hat{\Theta}_{\text{LMS}} = E[\Theta|X]$  minimizes the overall MSE  $E[(\Theta - g(X))^2]$  among *all possible estimators*  $g(X)$  that are functions of the observation  $X$ .
- That is,  $E[\Theta|X]$  is the function  $g(X)$  that is “closest” to  $\Theta$  in the mean squared error sense.



LMS estimate (given  $X = x$ ):  $\hat{\theta} = E[\Theta|X = x]$ , LMS estimator (function of  $X$ ):  $\hat{\Theta} = E[\Theta|X]$

- **Conditional MSE:** Expected performance after obtaining a specific measurement  $x$ .

- This is the minimum value of the conditional MSE criterion:

$$\text{MSE}_{\text{cond}}(x) = E[(\Theta - E[\Theta|X = x])^2 | X = x]$$

- By definition, this is the variance of the posterior (conditional) distribution of  $\Theta$  given  $X = x$ :

$$\text{MSE}_{\text{cond}}(x) = \text{var}(\Theta|X = x)$$

- Note: This value generally depends on the observed value  $x$ .
- **Overall MSE:** Expected performance of the estimator design *before* making an observation.
  - This is the expectation of the conditional MSE over all possible values of  $X$ :

$$\text{MSE}_{\text{overall}} = E[(\Theta - E[\Theta|X])^2] = E[\text{var}(\Theta|X)]$$

- This averages the conditional performance across all possible outcomes of the observation.

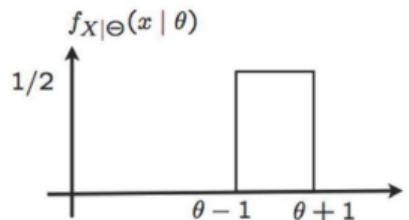
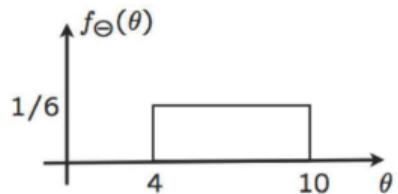
# Relationship Between LMS and MAP

- LMS estimation is focused on minimizing squared error, making it suitable for estimation problems where the magnitude of errors matters quadratically. It is generally not directly used for hypothesis testing.
- When does the LMS estimate  $\hat{\theta}_{\text{LMS}} = E[\Theta|X = x]$  coincide with the MAP estimate  $\hat{\theta}_{\text{MAP}}$  (which maximizes  $f_{\Theta|X}(\theta|x)$ )?
  - They are the same if the posterior distribution  $f_{\Theta|X}(\theta|x)$  is unimodal (has a single peak) and is symmetric around its mean.
  - A key example is when the posterior distribution is normal. In this case, the mean, median, and mode are all identical.
  - Recall from the previous lecture that linear models with normal priors and normal noise result in normal posteriors. Therefore, in “linear-normal” models, LMS and MAP estimates coincide.

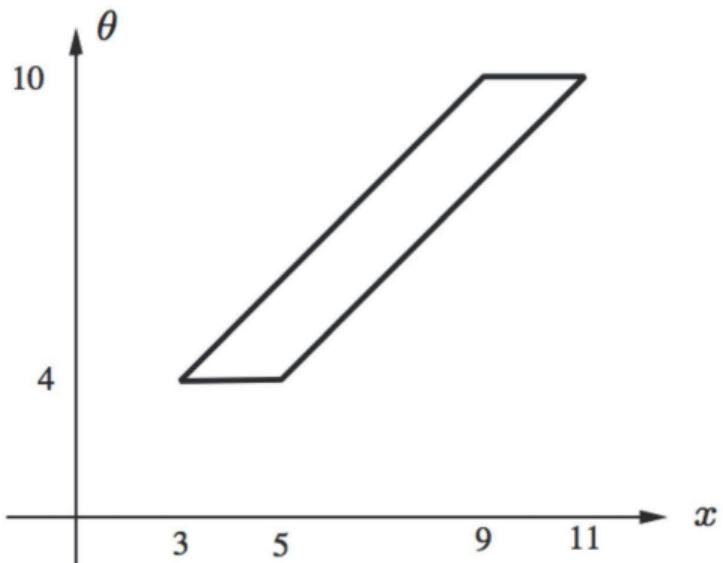


## Example: Uniform Prior and Uniform Noise

Prior:  $f_{\Theta}(\theta) = 1/6$  for  $4 \leq \theta \leq 10$ .



Likelihood:  $f_{X|\Theta}(x|\theta) = 1/2$  for  $\theta - 1 \leq x \leq \theta + 1$ .



Joint:  $(X, \Theta)$  is uniform over the parallelogram defined by  $4 \leq \theta \leq 10$  and  $\theta - 1 \leq x \leq \theta + 1$ .

## Example: Calculating the LMS Estimate

To find  $\hat{\Theta}_{LMS} = E[\Theta|X = x]$ , we need the posterior PDF  $f_{\Theta|x}(\theta|x)$ .

$$f_{\Theta|x}(\theta|x) = \frac{f_{\Theta,x}(\theta, x)}{f_X(x)}$$

The joint PDF  $f_{\Theta,x}(\theta, x) = f_\Theta(\theta)f_{X|\Theta}(x|\theta)$  is constant ( $1/6 \times 1/2 = 1/12$ ) over the parallelogram, and 0 elsewhere.

To find  $f_{\Theta|x}(\theta|x)$ , we fix  $x$  and consider the shape of the joint PDF along the vertical line at  $x$ .

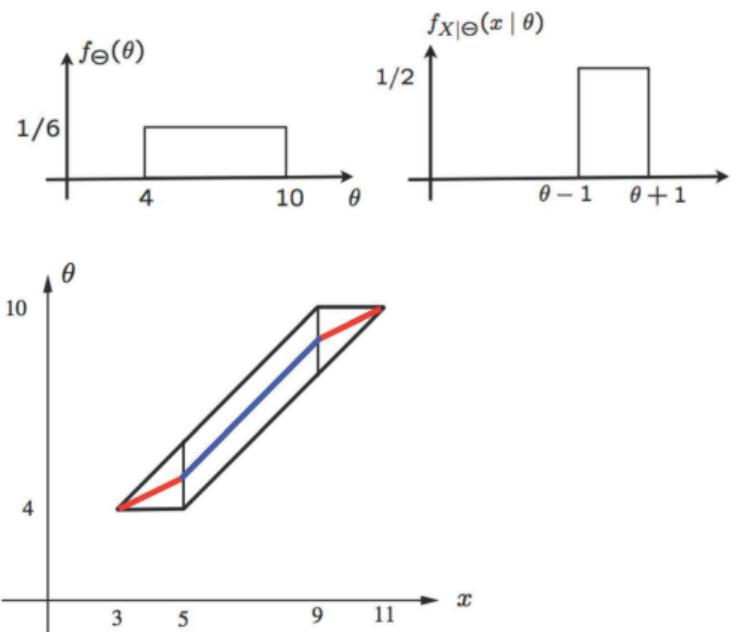
- If  $3 < x < 5$ :  $\theta$  must satisfy  $4 \leq \theta \leq x + 1$ .  $f_{\Theta|x}(\theta|x)$  is uniform on  $[4, x + 1]$ .
- If  $5 \leq x \leq 9$ :  $\theta$  must satisfy  $x - 1 \leq \theta \leq x + 1$ .  $f_{\Theta|x}(\theta|x)$  is uniform on  $[x - 1, x + 1]$ .
- If  $9 < x < 11$ :  $\theta$  must satisfy  $x - 1 \leq \theta \leq 10$ .  $f_{\Theta|x}(\theta|x)$  is uniform on  $[x - 1, 10]$ .

The LMS estimate is the mean of the conditional (posterior) uniform distribution:

- If  $3 < x < 5$ :  $E[\Theta|X = x] = \frac{4+(x+1)}{2} = \frac{x+5}{2}$ .
- If  $5 \leq x \leq 9$ :  $E[\Theta|X = x] = \frac{(x-1)+(x+1)}{2} = x$ .
- If  $9 < x < 11$ :  $E[\Theta|X = x] = \frac{(x-1)+10}{2} = \frac{x+9}{2}$ .

## Example: Conditional Mean Squared Error

The conditional MSE is  $\text{var}(\Theta|X = x)$ . Since the posterior distribution is uniform over an interval  $[a, b]$ , its variance is  $\frac{(b-a)^2}{12}$ .



- If  $3 < x < 5$ : Uniform on  $[4, x + 1]$ . Length  $b - a = (x + 1) - 4 = x - 3$ . Variance =  $\frac{(x-3)^2}{12}$ .
- If  $5 \leq x \leq 9$ : Uniform on  $[x - 1, x + 1]$ . Length  $b - a = (x + 1) - (x - 1) = 2$ . Variance =  $\frac{2^2}{12} = \frac{4}{12} = \frac{1}{3}$ .
- If  $9 < x < 11$ : Uniform on  $[x - 1, 10]$ . Length  $b - a = 10 - (x - 1) = 11 - x$ . Variance =  $\frac{(11-x)^2}{12}$ .

# LMS with Multiple Observations or Unknowns

The LMS principle extends directly to more complex scenarios.

- We have an unknown parameter  $\Theta$  (scalar or vector) with prior  $p_\Theta(\theta)$ .
- We have multiple observations  $X = (X_1, X_2, \dots, X_n)$  with a model  $p_{X|\Theta}(x|\theta)$ .
- We observe the specific vector  $X = x$ .
- The relevant probability space is now conditioned on the event  $\{X = x\}$ .

**LMS Estimate:** The estimate minimizing the conditional MSE  $E[(\Theta - \hat{\theta})^2 | X = x]$  is the conditional expectation:

$$\hat{\theta}_{\text{LMS}} = E[\Theta | X_1 = x_1, \dots, X_n = x_n]$$

If  $\Theta = (\Theta_1, \dots, \Theta_m)$  is a vector of unknown parameters, the LMS estimate  $\hat{\Theta} = (\hat{\Theta}_1, \dots, \hat{\Theta}_m)$  is found by applying the principle to each component separately:

$$\hat{\Theta}_{j,\text{LMS}} = E[\Theta_j | X = x]$$

We must calculate the conditional expectation for each parameter given all the observations.

# Properties of the Estimation Error I

Let  $\hat{\Theta} = E[\Theta|X]$  be the LMS estimator. Define the estimation error as  $\tilde{\Theta} = \Theta - \hat{\Theta}$ .

- **1. Conditional Error Mean is Zero:** The expected error, given the observation, is zero.

$$E[\tilde{\Theta}|X = x] = E[\Theta - \hat{\Theta}|X = x] = E[\Theta|X = x] - E[\hat{\Theta}|X = x]$$

Since  $\hat{\Theta} = E[\Theta|X]$  is a function of  $X$ ,  $E[\hat{\Theta}|X = x] = \hat{\Theta}|_{X=x} = E[\Theta|X = x]$ .

$$E[\tilde{\Theta}|X = x] = E[\Theta|X = x] - E[\Theta|X = x] = 0$$

This holds for any specific value  $x$ , so  $E[\tilde{\Theta}|X] = 0$ .

## Properties of the Estimation Error II

- **2. Orthogonality Principle:** The error  $\tilde{\Theta}$  is uncorrelated with the estimator  $\hat{\Theta}$ . More strongly, the error is uncorrelated with any function  $h(X)$  of the observations.

$$E[\tilde{\Theta}h(X)] = 0$$

Proof (using iterated expectations):

$$E[\tilde{\Theta}h(X)] = E[E[\tilde{\Theta}h(X)|X]] = E[h(X)E[\tilde{\Theta}|X]] = E[h(X) \cdot 0] = 0.$$

Setting  $h(X) = \hat{\Theta}$  shows  $E[\tilde{\Theta}\hat{\Theta}] = 0$ .

If  $E[\tilde{\Theta}] = 0$  (which it is,  $E[\tilde{\Theta}] = E[E[\tilde{\Theta}|X]] = E[0] = 0$ ) and  $E[\hat{\Theta}]$  exists, this implies  $\text{cov}(\tilde{\Theta}, \hat{\Theta}) = 0$ .

## Properties of the Estimation Error III

- **3. Variance Decomposition (Law of Total Variance):** The variance of the original parameter  $\Theta$  can be decomposed:

$$\text{var}(\Theta) = E[\text{var}(\Theta|X)] + \text{var}(E[\Theta|X])$$

Substituting  $\hat{\Theta} = E[\Theta|X]$  and  $E[\text{var}(\Theta|X)] = E[(\Theta - \hat{\Theta})^2] = \text{MSE}_{\text{overall}}$ :

$$\text{var}(\Theta) = E[(\Theta - \hat{\Theta})^2] + \text{var}(\hat{\Theta})$$

Or:  $\text{var}(\Theta) = \text{var}(\tilde{\Theta}) + \text{var}(\hat{\Theta})$ .

This shows that the variance of the estimator is always less than or equal to the variance of the original parameter.

## **Lecture 19: Linear Least Mean Squares (LLMS) Estimation**

---

# Challenges in LMS Estimation

- **Model Accuracy:** The calculation relies on having the correct prior  $f_{\Theta}(\theta)$  and likelihood model  $f_{X|\Theta}(x|\theta)$ . In reality, these models might be approximations or partially unknown.

$$f_{\Theta|x}(\theta|x) = \frac{f_{\Theta}(\theta)f_{X|\Theta}(x|\theta)}{f_X(x)}, \quad f_X(x) = \int f_{\Theta}(\theta')f_{X|\Theta}(x|\theta')d\theta'$$

Errors in the assumed models will lead to a suboptimal estimate.

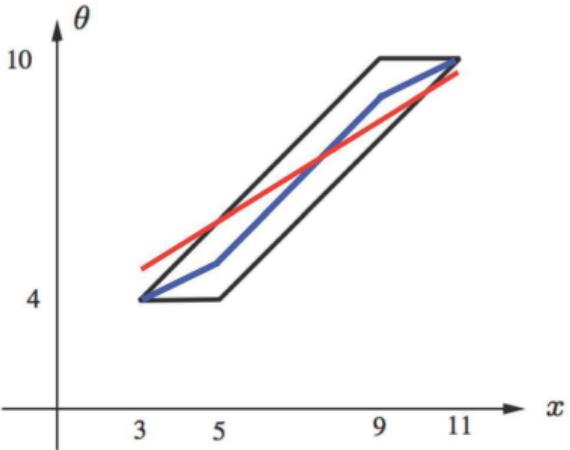
- **Computational Complexity:**

- Calculating the posterior PDF  $f_{\Theta|x}(\theta|x)$  often requires computing the normalization constant  $f_X(x)$ , which involves integration (potentially high-dimensional if  $\Theta$  is a vector).
  - Calculating the conditional expectation  $E[\Theta|X = x] = \int \theta f_{\Theta|x}(\theta|x)d\theta$  requires another integration step.
  - These integrations can be analytically intractable for complex models, requiring numerical methods (e.g., Monte Carlo simulations).
- **Implementation and Analysis:** The resulting estimator  $\hat{\Theta} = E[\Theta|X]$  might be a complex, nonlinear function of  $X$ , making its analysis difficult. Linear-normal models are an exception where the estimator is simple (linear) and analysis is easier.

# LLMS Formulation

- Unknown parameter  $\Theta$ , observation  $X$ .
- Overall Goal: Minimize MSE  $E[(\hat{\Theta} - \Theta)^2]$ .
- General estimators:  $\hat{\Theta} = g(X)$ .
- Optimal general estimator:  $\hat{\Theta}_{LMS} = E[\Theta|X]$ .
- **LLMS:** Consider only linear estimators  

$$\hat{\Theta} = aX + b.$$
- Find best  $a, b$   
 by minimizing  $E[(\Theta - (aX + b))^2]$ .
- If  $E[\Theta|X]$  happens to be linear in  $X$ , then the optimal LMS estimator is already linear, so  $\hat{\Theta}_{LMS} = \hat{\Theta}_{LLMS}$ .



# Solution to the LLMS Problem

Objective: Minimize  $J(a, b) = E[(\Theta - aX - b)^2]$  with respect to  $a$  and  $b$ .

Steps:

1. Take partial derivatives:  $\frac{\partial J}{\partial b} = 0$  and  $\frac{\partial J}{\partial a} = 0$ .
2. Solving  $\frac{\partial J}{\partial b} = 0$ :  $E[-2(\Theta - aX - b)] = 0 \implies E[\Theta] - aE[X] - b = 0$ .  $b = E[\Theta] - aE[X]$ . (Optimal  $b$  depends on optimal  $a$ ).
3. Substitute  $b$  back into  $J(a, b)$  and solve  $\frac{\partial J}{\partial a} = 0$ :

$$J(a) = E[(\Theta - aX - (E[\Theta] - aE[X]))^2] = E[((\Theta - E[\Theta]) - a(X - E[X]))^2].$$

$$\frac{\partial J}{\partial a} = E[-2(X - E[X])((\Theta - E[\Theta]) - a(X - E[X]))] = 0.$$

$$E[(X - E[X])(\Theta - E[\Theta])] - aE[(X - E[X])^2] = 0. \text{ Cov}(\Theta, X) - a\text{var}(X) = 0. \text{ Optimal } a = \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}.$$

# Solution to the LLMS Problem

Optimal Linear Estimator  $\hat{\Theta}_L$  (or  $\hat{\Theta}_{LLMS}$ ): Substitute optimal  $a$  back into expression for  $b$ :

$$b = E[\Theta] - \frac{\text{Cov}(\Theta, X)}{\text{var}(X)} E[X]. \quad \hat{\Theta}_L = aX + b = \frac{\text{Cov}(\Theta, X)}{\text{var}(X)} X + E[\Theta] - \frac{\text{Cov}(\Theta, X)}{\text{var}(X)} E[X].$$

$$\hat{\Theta}_L = E[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)} (X - E[X])$$

Using correlation coefficient  $\rho = \frac{\text{Cov}(\Theta, X)}{\sigma_\Theta \sigma_X}$  and  $\text{var}(X) = \sigma_X^2$ :

$$\hat{\Theta}_L = E[\Theta] + \rho \frac{\sigma_\Theta}{\sigma_X} (X - E[X])$$

- **Depends only on second moments:** The LLMS estimator  $\hat{\Theta}_L$  depends only on the means ( $E[\Theta], E[X]$ ), variances ( $\text{var}(\Theta), \text{var}(X)$ ), and covariance ( $\text{Cov}(\Theta, X)$  or  $\rho$ ). No need for full distribution details.

## Remarks on the LLMS Solution

$$\hat{\Theta}_L = E[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - E[X]) = E[\Theta] + \rho \frac{\sigma_\Theta}{\sigma_X}(X - E[X])$$

- **Interpretation:** Starts with the prior mean  $E[\Theta]$  and adds a correction term proportional to how much  $X$  deviates from its mean  $E[X]$ . The scaling factor depends on the correlation and variances.
- If  $\rho > 0$  (positively correlated): If  $X > E[X]$ , estimate increases. If  $X < E[X]$ , estimate decreases.
- If  $\rho < 0$  (negatively correlated): If  $X > E[X]$ , estimate decreases.
- If  $\rho = 0$  (uncorrelated):  $\hat{\Theta}_L = E[\Theta]$ . Observation  $X$  provides no useful linear information.

## Remarks on the LLMS Solution

$$\hat{\Theta}_L = E[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - E[X]) = E[\Theta] + \rho \frac{\sigma_\Theta}{\sigma_X}(X - E[X])$$

**LLMS Error Variance (MSE):** The minimum MSE achieved by the linear estimator is:

$$E[(\Theta - \hat{\Theta}_L)^2] = (1 - \rho^2)\text{var}(\Theta)$$

- MSE reduction depends on  $|\rho|$ . Larger  $|\rho|$  means better estimation.
- If  $|\rho| = 1$  (perfect linear relationship): MSE = 0.  $\Theta$  can be perfectly determined linearly from  $X$ .
- If  $\rho = 0$ : MSE =  $\text{var}(\Theta)$ . No improvement over prior variance.

## Example Revisited: LLMS Calculation

Recall  $\Theta \sim U[4, 10]$  and  $X|\{\Theta = \theta\} \sim U[\theta - 1, \theta + 1]$ .

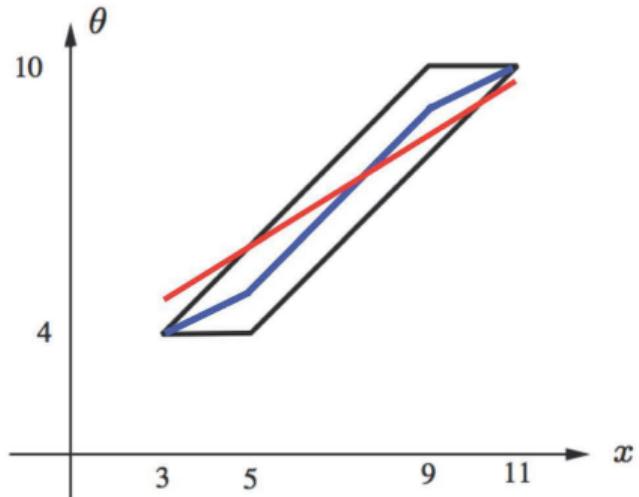
We need  $E[\Theta]$ ,  $E[X]$ ,  $\text{var}(\Theta)$ ,  $\text{var}(X)$ ,  $\text{Cov}(\Theta, X)$ .

- $E[\Theta] = (4 + 10)/2 = 7$ .
- $\text{var}(\Theta) = (10 - 4)^2/12 = 36/12 = 3$ . So  $\sigma_\Theta = \sqrt{3}$ .
- $E[X] = E[E[X|\Theta]]$ .

Since  $X|\Theta \sim U[\Theta - 1, \Theta + 1]$

$$E[X|\Theta] = \frac{(\Theta-1)+(\Theta+1)}{2} = \Theta.$$

$$E[X] = E[\Theta] = 7.$$



## Example Revisited: LLMS Calculation

- $\text{var}(X) = E[\text{var}(X|\Theta)] + \text{var}(E[X|\Theta]).$

$$\text{var}(X|\Theta) = \frac{((\Theta+1) - (\Theta-1))^2}{12} = \frac{2^2}{12} = \frac{1}{3}.$$

$$E[\text{var}(X|\Theta)] = E[1/3] = 1/3.$$

$$\text{var}(E[X|\Theta]) = \text{var}(\Theta) = 3.$$

$$\text{var}(X) = 1/3 + 3 = 10/3. \text{ So } \sigma_X = \sqrt{10/3}.$$

- $\text{Cov}(\Theta, X) = E[\Theta X] - E[\Theta]E[X].$

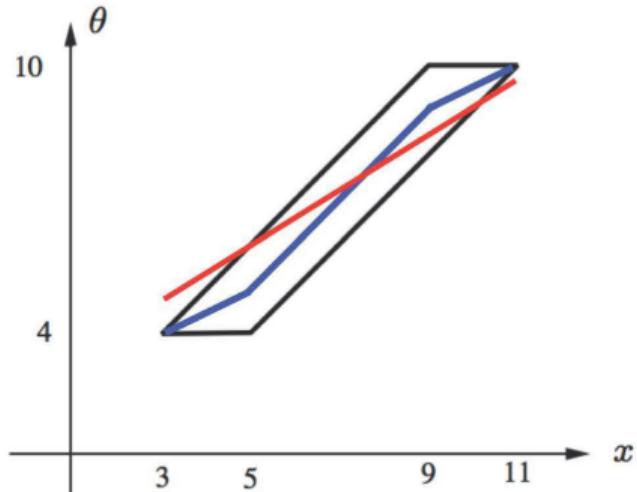
$$E[\Theta X] = E[E[\Theta X|\Theta]] = E[\Theta E[X|\Theta]] = E[\Theta \cdot \Theta] = E[\Theta^2].$$

$$E[\Theta^2] = \text{var}(\Theta) + (E[\Theta])^2 = 3 + 7^2 = 52.$$

$$\text{Cov}(\Theta, X) = 52 - (7)(7) = 52 - 49 = 3.$$

LLMS Estimator:

$$\hat{\Theta}_L = E[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - E[X]) = 7 + \frac{3}{10/3}(X - 7) = 7 + \frac{9}{10}(X - 7)$$



# LLMS for Inferring Coin Bias

- Unknown coin bias  $\Theta$ . Prior  $f_\Theta(\theta)$ .
- Experiment: Flip coin  $n$  times.
- Observation  $X$ : Number of heads obtained.  $X|\Theta \sim \text{Binomial}(n, \Theta)$ .
- Goal: Estimate  $\Theta$  based on  $X$ .
- Assume a uniform prior:  $\Theta \sim U[0, 1]$ .

LMS Estimator (derived previously using Beta posterior):

$$\hat{\Theta}_{LMS} = E[\Theta|X] = \frac{X + 1}{n + 2}$$

Notice that this LMS estimator is *already linear* in  $X$  ( $a = 1/(n+2)$ ,  $b = 1/(n+2)$ ).

Since the optimal LMS estimator is linear, it must also be the optimal *linear* LMS estimator.

Let's verify this using the LLMS formula:

$$\hat{\Theta}_{LLMS} = E[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - E[X])$$

# LLMS for Coin Bias: Moment Calculations

Need  $E[\Theta]$ ,  $\text{var}(\Theta)$ ,  $E[X]$ ,  $\text{var}(X)$ ,  $\text{Cov}(\Theta, X)$ .

- Prior  $\Theta \sim U[0, 1]$ :  $E[\Theta] = (0 + 1)/2 = 1/2$ .  $\text{var}(\Theta) = (1 - 0)^2/12 = 1/12$ .  
 $E[\Theta^2] = \text{var}(\Theta) + (E[\Theta])^2 = 1/12 + (1/2)^2 = 1/12 + 1/4 = 1/12 + 3/12 = 4/12 = 1/3$ .
- Conditional distribution  $X|\Theta \sim \text{Binomial}(n, \Theta)$ :  $E[X|\Theta] = n\Theta$ .  $\text{var}(X|\Theta) = n\Theta(1 - \Theta)$ .  
 $E[X^2|\Theta] = \text{var}(X|\Theta) + (E[X|\Theta])^2 = n\Theta(1 - \Theta) + (n\Theta)^2 = n\Theta - n\Theta^2 + n^2\Theta^2$ .
- Unconditional moments of  $X$ :  $E[X] = E[E[X|\Theta]] = E[n\Theta] = nE[\Theta] = n/2$ .  
 $\text{var}(X) = E[\text{var}(X|\Theta)] + \text{var}(E[X|\Theta])$ .  
 $E[\text{var}(X|\Theta)] = E[n\Theta(1 - \Theta)] = E[n\Theta - n\Theta^2] = nE[\Theta] - nE[\Theta^2] = n(1/2) - n(1/3) = n/6$ .  
 $\text{var}(E[X|\Theta]) = \text{var}(n\Theta) = n^2\text{var}(\Theta) = n^2/12$ .  
 $\text{var}(X) = n/6 + n^2/12 = 2n/12 + n^2/12 = \frac{n(n+2)}{12}$ .
- Covariance:  $\text{Cov}(\Theta, X) = E[\Theta X] - E[\Theta]E[X]$ .  
 $E[\Theta X] = E[E[\Theta X|\Theta]] = E[\Theta E[X|\Theta]] = E[\Theta(n\Theta)] = E[n\Theta^2] = nE[\Theta^2] = n/3$ .  
 $\text{Cov}(\Theta, X) = n/3 - (1/2)(n/2) = n/3 - n/4 = 4n/12 - 3n/12 = n/12$ .

# LLMS for Coin Bias: Final Result

Substitute moments into the LLMS formula:

$$\hat{\Theta}_{LLMS} = E[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - E[X])$$

We found:  $E[\Theta] = 1/2$ .  $\text{Cov}(\Theta, X) = n/12$ .  $\text{var}(X) = \frac{n(n+2)}{12}$ .  $E[X] = n/2$ .

$$\frac{\text{Cov}(\Theta, X)}{\text{var}(X)} = \frac{n/12}{n(n+2)/12} = \frac{1}{n+2}$$

$$\begin{aligned}\hat{\Theta}_{LLMS} &= \frac{1}{2} + \frac{1}{n+2}(X - \frac{n}{2}) = \frac{1}{2} + \frac{X}{n+2} - \frac{n}{2(n+2)} \\ &= \frac{n+2}{2(n+2)} + \frac{2X}{2(n+2)} - \frac{n}{2(n+2)} = \frac{(n+2) + 2X - n}{2(n+2)} = \frac{2X + 2}{2(n+2)} = \frac{X + 1}{n+2}\end{aligned}$$

This confirms  $\hat{\Theta}_{LLMS}$  is identical to  $\hat{\Theta}_{LMS}$  for this specific prior and likelihood.

# LLMS with Multiple Observations

- Unknown parameter  $\Theta$ . Observations  $X_1, \dots, X_n$ .
- Restrict to estimators that are linear combinations of the observations, plus a constant:

$$\hat{\Theta} = a_1 X_1 + \cdots + a_n X_n + b = \mathbf{a}^T \mathbf{X} + b$$

- Goal: Find coefficients  $a_1, \dots, a_n$  and constant  $b$  that minimize the MSE:

$$\min_{a_1, \dots, a_n, b} E[(\Theta - (a_1 X_1 + \cdots + a_n X_n + b))^2]$$

- Solution involves setting partial derivatives w.r.t.  $b$  and each  $a_i$  to zero. This results in a system of  $n + 1$  linear equations for  $b, a_1, \dots, a_n$ .
- The coefficients  $a_i$  and  $b$  depend only on the means, variances, and covariances involving  $\Theta$  and  $X_i$ . ( $E[\Theta], E[X_i], \text{var}(\Theta), \text{var}(X_i), \text{Cov}(\Theta, X_i), \text{Cov}(X_i, X_j)$ ).
- If  $E[\Theta|X]$  happens to be linear in  $X_1, \dots, X_n$ , then  $\hat{\Theta}_{LMS} = \hat{\Theta}_{LLMS}$ .
- If estimating multiple unknown parameters  $\Theta_j$ , apply the LLMS framework separately for each  $\Theta_j$ .

# Simplest LLMS Example with Multiple Observations

Model:  $X_i = \Theta + W_i, \quad i = 1, \dots, n$

- $\Theta$  has mean  $x_0$  and variance  $\sigma_0^2$ . (Note:  $x_0$  here is prior mean, not an observation).
- $W_i$  have mean 0 and variance  $\sigma_i^2$ .
- All variables  $\Theta, W_1, \dots, W_n$  are **uncorrelated**.

Case 1: Assume  $\Theta, W_1, \dots, W_n$  are independent and **normal**.

- From Lecture 17, we know the LMS estimator is:

$$\hat{\Theta}_{LMS} = E[\Theta|X] = \frac{\frac{x_0}{\sigma_0^2} + \sum_{i=1}^n \frac{X_i}{\sigma_i^2}}{\frac{1}{\sigma_0^2} + \sum_{i=1}^n \frac{1}{\sigma_i^2}}$$

- This estimator  $\hat{\Theta}_{LMS}$  is already a linear function of  $X_1, \dots, X_n$ .
- Therefore, it must be the best linear estimator as well:  $\hat{\Theta}_{LLMS} = \hat{\Theta}_{LMS}$ .

# Simplest LLMS Example with Multiple Observations

Case 2: Assume general distributions (not necessarily normal).

- Assume only the means, variances are the same as in the normal case, and the variables are uncorrelated.
- The LLMS solution only depends on these first and second moments (means, variances, covariances).
- Since all relevant moments are the same as in the normal case, the LLMS solution must be the same linear function found in Case 1.

$$\hat{\Theta}_{LLMS} = \frac{\frac{x_0}{\sigma_0^2} + \sum_{i=1}^n \frac{X_i}{\sigma_i^2}}{\frac{1}{\sigma_0^2} + \sum_{i=1}^n \frac{1}{\sigma_i^2}}$$

- In this non-normal case,  $\hat{\Theta}_{LLMS}$  is the best *linear* estimator, but it might not be the true LMS estimator  $E[\Theta|X]$  (which could be nonlinear).

# The Representation of Data Matters in LLMS

The restriction to linear estimators makes LLMS sensitive to how the data is represented.

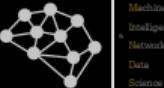
- **LMS:** The optimal estimator is  $E[\Theta|X]$ . Since the information contained in  $X$  is the same as the information in  $X^3$  (assuming  $X \mapsto X^3$  is invertible, or carefully handling non-invertibility), conditioning on  $X$  is equivalent to conditioning on  $X^3$ .

$$E[\Theta|X] \text{ is the same as } E[\Theta|X^3]$$

- **LLMS:** The class of estimators considered depends explicitly on the data representation.
  - LLMS based on  $X$ :  $\hat{\Theta} = aX + b$ . While LLMS based on  $X^3$ :  $\hat{\Theta}' = cX^3 + d$
  - Generally,  $\hat{\Theta} \neq \hat{\Theta}'$ . The best linear function of  $X$  is not necessarily the best linear function of  $X^3$ .
- **Extended Linear Models:** We can enhance LLMS by considering linearity in functions (features) of the data:
  - E.g.,  $\hat{\Theta} = a_1X + a_2X^2 + a_3X^3 + a_4e^X + a_5 \log X + b$ . (Linear in  $X, X^2, X^3, e^X, \log X$ ). crucial for the performance of LLMS if the true relationship (i.e.,  $E[\Theta|X]$ ) is nonlinear.

# Lecture 20: Inequalities, Convergence, and the Weak Law of Large Numbers

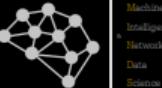
---



# The Markov Inequality: Concept

- Goal: Use minimal information (like the mean) about a distribution to bound probabilities of “extreme events”.
- Intuition: “If a non-negative random variable  $X$  has a small mean  $E[X]$ , then  $X$  is unlikely to take on very large values.”
- **Markov Inequality:** If  $X \geq 0$  and  $a > 0$ , then:

$$P(X \geq a) \leq \frac{E[X]}{a}$$



# The Markov Inequality: Examples

**Markov Inequality:** If  $X \geq 0$  and  $a > 0$ , then  $P(X \geq a) \leq \frac{E[X]}{a}$ .

- Example 1:  $X \sim \text{Exponential}(\lambda = 1)$ .
  - $X \geq 0$ .
  - $E[X] = 1/\lambda = 1$ .
  - Markov bound:  $P(X \geq a) \leq \frac{1}{a}$ .
  - (Exact probability:  $P(X \geq a) = e^{-a}$ ). The bound is simple but can be loose.
- Example 2:  $X \sim \text{Uniform}[-4, 4]$ .
  - $X$  is not non-negative. Markov inequality *cannot* be directly applied.
  - If we wanted  $P(X \geq 3)$ : Markov does not apply.
  - (Need other tools, or apply to  $|X|$  or similar if applicable).

# The Chebyshev Inequality: Concept

- Uses more information: Mean  $\mu$  and variance  $\sigma^2$ .
- Applicable to any random variable  $X$  with finite mean and variance (need not be non-negative).
- Intuition: “If the variance  $\sigma^2$  is small, then  $X$  is unlikely to be far from its mean  $\mu$ .”
- **Chebyshev Inequality:** For any  $c > 0$ :

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

- Proof idea: Apply Markov inequality to the non-negative random variable  $Y = (X - \mu)^2$ . Let  $a = c^2$ . Then  $Y \geq a$  if and only if  $|X - \mu| \geq c$ .  $P(|X - \mu| \geq c) = P((X - \mu)^2 \geq c^2)$ . By Markov:  $P((X - \mu)^2 \geq c^2) \leq \frac{E[(X - \mu)^2]}{c^2} = \frac{\sigma^2}{c^2}$ .

# The Chebyshev Inequality: Examples

**Chebyshev Inequality:**  $P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$ .

Alternative form (let  $c = k\sigma$  where  $k > 0$ ):

$$P(|X - \mu| \geq k\sigma) \leq \frac{\sigma^2}{(k\sigma)^2} = \frac{1}{k^2}$$

Probability of being  $k$  or more standard deviations away from the mean is at most  $1/k^2$ .

- Example:  $X \sim \text{Exponential}(\lambda = 1)$ .
  - $\mu = E[X] = 1$ .
  - $\sigma^2 = \text{var}(X) = 1/\lambda^2 = 1$ .
  - Chebyshev bound on  $P(X \geq a)$ ? Need to relate to  $P(|X - \mu| \geq c)$ . If  $a > \mu = 1$ , then  $X \geq a$  implies  $|X - 1| \geq a - 1$ .  $P(X \geq a) \leq P(|X - 1| \geq a - 1)$ . Using Chebyshev with  $c = a - 1$ :  
 $P(|X - 1| \geq a - 1) \leq \frac{\sigma^2}{(a-1)^2} = \frac{1}{(a-1)^2}$  (for  $a > 1$ ).
  - Compare: Markov:  $P(X \geq a) \leq 1/a$ . Chebyshev:  $P(X \geq a) \leq 1/(a-1)^2$ . Exact:  $P(X \geq a) = e^{-a}$ .
  - Chebyshev is often tighter than Markov, but still can be loose.

# The Weak Law of Large Numbers (WLLN)

- $X_1, X_2, \dots$  sequence of independent and identically distributed (i.i.d.) random variables.
- Assume finite mean  $E[X_i] = \mu$  and finite variance  $\text{var}(X_i) = \sigma^2$ .
- Sample Mean:  $M_n = \frac{X_1 + \dots + X_n}{n}$ .

Properties of the Sample Mean:

- Mean:  $E[M_n] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n}(n\mu) = \mu$ . (Sample mean is unbiased estimator of population mean).
- Variance:  $\text{var}(M_n) = \text{var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \text{var}\left(\sum_{i=1}^n X_i\right)$ . Due to independence:  
 $\text{var}(M_n) = \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$ . (Variance decreases as  $n$  increases).

Applying Chebyshev to  $M_n$ :  $M_n$  has mean  $\mu$  and variance  $\sigma^2/n$ . For any  $\epsilon > 0$ :

$$P(|M_n - \mu| \geq \epsilon) \leq \frac{\text{var}(M_n)}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

# The Weak Law of Large Numbers (WLLN)

$$P(|M_n - \mu| \geq \epsilon) \leq \frac{\text{var}(M_n)}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

**WLLN Statement:** As  $n \rightarrow \infty$ , the Chebyshev bound  $\frac{\sigma^2}{n\epsilon^2} \rightarrow 0$ . Since probability is non-negative:

$$\lim_{n \rightarrow \infty} P(|M_n - \mu| \geq \epsilon) = 0$$

For any small  $\epsilon > 0$ , the probability that the sample mean  $M_n$  is further than  $\epsilon$  away from the true mean  $\mu$  approaches zero as the sample size  $n$  grows.

# Interpreting the WLLN

WLLN:  $M_n = (X_1 + \dots + X_n)/n$ . For any  $\epsilon > 0$ ,  $P(|M_n - \mu| \geq \epsilon) \rightarrow 0$  as  $n \rightarrow \infty$ .

- **Scenario 1: Repeated Measurements**

- Experiment: Measure a fixed quantity  $\mu$ .
- $X_i = \mu + W_i$ , where  $W_i$  is i.i.d. measurement noise with  $E[W_i] = 0$ . Then  $E[X_i] = \mu$ .
- $M_n = \frac{1}{n} \sum (\mu + W_i) = \mu + \frac{1}{n} \sum W_i$ .
- WLLN implies  $P(|\mu + \frac{1}{n} \sum W_i - \mu| \geq \epsilon) = P(|\frac{1}{n} \sum W_i| \geq \epsilon) \rightarrow 0$ .
- The average measurement  $M_n$  becomes arbitrarily close to the true value  $\mu$  with high probability, as  $n$  increases. Averaging reduces noise.

- **Scenario 2: Repeated Independent Experiments**

- Experiment: Bernoulli trial (e.g., coin flip, event occurs/doesn't occur).
- Event  $A$ , with  $P(A) = p$ .
- $X_i = 1$  if  $A$  occurs on trial  $i$ ,  $X_i = 0$  otherwise. ( $X_i$  are i.i.d. Bernoulli( $p$ )). Then,  $E[X_i] = p$ .
- $M_n = \frac{\sum X_i}{n}$  is the fraction of times  $A$  occurred in  $n$  trials (empirical frequency).
- WLLN implies  $P(|M_n - p| \geq \epsilon) \rightarrow 0$ .
- The empirical frequency  $M_n$  converges to the true probability  $p$  (in a sense defined by WLLN).  
Connects probability theory to relative frequencies.

# The Pollster's Problem

- $p$ : True fraction of population that will vote “yes”. (Unknown).
- Randomly select  $n$  people for polling.
- $X_i = 1$  if person  $i$  says “yes”,  $X_i = 0$  if “no”. (Assume  $X_i \sim \text{Bernoulli}(p)$ , i.i.d.).
- $M_n = (X_1 + \dots + X_n)/n$  = fraction of “yes” in the sample. (Our estimate of  $p$ ).
- Goal: Ensure estimate  $M_n$  is close to true  $p$  with high probability.
- E.g., want error less than 1

Using Chebyshev/WLLN bound:  $P(|M_n - p| \geq \epsilon) \leq \frac{\text{var}(M_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$ . Here  $\mu = p$ ,  $\sigma^2 = \text{var}(X_i) = p(1-p)$ .  $\epsilon = 0.01$ .

$$P(|M_n - p| \geq 0.01) \leq \frac{p(1-p)}{n(0.01)^2}$$

Bound depends on unknown  $p$ . Worst case:  $p(1-p) = 1/4$  happens at  $p = 1/2$ .

$$P(|M_n - p| \geq 0.01) \leq \frac{1/4}{n(0.01)^2} = \frac{1}{4n(0.0001)} = \frac{1}{0.0004n} = \frac{2500}{n}$$

# Convergence “in Probability”

WLLN states:  $P(|M_n - \mu| \geq \epsilon) \rightarrow 0$  as  $n \rightarrow \infty$ , for any  $\epsilon > 0$ . This prompts a formal definition of convergence for sequences of random variables.

We want to say “ $M_n$  converges to  $\mu$ ”. What does “converges” mean here?

Consider a sequence of random variables  $Y_1, Y_2, \dots$  (not necessarily independent).

**Definition: Convergence in Probability** A sequence of random variables  $Y_n$  converges in probability to a number  $a$  if, for any  $\epsilon > 0$ :

$$\lim_{n \rightarrow \infty} P(|Y_n - a| \geq \epsilon) = 0$$

Notation:  $Y_n \xrightarrow{P} a$ .

WLLN can be restated as: If  $X_i$  are i.i.d. with mean  $\mu$  and finite variance, then the sample mean  $M_n$  converges in probability to  $\mu$ . ( $M_n \xrightarrow{P} \mu$ ). (Note: WLLN holds even if variance is infinite, but proof is harder).

Compare with ordinary convergence of a sequence of numbers  $a_n$ .

- **Ordinary Convergence:**  $a_n \rightarrow a$ .

- Meaning: “ $a_n$  eventually gets and stays arbitrarily close to  $a$ ”.
- Formal: For every  $\epsilon > 0$ , there exists  $n_0$  such that for all  $n \geq n_0$ , we have  $|a_n - a| \leq \epsilon$ .

- **Convergence in Probability:**  $Y_n \xrightarrow{P} a$ .

- Meaning: “Almost all of the probability mass (PMF/PDF) of  $Y_n$  eventually gets concentrated arbitrarily close to  $a$ ”.
- Formal: For every  $\epsilon > 0$ ,  $P(|Y_n - a| \geq \epsilon) \rightarrow 0$ .
- $Y_n$  can still take values far from  $a$ , but the probability of doing so becomes vanishingly small as  $n$  increases.

# Some Properties of Convergence in Probability

Let  $X_n \xrightarrow{P} a$  and  $Y_n \xrightarrow{P} b$ .

- **Sum:**  $X_n + Y_n \xrightarrow{P} a + b$ .
- **Product:**  $X_n Y_n \xrightarrow{P} ab$ .
- **Continuous Mapping Theorem:** If  $g$  is a continuous function, then  $g(X_n) \xrightarrow{P} g(a)$ . Example:  
 $X_n^2 \xrightarrow{P} a^2$ .
- **Expectation Limitation:** Convergence in probability  $X_n \xrightarrow{P} a$  does *not* necessarily imply that  $E[X_n] \rightarrow a$ . The expectation might not converge, or might converge to a different value.

# Convergence in Probability: Example (Expectation)

Consider a sequence  $Y_n$  with the following PMF:



$$p_{Y_n}(y) = \begin{cases} 1 - \frac{1}{n} & \text{if } y = 0 \\ \frac{1}{n} & \text{if } y = n^2 \\ 0 & \text{otherwise} \end{cases}$$

Does  $Y_n$  converge in probability to 0? Check the definition: For any  $\epsilon > 0$ :

$P(|Y_n - 0| \geq \epsilon) = P(Y_n \geq \epsilon)$ . If  $n$  is large enough such that  $n^2 \geq \epsilon$ , then the only way  $|Y_n| \geq \epsilon$  is if  $Y_n = n^2$ .  $P(|Y_n - 0| \geq \epsilon) = P(Y_n = n^2) = 1/n$ .

$$\lim_{n \rightarrow \infty} P(|Y_n - 0| \geq \epsilon) = \lim_{n \rightarrow \infty} 1/n = 0. \text{ Yes, } Y_n \xrightarrow{P} 0.$$

What about the expectation?  $E[Y_n] = 0 \cdot (1 - 1/n) + n^2 \cdot (1/n) = n$ .

$\lim_{n \rightarrow \infty} E[Y_n] = \lim_{n \rightarrow \infty} n = \infty$ . Here,  $Y_n \xrightarrow{P} 0$ , but  $E[Y_n] \rightarrow \infty$ . Convergence in probability does not imply convergence of expectations.

# Convergence in Probability: Example (Min)

- $X_1, X_2, \dots$  i.i.d.,  $X_i \sim \text{Uniform}[0, 1]$ .
- $Y_n = \min\{X_1, \dots, X_n\}$ .

Intuitively, as  $n$  increases, it's more likely one of the  $X_i$  will be very small, so  $Y_n$  should approach 0. Does  $Y_n \xrightarrow{P} 0$ ?

Check the definition: For any  $\epsilon > 0$  (assume  $0 < \epsilon < 1$ ):  $P(|Y_n - 0| \geq \epsilon) = P(Y_n \geq \epsilon)$ .  $Y_n \geq \epsilon$  if and only if  $X_i \geq \epsilon$  for ALL  $i = 1, \dots, n$ . Due to independence:

$$P(Y_n \geq \epsilon) = P(X_1 \geq \epsilon, \dots, X_n \geq \epsilon) = P(X_1 \geq \epsilon) \cdots P(X_n \geq \epsilon). \text{ For a single } X_i \sim U[0, 1], \\ P(X_i \geq \epsilon) = 1 - P(X_i < \epsilon) = 1 - \epsilon. \text{ So, } P(Y_n \geq \epsilon) = (1 - \epsilon)^n.$$

Now take the limit:  $\lim_{n \rightarrow \infty} P(|Y_n - 0| \geq \epsilon) = \lim_{n \rightarrow \infty} (1 - \epsilon)^n$ . Since  $0 < \epsilon < 1$ , we have  $0 < 1 - \epsilon < 1$ . The limit is 0. Yes,  $Y_n = \min\{X_1, \dots, X_n\} \xrightarrow{P} 0$ .

# Related Topics and Other Convergence Types

Beyond Markov and Chebyshev:

- Tighter bounds on tail probabilities exist, e.g., Chernoff bound (uses the moment generating function).
- Central Limit Theorem (CLT): Provides an approximation for the *distribution* of sums/averages (often Gaussian), not just bounds on probabilities.

Other Forms of Convergence for Random Variables:

- **Convergence in Probability** (WLLN):  $P(|Y_n - a| \geq \epsilon) \rightarrow 0$ . Probability of being far from limit goes to zero.
- **Almost Sure Convergence** (Convergence with Probability 1):  $P(\lim_{n \rightarrow \infty} Y_n = a) = 1$ . The sequence of outcomes  $Y_n(\omega)$  converges to  $a$  for almost every outcome  $\omega$ . Stronger than convergence in probability.
  - Strong Law of Large Numbers (SLLN): Under similar conditions as WLLN,  $M_n \rightarrow \mu$  almost surely.
- **Convergence in Distribution**: The CDF of  $Y_n$  converges to the CDF of a limiting random variable  $Y$ ,  $F_{Y_n}(y) \rightarrow F_Y(y)$  at continuity points. (Related to CLT).

## **Lecture 21: The Central Limit Theorem (CLT)**

---

# Different Scalings of the Sum of i.i.d. Random Variables

- $X_1, \dots, X_n$  i.i.d., finite mean  $\mu$  and variance  $\sigma^2$
- $S_n = X_1 + \dots + X_n$ 
  - variance:  $n\sigma^2$  (variance grows)
- $M_n = \frac{S_n}{n} = \frac{X_1 + \dots + X_n}{n}$ 
  - variance:  $\frac{\sigma^2}{n}$  (variance shrinks to 0  $\implies$  WLLN)
- $\frac{S_n}{\sqrt{n}} = \frac{X_1 + \dots + X_n}{\sqrt{n}}$ 
  - variance:  $\sigma^2$  (variance is constant)

# The Central Limit Theorem (CLT)

- $X_1, \dots, X_n$  i.i.d., finite mean  $\mu$  and variance  $\sigma^2$
- $S_n = X_1 + \dots + X_n$  (variance:  $n\sigma^2$ )
- $\frac{S_n}{\sqrt{n}} = \frac{X_1 + \dots + X_n}{\sqrt{n}}$  (variance:  $\sigma^2$ )

Standardized sum:

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

- $E[Z_n] = \frac{E[S_n] - n\mu}{\sqrt{n}\sigma} = \frac{n\mu - n\mu}{\sqrt{n}\sigma} = 0$
- $\text{var}(Z_n) = \text{var}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma}\right) = \frac{\text{var}(S_n)}{n\sigma^2} = \frac{n\sigma^2}{n\sigma^2} = 1$

Let  $Z$  be a standard normal r.v. (zero mean, unit variance),  $Z \sim N(0, 1)$ .

## Central Limit Theorem

For every  $z$ :  $\lim_{n \rightarrow \infty} P(Z_n \leq z) = P(Z \leq z) = \Phi(z)$

- $P(Z \leq z)$  is the standard normal CDF,  $\Phi(z)$ , available from normal tables.

# Usefulness of the CLT

$$S_n = X_1 + \cdots + X_n$$

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} \quad Z \sim N(0, 1)$$

Central Limit Theorem: For every  $z$ :  $\lim_{n \rightarrow \infty} P(Z_n \leq z) = P(Z \leq z)$

- Universal: Applies regardless of the underlying distribution of  $X_i$  (as long as  $\mu, \sigma^2$  are finite).
- Easy to apply: Only requires means and variances.
- Fairly accurate computational shortcut for sums of many RVs.
- Provides a justification for using normal models in many real-world scenarios (where noise is the sum of many small effects).

## What exactly does the CLT say? - Theory

$$S_n = X_1 + \cdots + X_n \quad Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} \quad Z \sim N(0, 1)$$

Central Limit Theorem: For every  $z$ :  $\lim_{n \rightarrow \infty} P(Z_n \leq z) = P(Z \leq z)$

- The CDF of  $Z_n$  converges to the standard normal CDF.
- There are further results for convergence of PDFs or PMFs (requires more assumptions).
- There are versions of the CLT that do not require the  $X_i$  to be identically distributed.
- There are versions that hold even under “weak dependence” (not fully independent).
- Proof: Typically uses “transforms” (Moment Generating Functions):

$$E[e^{sZ_n}] \rightarrow E[e^{sZ}] = e^{s^2/2}, \text{ for all } s$$

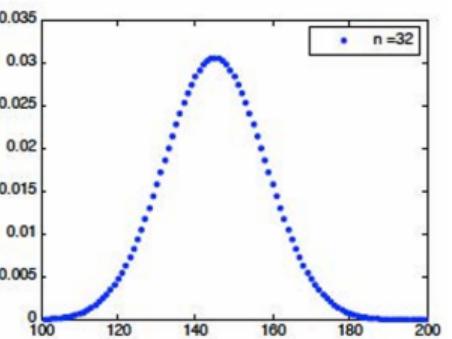
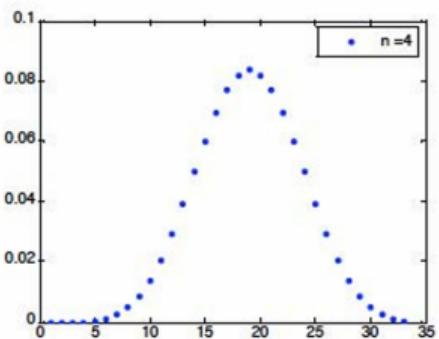
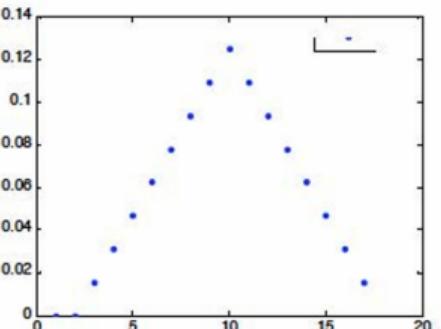
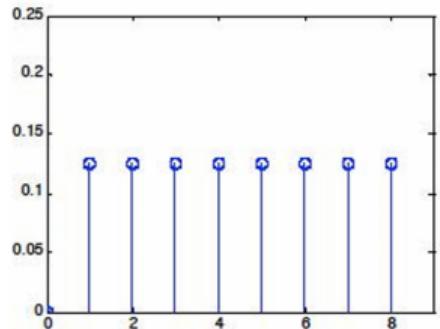
# What exactly does the CLT say? - Practice

$$S_n = X_1 + \cdots + X_n \quad Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} \quad Z \sim N(0, 1)$$

Central Limit Theorem: For every  $z$ :  $\lim_{n \rightarrow \infty} P(Z_n \leq z) = P(Z \leq z)$

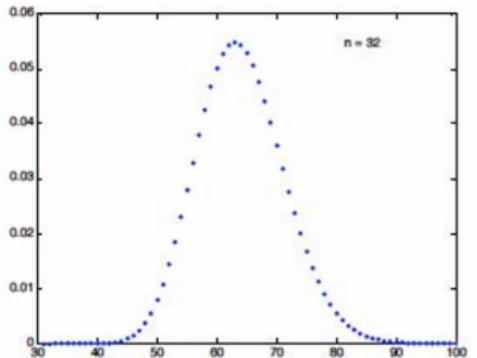
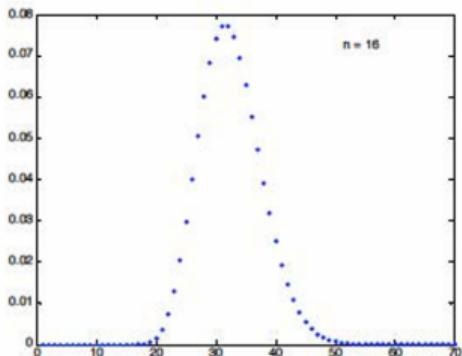
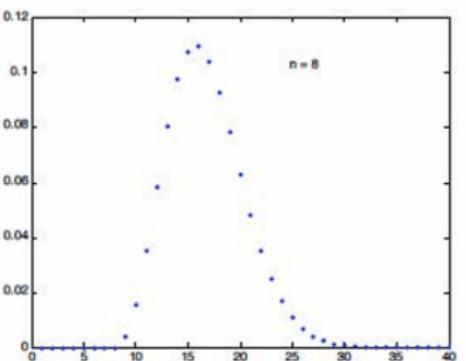
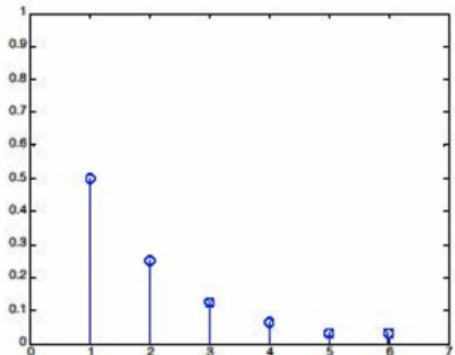
- The practice of normal approximations:
  - Treat  $Z_n$  as if it were a standard normal random variable.
  - Hence, treat  $S_n$  as if it were a normal random variable:  $S_n \approx N(n\mu, n\sigma^2)$ .
- Can we use the CLT when  $n$  is “moderate”? (e.g.,  $n=30$  or  $n=50$ )
  - Usually, yes. The approximation is often good enough.
  - If the underlying distribution  $f_X(x)$  is symmetric and unimodal (bell-shaped), the convergence is very fast.

# CLT Illustration: Sum of Symmetric (Uniform) RVs





# CLT Illustration: Sum of Non-Symmetric (Geometric) RVs



## Example 1: Find Probability

Standard problem:  $P(S_n \leq a) \approx b$ . Given two parameters, find the third.

- Package weights  $X_i$  are i.i.d. exponential,  $\lambda = 1/2$ . From formula sheet:  $\mu = E[X_i] = 1/\lambda = 2$ .
- $\sigma^2 = \text{var}(X_i) = 1/\lambda^2 = 4$ . So,  $\sigma = 2$ .
- Load container with  $n = 100$  packages. Find  $P(S_n \geq 210)$ .

Standardize the sum  $S_n$ :

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{S_n - 100(2)}{\sqrt{100}(2)} = \frac{S_n - 200}{20}$$

$$P(S_n \geq 210) = P\left(\frac{S_n - 200}{20} \geq \frac{210 - 200}{20}\right) = P(Z_n \geq 0.5)$$

Using CLT, approximate  $P(Z_n \geq 0.5)$  with  $P(Z \geq 0.5)$ , where  $Z \sim N(0, 1)$ .

$$P(Z \geq 0.5) = 1 - P(Z < 0.5) = 1 - \Phi(0.5)$$

From table:  $\Phi(0.5) = 0.6915$ . Thus,  $P(S_n \geq 210) \approx 1 - 0.6915 = 0.3085$



# Normal Table

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
<b>0.0</b>	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
<b>0.1</b>	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
<b>0.2</b>	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
<b>0.3</b>	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
<b>0.4</b>	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
<b>0.5</b>	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
<b>0.6</b>	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
<b>0.7</b>	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
<b>0.8</b>	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
<b>0.9</b>	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
<b>1.0</b>	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
<b>1.1</b>	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
<b>1.2</b>	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
<b>1.3</b>	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
<b>1.4</b>	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
<b>1.5</b>	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
<b>1.6</b>	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
<b>1.7</b>	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
<b>1.8</b>	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
<b>1.9</b>	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
<b>2.0</b>	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
<b>2.1</b>	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
<b>2.2</b>	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
<b>2.3</b>	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
<b>2.4</b>	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
<b>2.5</b>	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520

## Example 2: Find Threshold

- Same setup:  $X_i$  i.i.d. exponential,  $\lambda = 1/2$ ,  $\mu = 2$ ,  $\sigma = 2$ .
- $n = 100$ .
- Choose capacity  $a$  so that  $P(S_n \geq a) \approx 0.05$ .

We want to find  $a$  such that:

$$P(S_n \geq a) = P\left(\frac{S_n - 200}{20} \geq \frac{a - 200}{20}\right) \approx 0.05, \quad \equiv \quad P\left(Z \geq \frac{a - 200}{20}\right) \approx 0.05$$

This is equivalent to:

$$1 - \Phi\left(\frac{a - 200}{20}\right) \approx 0.05 \implies \Phi\left(\frac{a - 200}{20}\right) \approx 0.95$$

From the normal table, find  $z$  such that  $\Phi(z) \approx 0.95$ .  $\Phi(1.64) = 0.9495$ ,  $\Phi(1.65) = 0.9505$ .

Let's use  $z \approx 1.645$ .

$$\frac{a - 200}{20} \approx 1.645 \quad \Rightarrow \quad a \approx 200 + 20(1.645) = 200 + 32.9 = 232.9$$

## Example 3: Find Sample Size

- Same setup:  $X_i$  i.i.d. exponential,  $\lambda = 1/2$ ,  $\mu = 2$ ,  $\sigma = 2$ .
- How large can  $n$  be so that  $P(S_n \geq 210) \approx 0.05$ ?

We want to find  $n$  such that:

$$P(S_n \geq 210) = P\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \geq \frac{210 - n\mu}{\sqrt{n}\sigma}\right) \approx 0.05 \quad \equiv \quad P\left(Z \geq \frac{210 - 2n}{2\sqrt{n}}\right) \approx 0.05$$

From Example 2, we know this means the argument must be  $\approx 1.645$ .

$$\frac{210 - 2n}{2\sqrt{n}} \approx 1.645 \quad \equiv \quad 210 - 2n \approx 3.29\sqrt{n} \quad \equiv \quad 2n + 3.29\sqrt{n} - 210 \approx 0$$

This is a quadratic equation in  $\sqrt{n}$ . Let  $y = \sqrt{n}$ .  $2y^2 + 3.29y - 210 = 0$ .

## Example 3: Find Sample Size (Cont'd)

This is a quadratic equation in  $\sqrt{n}$ . Let  $y = \sqrt{n}$ .  $2y^2 + 3.29y - 210 = 0$ .

$$y = \frac{-3.29 \pm \sqrt{3.29^2 - 4(2)(-210)}}{2(2)} = \frac{-3.29 \pm \sqrt{10.82 + 1680}}{4}$$

Since  $y = \sqrt{n} > 0$ :

$$y \approx \frac{-3.29 + \sqrt{1690.82}}{4} \approx \frac{-3.29 + 41.12}{4} = \frac{37.83}{4} \approx 9.46$$

$$n = y^2 \approx 9.46^2 \approx 89.5$$

So,  $n \approx 89$  or  $90$ .

## Example 4: Alternative Formulation

- Same setup:  $X_i$  i.i.d. exponential,  $\lambda = 1/2$ ,  $\mu = 2$ ,  $\sigma = 2$ .
- Load container until weight exceeds 210.
- $N$ : number of packages loaded.
- Find  $P(N > 100)$ .

The event  $N > 100$  (it takes more than 100 packages to exceed 210) is the same as the event  $S_{100} \leq 210$  (the sum of the first 100 packages is less than or equal to 210).

$$P(N > 100) = P(S_{100} \leq 210)$$

This is the complement of the probability from Example 1.

$$P(S_{100} \leq 210) = P\left(Z_{100} \leq \frac{210 - 200}{20}\right) = P(Z_{100} \leq 0.5)$$

Using CLT, we found  $P(N > 100) \approx \Phi(0.5)$ . From table:  $\Phi(0.5) = 0.6915$ . Thus,  
 $P(N > 100) \approx 0.6915$ .

# Normal Approximation to the Binomial

- $X_i$ : independent, Bernoulli( $p$ );  $0 < p < 1$
- $S_n = X_1 + \dots + X_n$ : Binomial( $n$ ,  $p$ )
- $E[S_n] = n\mu = np$ ,  $\text{var}(S_n) = n\sigma^2 = np(1 - p)$

Applying the CLT to  $S_n$ : The CDF of  $Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}$  converges to the standard normal CDF.

Example:  $n = 36$ ,  $p = 0.5$ . Find  $P(S_n \leq 21)$ .

- Mean:  $np = 36(0.5) = 18$ , Variance:  $np(1 - p) = 36(0.5)(0.5) = 9$ , Standard Deviation:  $\sqrt{np(1 - p)} = 3$

Exact answer:  $\sum_{k=0}^{21} \binom{36}{k} \left(\frac{1}{2}\right)^{36} = 0.8785$

CLT approximation:

$$P(S_n \leq 21) = P\left(\frac{S_n - 18}{3} \leq \frac{21 - 18}{3}\right) = P(Z_n \leq 1) \approx \Phi(1) = 0.8413$$

This approximation is not very good. We can improve it.

# The 1/2 Correction for Integer Random Variables

- $S_n$  is an integer random variable.
- The event  $P(S_n \leq 21)$  is the same as  $P(S_n < 22)$ .
- The normal approximation (a continuous distribution) assigns probability to the interval  $(21, 21.5)$ .
- A better approximation is to use the midpoint 21.5.
- $P(S_n \leq 21) = P(S_n \leq 21.5)$  (since  $S_n$  is integer).

Revised CLT approximation (with 1/2 correction):

$$P(S_n \leq 21.5) = P\left(\frac{S_n - 18}{3} \leq \frac{21.5 - 18}{3}\right) = P(Z_n \leq \frac{3.5}{3}) = P(Z_n \leq 1.166\dots)$$
$$\approx \Phi(1.17) = 0.8790$$

This is much closer to the true value of 0.8785.

# De Moivre-Laplace CLT to the Binomial PMF

- We can also approximate the PMF value  $P(S_n = k)$ .
- We approximate  $P(S_n = k)$  as the area under the normal curve over the interval  $[k - 0.5, k + 0.5]$ .

$$P(S_n = k) = P(k - 0.5 \leq S_n \leq k + 0.5)$$

$$\approx \Phi\left(\frac{k + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - 0.5 - np}{\sqrt{np(1-p)}}\right)$$

Example:  $n = 36$ ,  $p = 0.5$ . Find  $P(S_n = 19)$ .

- Exact answer:  $\binom{36}{19} \left(\frac{1}{2}\right)^{36} = 0.1251$
- CLT approximation:

$$P(18.5 \leq S_n \leq 19.5) = P\left(\frac{18.5 - 18}{3} \leq Z_n \leq \frac{19.5 - 18}{3}\right)$$

$$= P(0.166 \leq Z_n \leq 0.5) \approx \Phi(0.5) - \Phi(0.17) \approx 0.6915 - 0.5675 = 0.1240$$

# The Pollster's Problem Revisited

- $p$ : fraction of population that will vote “yes”.
- $X_i \sim \text{Bernoulli}(p)$ ,  $E[X_i] = p$ ,  $\sigma^2 = p(1 - p)$ .
- $M_n = S_n/n$ : fraction of “yes” in our sample.
- Goal:  $P(|M_n - p| \geq 0.01)$

This is  $P(M_n - p \geq 0.01) + P(M_n - p \leq -0.01)$ .

$$P(M_n - p \geq 0.01) = P(S_n/n - p \geq 0.01) = P(S_n - np \geq 0.01n)$$

$$= P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \geq \frac{0.01n}{\sqrt{np(1-p)}}\right) = P\left(Z_n \geq \frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right)$$

Worst case variance  $\sigma^2 = p(1 - p)$  is when  $p = 0.5$ ,  $\sigma^2 = 0.25$ ,  $\sigma = 0.5$ .

$$P\left(Z_n \geq \frac{0.01\sqrt{n}}{0.5}\right) = P(Z_n \geq 0.02\sqrt{n})$$

## The Pollster's Problem Revisited (Cont'd)

So,  $P(|M_n - p| \geq 0.01) \approx P(|Z| \geq 0.02\sqrt{n}) = 2 \cdot P(Z \geq 0.02\sqrt{n})$ .

- Try  $n = 10,000$ :

$$0.02\sqrt{n} = 0.02\sqrt{10000} = 0.02 \times 100 = 2$$

$$P(|M_{10000} - p| \geq 0.01) \approx 2 \cdot P(Z \geq 2) = 2(1 - \Phi(2)) \approx 2(1 - 0.9772) = 2(0.0228) = 0.0456$$

This is  $\approx 4.6\%$ . (Chebyshev gave  $\leq 25\%$ ).

- Specs: Find  $n$  so that  $P(|M_n - p| \geq 0.01) \leq 0.05$ .

$$2P(Z \geq 0.02\sqrt{n}) \leq 0.05 \quad \equiv \quad P(Z \geq 0.02\sqrt{n}) \leq 0.025 \quad \equiv \quad 1 - \Phi(0.02\sqrt{n}) \leq 0.025$$

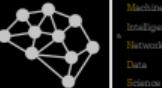
So,  $\Phi(0.02\sqrt{n}) \geq 0.975$ . From the table,  $\Phi(1.96) = 0.975$ .

$$0.02\sqrt{n} \geq 1.96 \quad \equiv \quad \sqrt{n} \geq \frac{1.96}{0.02} = 98 \quad \equiv \quad n \geq 98^2 = 9604$$

CLT suggests  $n \approx 9604$ , whereas Chebyshev required  $n \geq 25000$ .

## Lecture 22: Classical Statistics I

---



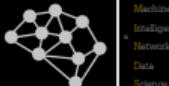
# Classical Statistics

- **Bayesian Inference (Recap):**

- Unknown parameter  $\theta$  is a random variable.
- Observation  $X$  is a random variable.
- Goal: Find posterior  $p_{\theta|X}(\theta|x)$  or  $f_{\theta|X}(\theta|x)$ .

- **Classical Statistics:**

- Unknown parameter  $\theta$  is a fixed, deterministic constant.
- Observation  $X$  is a random variable.
- Model:  $p_X(x; \theta)$  or  $f_X(x; \theta)$ .
- These are **not** conditional probabilities;  $\theta$  is not random.
- We have a family of models, one for each possible  $\theta$ .
- Estimator  $\hat{\Theta} = g(X)$  is a function of  $X$  used to guess  $\theta$ .
- Extends to vectors:  $p_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta_1, \dots, \theta_m)$ .



# Problem Types in Classical Statistics

Setup: Unknown constant  $\theta \implies$  Model  $p_X(x; \theta) \implies$  Estimator  $\hat{\Theta}$

- **Hypothesis Testing (Binary):**

- $H_0 : \theta = 1/2$  versus  $H_1 : \theta = 3/4$ .

- **Composite Hypotheses:**

- $H_0 : \theta = 1/2$  versus  $H_1 : \theta \neq 1/2$ .

- **Estimation:**

- Design an estimator  $\hat{\Theta} = g(X)$ .
- Goal: Keep estimation error  $\hat{\Theta} - \theta$  “small”.

# Estimating a Mean

- $X_1, \dots, X_n$  are i.i.d.
- Unknown mean:  $\theta = \mathbb{E}[X_i]$ .
- Unknown variance:  $\sigma^2 = \text{Var}(X_i)$ .

**Estimator:** Sample Mean

$$\hat{\Theta}_n = M_n = \frac{X_1 + \dots + X_n}{n}$$

## Properties and Terminology

- $\hat{\Theta}_n$  is the **estimator** (a random variable).
- $E[\hat{\Theta}_n] = \theta$ . This estimator is **unbiased**.
- WLLN:  $\hat{\Theta}_n \xrightarrow{P} \theta$ . This estimator is **consistent**.
- **Mean Squared Error (MSE):**  $E[(\hat{\Theta}_n - \theta)^2]$ .

# On the Mean Squared Error of an Estimator

For any estimator  $\hat{\Theta}$  of a constant  $\theta$ :

- Use the property  $E[Z^2] = \text{Var}(Z) + (E[Z])^2$ .
- Let  $Z = \hat{\Theta} - \theta$ .
- $E[(\hat{\Theta} - \theta)^2] = \text{Var}(\hat{\Theta} - \theta) + (E[\hat{\Theta} - \theta])^2$
- Since  $\theta$  is constant,  $\text{Var}(\hat{\Theta} - \theta) = \text{Var}(\hat{\Theta})$ .
- The **bias** is  $b(\hat{\Theta}) = E[\hat{\Theta}] - \theta$ .
- **MSE** =  $\text{Var}(\hat{\Theta}) + (\text{bias})^2$
- $\sqrt{\text{Var}(\hat{\Theta})}$  is called the **standard error**.

# Confidence Intervals (CIs)

- A point estimate  $\hat{\theta}$  alone may not be informative enough.
- A  $1 - \alpha$  **confidence interval** is a random interval  $[\hat{\Theta}^-, \hat{\Theta}^+]$  computed from the data.
- It must satisfy:  $P(\hat{\Theta}^- \leq \theta \leq \hat{\Theta}^+) \geq 1 - \alpha$ , for all possible values of  $\theta$ .
- $\alpha$  is the error probability (e.g.,  $\alpha = 0.05$  for a 95% CI).
- **Interpretation (Subtle):**  $\theta$  is fixed. The interval is random. If we repeat the experiment many times, the computed interval will contain the true  $\theta$  at least  $100(1 - \alpha)\%$  of the time.

# CI for the Estimation of the Mean ( $\sigma$ Known)

- Estimator:  $\hat{\Theta}_n = M_n = \frac{X_1 + \dots + X_n}{n}$
- Standardized estimator (by CLT):  $Z_n = \frac{\hat{\Theta}_n - \theta}{\sigma/\sqrt{n}} \approx N(0, 1)$

For a 95% CI ( $\alpha = 0.05$ ):

- Find  $z$  such that  $P(-z \leq Z \leq z) = 0.95$ .
- This implies  $P(Z \leq z) = 0.975$ . From normal tables,  $z = 1.96$ .
- $P\left(-1.96 \leq \frac{\hat{\Theta}_n - \theta}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95$

Invert the inequality to isolate  $\theta$ :

$$P\left(\hat{\Theta}_n - 1.96 \frac{\sigma}{\sqrt{n}} \leq \theta \leq \hat{\Theta}_n + 1.96 \frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

The 95% CI is:  $\left[\hat{\Theta}_n - 1.96 \frac{\sigma}{\sqrt{n}}, \hat{\Theta}_n + 1.96 \frac{\sigma}{\sqrt{n}}\right]$

# CIs for the Mean when $\sigma$ is Unknown

The formula  $\left[ \hat{\Theta}_n \pm 1.96 \frac{\sigma}{\sqrt{n}} \right]$  requires  $\sigma$ . What if  $\sigma$  is unknown?

- **Option 1: Use an upper bound on  $\sigma$ .**

- E.g., if  $X_i$  are Bernoulli( $\theta$ ),  $\sigma^2 = \theta(1 - \theta) \leq 1/4$ , so  $\sigma \leq 1/2$ .
- This gives a conservative (wider) interval.

- **Option 2: Use an ad hoc estimate  $\hat{\sigma}$ .**

- E.g., if  $X_i$  are Bernoulli( $\theta$ ), estimate  $\hat{\sigma} = \sqrt{\hat{\Theta}_n(1 - \hat{\Theta}_n)}$ .
- This is a “plug-in” estimator.

# CLs for the Mean when $\sigma$ is Unknown (cont.)

- **Option 3: Use the sample variance estimate.**

- We know  $\sigma^2 = E[(X_i - \theta)^2]$ .
- By WLLN,  $\frac{1}{n} \sum_{i=1}^n (X_i - \theta)^2 \xrightarrow{P} \sigma^2$ .
- Since we don't know  $\theta$ , we plug in our estimate  $\hat{\Theta}_n$ :

$$\hat{v}_n = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\Theta}_n)^2$$

- This estimate  $\hat{v}_n$  also converges to  $\sigma^2$  and is consistent.
- This CI relies on two approximations:
  1. The CLT approximation (assuming  $Z_n$  is normal).
  2. The variance approximation (using  $\hat{v}_n$  for  $\sigma^2$ ).
- For small  $n$ , if  $X_i$  are *exactly* normal, the statistic  $\frac{\hat{\Theta}_n - \theta}{\sqrt{\hat{S}_n^2/n}}$  (using unbiased variance  $\hat{S}_n^2 = \frac{1}{n-1} \sum (X_i - \hat{\Theta}_n)^2$ ) follows a **t-distribution**, which is wider and accounts for the uncertainty in the variance estimate.

# Other Natural Estimators (Method of Moments)

We can estimate any moment by its corresponding sample average.

- **Mean:**  $\theta_X = E[X]$ 
  - **Estimator:**  $\hat{\Theta}_X = \frac{1}{n} \sum_{i=1}^n X_i$
- **Variance:**  $\nu_X = E[(X - \theta_X)^2]$ 
  - **Estimator:**  $\hat{\nu}_X = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\Theta}_X)^2$
- **Covariance:**  $\text{Cov}(X, Y) = E[(X - \theta_X)(Y - \theta_Y)]$ 
  - **Estimator:**  $\hat{\text{Cov}}(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\Theta}_X)(Y_i - \hat{\Theta}_Y)$
- **Function Mean:**  $\theta = E[g(X)]$ 
  - **Estimator:**  $\hat{\Theta} = \frac{1}{n} \sum_{i=1}^n g(X_i)$

The next step is to analyze the properties (MSE, CIs) of these estimators.

# Maximum Likelihood (ML) Estimation

A core principle of classical estimation.

- **Idea:** Pick the parameter  $\theta$  that “makes the observed data  $x$  most likely.”
- The **likelihood function** is  $p_X(x; \theta)$  or  $f_X(x; \theta)$ , viewed as a function of  $\theta$  for fixed  $x$ .
- The **ML estimate**  $\hat{\theta}_{ML}$  is the value of  $\theta$  that maximizes this function:

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} p_X(x; \theta) \quad (\text{or } f_X(x; \theta))$$

## Comparison with Bayesian MAP

- MAP maximizes:  $p_{X|\theta}(x|\theta)p_\theta(\theta)$  (likelihood  $\times$  prior)
- ML maximizes:  $p_X(x; \theta)$  (likelihood only)
- ML is equivalent to MAP estimation when the prior  $p_\theta(\theta)$  is uniform (flat).
- The interpretation is different:  $\theta$  is a fixed constant.

## Comments on ML Estimation

- In practice, we maximize the **log-likelihood**  $\log p_X(x; \theta)$  (since log is monotonic and turns products into sums).
- For  $n$  i.i.d. observations,  $\hat{\Theta}_n$  (the ML estimator) has strong properties:
  - **Consistent:**  $\hat{\Theta}_n \xrightarrow{P} \theta$ .
  - **Asymptotically Normal:** The error  $\hat{\Theta}_n - \theta$  is approximately normal.

$$\frac{\hat{\Theta}_n - \theta}{\sigma(\hat{\Theta}_n)} \xrightarrow{\text{dist}} N(0, 1)$$

- This allows for constructing CIs:  $[\hat{\Theta}_n \pm 1.96\hat{\sigma}(\hat{\Theta}_n)]$ .
- **Asymptotically Efficient:** For large  $n$ ,  $\hat{\Theta}_n$  has the smallest possible variance among “good” estimators. It is “best”.

# ML Example: Parameter of Binomial

- $K$ : number of successes (Binomial) in  $n$  (known) trials.
- $\theta$  (unknown) is the probability of success  $p$ .
- Likelihood function:  $p_K(k; \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$ .

Log-likelihood:  $L(\theta) = \log \binom{n}{k} + k \log \theta + (n - k) \log(1 - \theta)$ .

Differentiate w.r.t.  $\theta$  and set to 0:

$$\frac{dL}{d\theta} = \frac{k}{\theta} - \frac{n - k}{1 - \theta} = 0 \implies \frac{k}{\theta} = \frac{n - k}{1 - \theta}$$

$$k(1 - \theta) = (n - k)\theta \implies k - k\theta = n\theta - k\theta$$

$$k = n\theta \implies \hat{\theta}_{ML} = \frac{k}{n}$$

The ML estimator is the sample mean:  $\hat{\Theta}_{ML} = K/n$ . (This is the same result as the MAP estimator with a uniform prior  $U[0, 1]$ ).

# ML Example: Normal Mean and Variance

- $X_1, \dots, X_n$  are i.i.d.  $N(\mu, \nu)$ , where  $\nu = \sigma^2$ .
- Parameter vector  $\theta = (\mu, \nu)$ .
- Likelihood:  $f_X(x; \mu, \nu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\nu}} \exp \left\{ -\frac{(x_i - \mu)^2}{2\nu} \right\}$ .

Log-likelihood:  $L(\mu, \nu) = \log f_X(x; \mu, \nu) = \sum_{i=1}^n \left( -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log \nu - \frac{(x_i - \mu)^2}{2\nu} \right)$

$$L(\mu, \nu) = C - \frac{n}{2} \log \nu - \frac{1}{2\nu} \sum_{i=1}^n (x_i - \mu)^2$$

## ML Example: Normal Mean and Variance (Cont'd)

1. **Minimize w.r.t.  $\mu$ :** This means minimizing  $\sum(x_i - \mu)^2$ .

$$\frac{\partial L}{\partial \mu} = -\frac{1}{2v} \sum_{i=1}^n 2(x_i - \mu)(-1) = 0 \implies \sum_{i=1}^n (x_i - \mu) = 0$$

$$\sum x_i - n\mu = 0 \implies \hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^n x_i$$

2. **Minimize w.r.t.  $v$ :** Plug in  $\hat{\mu}_{ML}$  and set  $\frac{\partial L}{\partial v} = 0$ :

$$\frac{\partial L}{\partial v} = -\frac{n}{2v} + \frac{1}{2v^2} \sum_{i=1}^n (x_i - \hat{\mu}_{ML})^2 = 0$$

$$\frac{1}{2v^2} \sum (x_i - \hat{\mu}_{ML})^2 = \frac{n}{2v} \implies \hat{v}_{ML} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_{ML})^2$$

The ML estimators are the sample mean and the (biased) sample variance.

## **Lecture 23: Classical Statistics**

**II**

---

# Linear Regression

- We have  $n$  data pairs  $(x_1, y_1), \dots, (x_n, y_n)$ .
- We want to model the relationship between  $X$  and  $Y$ .
- Assume an approximately linear relationship:

$$Y \approx \theta_0 + \theta_1 X$$

- $\theta_0, \theta_1$  are unknown parameters to be estimated.

## The Least Squares Approach

- Find the line  $y = \hat{\theta}_0 + \hat{\theta}_1 x$  that “best fits” the data.
- **Residual:** The error for data point  $i$  is  $y_i - (\hat{\theta}_0 + \hat{\theta}_1 x_i)$ .
- **Goal:** Minimize the sum of the squared residuals over all  $\theta_0, \theta_1$ :

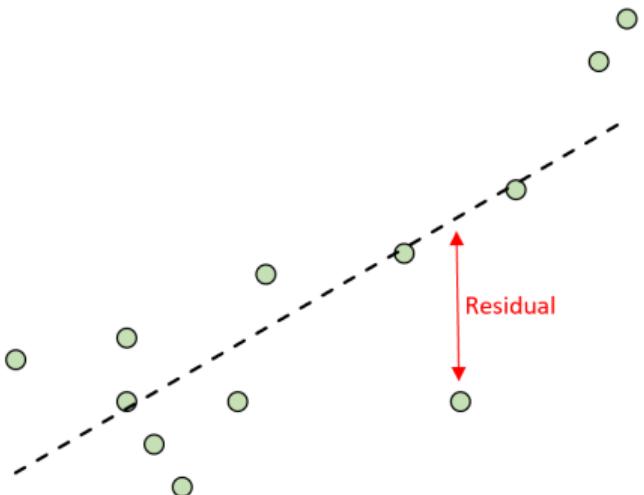
$$\min_{\theta_0, \theta_1} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2$$

# Visualizing Linear Regression

Cost Function to Minimize:

$$J(\theta_0, \theta_1) = \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2$$

This is a quadratic function of  $\theta_0$  and  $\theta_1$ .



# Linear Regression Solution

Minimize  $J(\theta_0, \theta_1)$  by setting partial derivatives to zero:

$$\frac{\partial J}{\partial \theta_0} = 0 \quad \text{and} \quad \frac{\partial J}{\partial \theta_1} = 0$$

This gives a system of two linear equations in  $\hat{\theta}_0$  and  $\hat{\theta}_1$ .

## Regression Estimates

Let  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  and  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  be the sample means.

The optimal estimates are:

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

Note: The line of best fit always passes through the “center of mass”  $(\bar{x}, \bar{y})$  of the data.

## Example: Leaning Tower of Pisa

Data: Lean (in meters) vs. Year.

- (1975, 2.9642)
- (1976, 2.9644)
- ...
- (1987, 2.9757)

Model:  $y \approx \theta_0 + \theta_1 x$ , where  $x$  is the year,  $y$  is the lean.

Calculations from data (13 points):

- $\bar{x} = 1981$
- $\bar{y} = 2.9694$

Estimates (from formulas):

- $\hat{\theta}_1 \approx 0.0009$  (meters per year)
- $\hat{\theta}_0 \approx 1.1233$

Estimated Model:  $Y = 1.1233 + 0.0009x$

# Probabilistic Justifications for Least Squares

Why minimize the *square* of the residuals?

- **1. Maximum Likelihood (ML) Estimation**

- Assume a linear model with normal noise:

$$Y_i = \theta_0 + \theta_1 x_i + W_i$$

- Assume  $W_i$  are i.i.d.  $N(0, \sigma^2)$ .
- Then  $Y_i \sim N(\theta_0 + \theta_1 x_i, \sigma^2)$ .
- The likelihood function is  $f_Y(y; \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2} \right\}$ .
- Maximizing this likelihood (or log-likelihood) w.r.t.  $\theta_0, \theta_1$  is **equivalent** to minimizing the sum of squared residuals:

$$\sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2$$

- **2. Approximate Bayesian LLMS**

- We can also show that the regression formulas are sample-based approximations of the Bayesian LLMS estimator coefficients.

# Bayesian Linear Regression

We can also treat  $\theta_0, \theta_1$  as random variables  $\theta_0, \theta_1$ .

- **Model:**  $Y_i = \theta_0 + \theta_1 x_i + W_i$
- **Priors:**  $\theta_0 \sim N(0, \sigma_0^2)$ ,  $\theta_1 \sim N(0, \sigma_1^2)$ ,  $W_i \sim N(0, \sigma^2)$
- Assume all are independent.

## MAP Estimation

Find  $(\hat{\theta}_0, \hat{\theta}_1)$  that maximize  $f_{\Theta|Y}(\theta|y) \propto f_{Y|\Theta}(y|\theta)f_{\Theta}(\theta)$ . This is equivalent to minimizing:

$$\sum_{i=1}^n \frac{(y_i - \theta_0 - \theta_1 x_i)^2}{\sigma^2} + \frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2}$$

- This is a “regularized” least squares problem.
- The prior terms  $\frac{\theta_0^2}{\sigma_0^2}$  and  $\frac{\theta_1^2}{\sigma_1^2}$  pull the estimates toward their mean (0) to prevent overfitting.
- If prior variances  $\sigma_0, \sigma_1 \rightarrow \infty$  (a “flat” prior), the solution becomes identical to the classical least squares / ML estimate.

# Multiple and Nonlinear Regression

The linear regression framework is very general.

- **Multiple Linear Regression**

- Model with multiple explanatory variables.
- E.g.,  $Y \approx \theta_0 + \theta_1 X_1 + \theta_2 X_2$ .
- Data:  $(x_{i,1}, x_{i,2}, y_i)$ .
- Goal:  $\min_{\theta_0, \theta_1, \theta_2} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_{i,1} - \theta_2 x_{i,2})^2$ .
- Solved by setting 3 partial derivatives to 0 (3 linear equations).

- **Polynomial Regression**

- Model:  $Y \approx \theta_0 + \theta_1 X + \theta_2 X^2$ .
- This is a special case of multiple regression.
- Let  $X_1 = X$  and  $X_2 = X^2$ .
- This is *still a linear regression model* because it is linear in the unknown parameters  $\theta_0, \theta_1, \theta_2$ .

# Binary Hypothesis Testing

- We must decide between two competing hypotheses,  $H_0$  and  $H_1$ .
- This is a choice between two probabilistic models for our data  $X$ .

## Setup

- **Null Hypothesis ( $H_0$ )**: The parameter is  $\theta_0$ .
  - $X \sim p_X(x; \theta_0)$  or  $f_X(x; \theta_0)$ .
- **Alternative Hypothesis ( $H_1$ )**: The parameter is  $\theta_1$ .
  - $X \sim p_X(x; \theta_1)$  or  $f_X(x; \theta_1)$ .
- $\theta$  is a fixed constant (not a random variable).

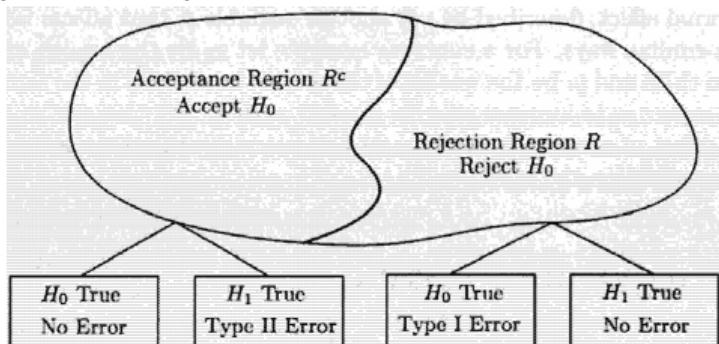
**Goal:** Design a decision rule based on  $X$  to choose  $H_0$  or  $H_1$ .

# Decision Rules and Types of Error

- A **decision rule** is a partition of the observation space into:
  - **Rejection Region ( $R$ )**: The set of observations  $x$  for which we decide to **reject  $H_0$**  (and accept  $H_1$ ).
  - **Acceptance Region ( $R^c$ )**: The set of observations  $x$  for which we **accept  $H_0$** .

Two types of error are possible:

- **Type I Error (False Rejection)**: We reject  $H_0$  when  $H_0$  is actually true.
  - Probability:  $\alpha(R) = P(X \in R; H_0)$ .
- **Type II Error (False Acceptance)**: We accept  $H_0$  when  $H_1$  is actually true.
  - Probability:  $\beta(R) = P(X \in R^c; H_1)$ .





# The Likelihood Ratio Test (LRT)

- How do we choose the “best” rejection region  $R$ ?
- **Bayesian MAP Motivation:**
  - If  $\theta$  were random, MAP rule says: Decide  $H_1$  if  $P(H_1|x) > P(H_0|x)$ .
  - $\implies P(x|H_1)P(H_1) > P(x|H_0)P(H_0)$ .
  - $\implies \frac{P(x|H_1)}{P(x|H_0)} > \frac{P(H_0)}{P(H_1)} = \text{threshold.}$

- **Classical LRT:** We adopt this same test structure.

- Define the **Likelihood Ratio**  $L(x)$ :

$$L(x) = \frac{f_X(x; \theta_1)}{f_X(x; \theta_0)} \quad (\text{or } \frac{p_X(x; \theta_1)}{p_X(x; \theta_0)})$$

- The LRT rule is: Reject  $H_0$  if  $L(x) > \xi$ .
- $R = \{x \mid L(x) > \xi\}$ , where  $\xi$  is the **critical value**.

# Choosing the Critical Value $\xi$

- There is an inherent **tradeoff** between  $\alpha$  and  $\beta$ .
- Increasing  $\xi \implies R$  gets smaller  $\implies \alpha$  decreases (good), but  $R^c$  gets larger  $\implies \beta$  increases (bad).

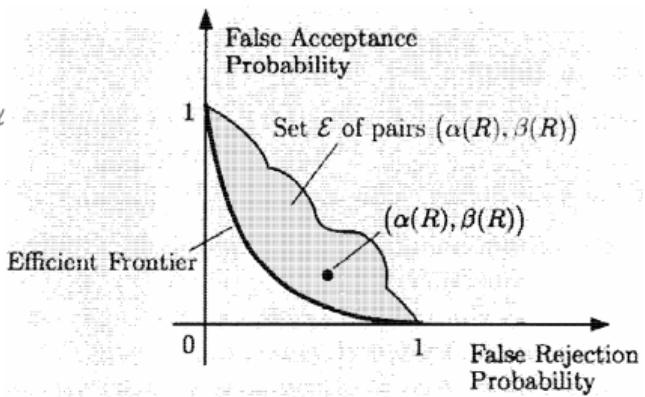
## The Neyman-Pearson Framework

The standard approach to setting  $\xi$  is:

1. Fix a maximum acceptable **significance level**  $\alpha$  for Type I error (e.g.,  $\alpha = 0.05$ ).
2. Find the critical value  $\xi$  that gives exactly this probability:

$$P(L(X) > \xi; H_0) = \alpha$$

3. This defines the rejection region  $R$ .



# Neyman-Pearson Lemma

## Neyman-Pearson Lemma

Among all possible decision rules (rejection regions  $R$ ) that have a false rejection probability  $\alpha(R) \leq \alpha$ , the Likelihood Ratio Test (LRT) with  $P(L(X) > \xi; H_0) = \alpha$  achieves the **smallest possible** false acceptance probability  $\beta$ .

- The LRT is the **most powerful** test.
- It gives the best “bang for your buck” in the  $\alpha$ - $\beta$  tradeoff.
- It is the optimal test in the classical framework.

## Example: Testing Normal Means

- Observe  $X \sim N(\theta, \sigma^2)$ , where  $\sigma^2$  is known.
- Test  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta = \theta_1$ . (Assume  $\theta_1 > \theta_0$ ).

### 1. Find the LRT structure:

$$L(x) = \frac{f_X(x; \theta_1)}{f_X(x; \theta_0)} = \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\theta_1)^2}{2\sigma^2}\right\}}{\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\theta_0)^2}{2\sigma^2}\right\}} = \exp\left\{\frac{(x - \theta_0)^2 - (x - \theta_1)^2}{2\sigma^2}\right\}$$

The test  $L(x) > \xi$  is equivalent to  $\log L(x) > \log \xi$ :

$$\frac{1}{2\sigma^2} ((x^2 - 2x\theta_0 + \theta_0^2) - (x^2 - 2x\theta_1 + \theta_1^2)) > \log \xi$$

$$\frac{1}{2\sigma^2} (2x(\theta_1 - \theta_0) + \theta_0^2 - \theta_1^2) > \log \xi$$

Since  $\theta_1 > \theta_0$ , this simplifies to  $x > \gamma$  for some threshold  $\gamma$ .

**The LRT is a simple threshold test: Reject  $H_0$  if  $x > \gamma$ .**

## Example: Finding $\gamma$ and $\beta$

**2. Find the threshold  $\gamma$  for a given  $\alpha$ :** We set the false rejection probability to  $\alpha$ .

$$P(\text{Reject } H_0; H_0) = P(X > \gamma; H_0) = \alpha$$

Under  $H_0$ ,  $X \sim N(\theta_0, \sigma^2)$ . We standardize  $X$ :

$$P\left(\frac{X - \theta_0}{\sigma} > \frac{\gamma - \theta_0}{\sigma}; H_0\right) = P\left(Z > \frac{\gamma - \theta_0}{\sigma}\right) = \alpha$$

Let  $z_\alpha$  be the value from the  $N(0, 1)$  table such that  $P(Z > z_\alpha) = \alpha$  (or  $\Phi(z_\alpha) = 1 - \alpha$ ).

$$\frac{\gamma - \theta_0}{\sigma} = z_\alpha \implies \gamma = \theta_0 + z_\alpha \sigma$$

## Example: Finding $\gamma$ and $\beta$ (Cont'd)

**3. Find the resulting Type II Error**  $\beta: \beta = P(\text{Accept } H_0; H_1) = P(X \leq \gamma; H_1)$ . Under  $H_1$ ,  $X \sim N(\theta_1, \sigma^2)$ . Standardize  $X$ :

$$\beta = P\left(\frac{X - \theta_1}{\sigma} \leq \frac{\gamma - \theta_1}{\sigma}; H_1\right) = P\left(Z \leq \frac{\gamma - \theta_1}{\sigma}\right)$$

Substitute the value of  $\gamma$ :

$$\beta = \Phi\left(\frac{(\theta_0 + z_\alpha \sigma) - \theta_1}{\sigma}\right) = \Phi\left(z_\alpha - \frac{\theta_1 - \theta_0}{\sigma}\right)$$

## Example: Discrete LRT (Coin Flips)

- $n = 25$  independent coin tosses.
- $X = \text{number of heads}$ .
- $H_0 : \theta_0 = 1/2$  (fair coin).  $S_n \sim \text{Binomial}(25, 1/2)$ .
- $H_1 : \theta_1 = 2/3$  (biased coin).  $S_n \sim \text{Binomial}(25, 2/3)$ .
- Set significance  $\alpha = 0.1$ .

Likelihood Ratio:

$$\begin{aligned}
 L(k) &= \frac{p_X(k; H_1)}{p_X(k; H_0)} = \frac{\binom{25}{k} (2/3)^k (1/3)^{25-k}}{\binom{25}{k} (1/2)^k (1/2)^{25-k}} = \frac{(2/3)^k (1/3)^{25-k}}{(1/2)^{25}} \\
 &= C \cdot \frac{(2/3)^k}{(1/3)^k} = C \cdot 2^k
 \end{aligned}$$

$L(k)$  is a monotonically increasing function of  $k$ .

## Example: Discrete LRT (Coin Flips) (Cont'd)

The LRT “Reject  $H_0$  if  $L(k) > \xi$ ” is equivalent to “Reject  $H_0$  if  $k > \gamma$ ”.

We find the smallest  $\gamma$  such that  $P(X > \gamma; H_0) \leq 0.1$ . Under  $H_0$ ,  $X \sim \text{Binomial}(25, 0.5)$ .

$E[X] = 12.5$ .  $\sigma = \sqrt{25(0.5)(0.5)} = 2.5$ . Using CLT approx. (with 1/2 correction):

$$P(X > \gamma; H_0) = P(X \geq \gamma + 1) \approx P\left(Z \geq \frac{(\gamma + 1) - 0.5 - 12.5}{2.5}\right)$$

$$P\left(Z \geq \frac{\gamma - 12}{2.5}\right) \leq 0.1$$

From table,  $P(Z \geq 1.28) \approx 0.1$ .

$$\frac{\gamma - 12}{2.5} \geq 1.28 \implies \gamma \geq 12 + 2.5(1.28) = 12 + 3.2 = 15.2$$

Since  $k$  must be an integer, we set  $\gamma = 16$ . **Decision Rule:** Reject  $H_0$  if  $X > 15$ .

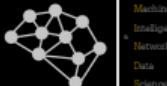
## **Lecture 24: The Bernoulli Process**

---

# The Bernoulli Process

- A sequence of **independent Bernoulli trials**,  $X_i$
- At each trial,  $i$ :
  - $P(X_i = 1) = P(\text{success at the } i\text{-th trial}) = p$
  - $P(X_i = 0) = P(\text{failure at the } i\text{-th trial}) = 1 - p$
- **Key assumptions:**
  - Independence of trials
  - Time-homogeneity (constant parameter  $p$ )
- Model of:
  - Sequence of lottery wins/losses
  - Arrivals (each second) to a bank
  - Arrivals (at each time slot) to a server





# Stochastic Processes

- **First view: sequence of random variables**  $X_1, X_2, \dots$
- Interested in:
  - Mean:  $E[X_i]$
  - Variance:  $\text{var}(X_i)$
  - Marginal distribution:  $p_{X_i}(x)$
  - Joint distribution:  $p_{X_1, \dots, X_n}(x_1, \dots, x_n)$
- **Second view: sample space**  $\Omega$
- $\Omega =$  (The space of all possible infinite sequences of outcomes)
- Example (for Bernoulli process):
  - $P(X_i = 1 \text{ for all } i)$

## Number of Successes/Arrivals $S$ in $n$ Time Slots

- $S = X_1 + X_2 + \cdots + X_n$
- $S$  follows a **Binomial distribution** with parameters  $n$  and  $p$ .
- $P(S = k) = \binom{n}{k} p^k (1 - p)^{n-k}$ , for  $k = 0, 1, \dots, n$ .
- $E[S] = np$
- $\text{var}(S) = np(1 - p)$

# Time Until the First Success/Arrival

- $T_1$  = time of the first success/arrival
- $T_1$  follows a **Geometric distribution** with parameter  $p$ .
- $P(T_1 = k) = (1 - p)^{k-1}p$ , for  $k = 1, 2, \dots$
- $E[T_1] = \frac{1}{p}$
- $\text{var}(T_1) = \frac{1-p}{p^2}$

# Independence, Memorylessness, and Fresh-Start Properties

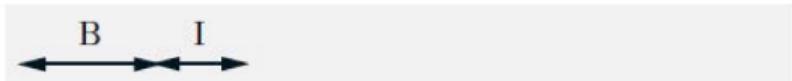
- **Memorylessness** is a key property of the Geometric distribution.
- **Fresh-start after time  $n$ :**
  - The sequence  $X_{n+1}, X_{n+2}, \dots$  is also a Bernoulli process with parameter  $p$ .
  - This future sequence is independent of  $X_1, \dots, X_n$ .
- **Fresh-start after time  $T_1$ :**
  - $T_1$  is the time of the first success.
  - The process immediately following the first success is statistically identical to the original process starting at time 1.

# Independence, Memorylessness, and Fresh-Start Properties

- **Fresh-start after a random time  $N$ :**
- Consider a random time  $N$  that is defined **causally** (i.e., its value depends only on  $X_1, \dots, X_N$ ).
- Examples of causal random times:
  - $N =$  time of the 3rd success
  - $N =$  first time that 3 successes in a row have been observed
  - $N =$  the time just before the first occurrence of 1, 1, 1
- The process  $X_{N+1}, X_{N+2}, \dots$  is:
  - a Bernoulli process with parameter  $p$ .
  - independent of  $N, X_1, \dots, X_N$ .

# The Distribution of Busy Periods

- At each slot, a server is **busy (B)** or **idle (I)**. This sequence  $X_i$  is a Bernoulli process.
- **First busy period:**
  - Starts with the first busy slot ( $X_i = 1$ ).
  - Ends just before the first subsequent idle slot ( $X_j = 0$ ).
- **Length of a busy period:** Number of consecutive successes. This follows a **Geometric distribution** shifted to start at 1.
- **Length of an idle period:** Number of consecutive failures. This also follows a **Geometric distribution** (with parameter  $1 - p$ ) shifted to start at 1.



# Time of the $k$ -th Success/Arrival

- $Y_k$  = time of the  $k$ -th arrival
- $Y_k$  is the sum of  $k$  inter-arrival times:  $Y_k = T_1 + T_2 + \cdots + T_k$
- $T_k = k$ -th inter-arrival time =  $Y_k - Y_{k-1}$  ( $k \geq 2$ ).
- **Distribution of  $T_i$ :**
  - The  $T_i$  are i.i.d., **Geometric( $p$ )** random variables.
  - This is a direct consequence of the fresh-start property after the occurrence of a success.
- **Distribution of  $Y_k$ :**
  - $Y_k$  follows a **Negative Binomial or Pascal distribution**.

## Time of the $k$ -th Success/Arrival (Negative Binomial)

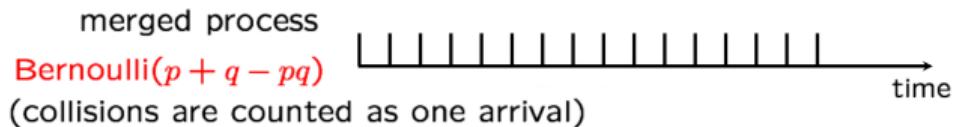
- $Y_k = T_1 + T_2 + \cdots + T_k$ , where  $T_i$  are i.i.d., Geometric( $p$ )
- Expected Value:  $E[Y_k] = E[\sum_{i=1}^k T_i] = kE[T_1] = \frac{k}{p}$
- Variance:  $\text{var}(Y_k) = \text{var}(\sum_{i=1}^k T_i) = \sum_{i=1}^k \text{var}(T_i) = k\frac{1-p}{p^2}$
- Probability Mass Function ( $p_{Y_k}(t)$ ):

$$p_{Y_k}(t) = P(Y_k = t) = \binom{t-1}{k-1} p^k (1-p)^{t-k}, \quad t = k, k+1, \dots$$

- Interpretation: The  $k$ -th success occurs at time  $t$  if there are exactly  $k-1$  successes in the first  $t-1$  trials AND the  $t$ -th trial is a success.

# Merging of Independent Bernoulli Processes

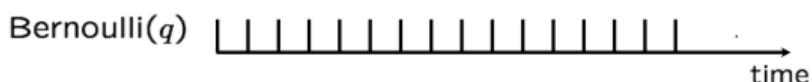
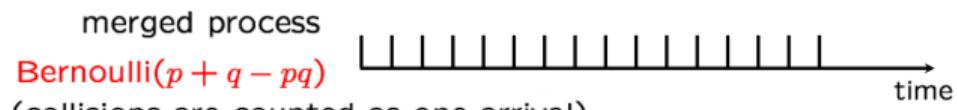
- Consider two independent Bernoulli processes:  $X_i \sim \text{Bernoulli}(p)$  and  $Z_i \sim \text{Bernoulli}(q)$ .
- **Merged Process ( $M_i = X_i$  or  $Z_i$ )**: An arrival occurs in the merged process if there is an arrival in  $X$  OR in  $Z$ .
- $P(M_i = 1) = P(X_i = 1 \cup Z_i = 1) = P(X_i = 1) + P(Z_i = 1) - P(X_i = 1 \cap Z_i = 1)$
- Since  $X_i$  and  $Z_i$  are independent we have  $P(M_i = 1) = p + q - pq$
- The merged process is also a Bernoulli process:  $\text{Bernoulli}(p + q - pq)$ .



# Merging of Independent Bernoulli Processes

- Collisions (when  $X_i = 1$  and  $Z_i = 1$ ) are counted as **one** arrival.
- If an arrival occurs in the merged process ( $M_i = 1$ ), the probability that the arrival came from the first process ( $X_i = 1$ ) is given by the conditional probability:

$$P(X_i = 1 \mid M_i = 1) = \frac{P(X_i = 1)}{P(M_i = 1)} = \frac{p}{p + q - pq}$$



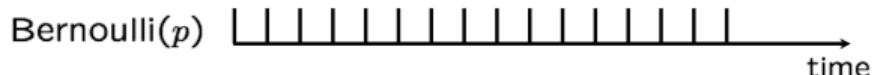
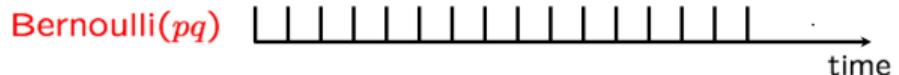
# Splitting of a Bernoulli Process

- Start with a Bernoulli process  $X_i \sim \text{Bernoulli}(p)$ .
- **Splitting rule:** Each success ( $X_i = 1$ ) is independently assigned to one of two streams (Stream 1 or Stream 2) based on an independent coin flip with bias  $q$ .
- Stream 1 (Successes assigned here):  $\text{Bernoulli}(p_1)$
- Stream 2 (Successes assigned to the complement):  $\text{Bernoulli}(p_2)$



# Splitting of a Bernoulli Process

- Probability of success in Stream 1 ( $p_1$ ):
  - $p_1 = P(X_i = 1 \text{ and assigned to Stream 1}) = P(X_i = 1) \cdot P(\text{assigned to 1}) = pq$
- Probability of success in Stream 2 ( $p_2$ ):
  - $p_2 = P(X_i = 1 \text{ and assigned to Stream 2}) = P(X_i = 1) \cdot P(\text{assigned to 2}) = p(1 - q)$
- **The two resulting streams are independent** Bernoulli processes: Bernoulli( $pq$ ) and Bernoulli( $p(1 - q)$ ).



# Poisson Approximation to Binomial

- This approximation is relevant in the **rare events** regime:
  - Large number of trials  $n$
  - Small probability of success  $p$
  - Moderate expected number of successes:  $\lambda = np$
- Number of successes  $S$  in  $n$  slots:

$$p_S(k) = P(S = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, \dots, n$$

- **Fact:** Substitute  $p = \lambda/n$  and take the limit as  $n \rightarrow \infty$ .
- For any fixed  $k \geq 0$ , the binomial probability  $p_S(k)$  converges to the **Poisson PMF**:

$$\lim_{n \rightarrow \infty} p_S(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \dots$$

- The term  $\lim_{n \rightarrow \infty} (1 - \lambda/n)^{n-k} = e^{-\lambda}$  is used in the derivation.

## **Lecture 25: The Poisson Process Part I**

---

# Definition of the Poisson process

- Poisson process is the continuous-time analogue of the Bernoulli process.
- **Rate  $\lambda$ :** “Arrival rate” per unit time.
- **Independent Increments:** Numbers of arrivals in disjoint time intervals are independent.
- **Time Homogeneity:**  $\lambda$  is constant over time.

## Bernoulli



- Independence
- **Time homogeneity:**  
Constant  $p$  at each slot

## Definition of the Poisson process (cont.)

- $P(k, \tau)$  = Probability of  $k$  arrivals in an interval of duration  $\tau$ .
- **Small Interval Probabilities:** For a VERY small duration  $\delta$ :

$$P(k, \delta) \approx \begin{cases} 1 - \lambda\delta & \text{if } k = 0 \\ \lambda\delta & \text{if } k = 1 \\ 0 & \text{if } k > 1 \end{cases}$$

- More rigorously, including terms of order  $\delta^2$ , denoted  $O(\delta^2)$ :

$$P(k, \delta) = \begin{cases} 1 - \lambda\delta + O(\delta^2) & \text{if } k = 0 \\ \lambda\delta + O(\delta^2) & \text{if } k = 1 \\ 0 + O(\delta^2) & \text{if } k > 1 \end{cases}$$

# Applications of the Poisson process

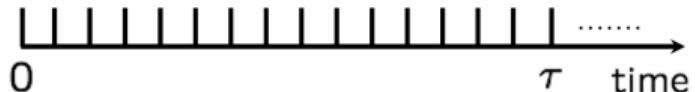
- Originally applied to rare events in the military, e.g., deaths from horse kicks in the Prussian army (1898).
- Particle emissions and radioactive decay.
- Photon arrivals from a weak source (optical communication).
- Financial market shocks (e.g., extreme price changes).
- Placement of phone calls, service requests to a help desk.



# The Poisson PMF for the number of arrivals

- Number of arrivals in  $[0, \tau]$  is  $N_\tau$ . We want  $P(k, \tau) = P(N_\tau = k)$ .
- Relate to Bernoulli: Divide time  $\tau$  into  $n = \tau/\delta$  small slots, each with success probability  $p = \lambda\delta + O(\delta^2)$ .
- $N_\tau$  is approximately Binomial( $n, p$ ).
- Limit: Let  $n \rightarrow \infty$  and  $p \rightarrow 0$  such that the mean  $np$  remains fixed at  $np = \lambda\tau$ .
- The Binomial PMF converges to the **Poisson PMF** (Poisson Approximation to Binomial):

$$P(k, \tau) = P(N_\tau = k) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \dots$$



# Mean and variance of the number of arrivals

- The number of arrivals  $N_\tau$  in a time interval of duration  $\tau$  follows Poisson( $\lambda\tau$ ).
- The mean and variance of a Poisson random variable are equal to its parameter,  $\lambda\tau$ .
- Expected Value:

$$E[N_\tau] = \lambda\tau$$

- Variance:

$$\text{var}(N_\tau) = \lambda\tau$$

- Derivation of  $E[N_\tau]$ :

$$E[N_\tau] = \sum_{k=0}^{\infty} k \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}$$

The calculation simplifies, yielding  $\lambda\tau$ .

## Example: Mail arrivals

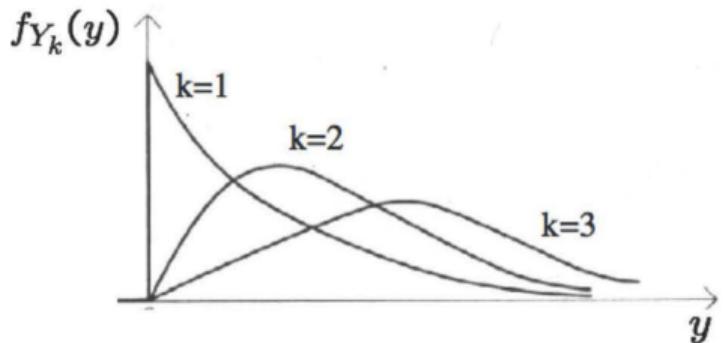
- Email arrivals follow a Poisson process at rate  $\lambda = 5$  messages per hour.
- **Time unit is 1 hour.**
- **Mean and variance of mails received during a day (24 hours):**
  - $\tau = 24$  hours.
  - $\lambda\tau = (5)(24) = 120$ .
  - $E[N_{24}] = 120$  messages;  $\text{var}(N_{24}) = 120$ .
- **P(one new message in the next hour):**
  - $\tau = 1$  hour.
  - $P(N_1 = 1) = \frac{(\lambda\tau)^1 e^{-\lambda\tau}}{1!} = \frac{5^1 e^{-5}}{1!} \approx 0.0337$ .
- **P(exactly two messages during each of the next three hours):**
  - $N_{H_1}, N_{H_2}, N_{H_3}$  are independent (disjoint intervals).
  - $P(N_{H_1} = 2, N_{H_2} = 2, N_{H_3} = 2) = P(N_{H_1} = 2) \cdot P(N_{H_2} = 2) \cdot P(N_{H_3} = 2)$
  - $P(N_1 = 2) = \frac{5^2 e^{-5}}{2!} = \frac{25 e^{-5}}{2} \approx 0.0842$ .
  - Total probability  $\approx (0.0842)^3 \approx 0.000597$ .

# The time $T_1$ until the first arrival

- $T_1$  is the time of the first arrival.  $T_1$  is a continuous random variable.
- Find the CDF:  $P(T_1 \leq t)$ :
  - The event  $\{T_1 > t\}$  means “no arrivals in  $[0, t]$ ”.
  - $P(T_1 > t) = P(N_t = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$ .
  - $P(T_1 \leq t) = 1 - P(T_1 > t) = 1 - e^{-\lambda t}$ , for  $t \geq 0$ .
- This is the CDF of the **Exponential distribution** with parameter  $\lambda$ .
- PDF is  $f_{T_1}(t) = \frac{d}{dt} P(T_1 \leq t) = \lambda e^{-\lambda t}$  for  $t \geq 0$ .

# The time $Y_k$ of the $k$ -th arrival

- $Y_k$  is the time of the  $k$ -th arrival.
- Can be derived by first finding the CDF:  $P(Y_k \leq y) = P(N_y \geq k)$ .
- **More intuitive argument** (using small time interval  $\delta$ ):
  - $f_{Y_k}(y)\delta \approx P(y \leq Y_k \leq y + \delta)$
  - This event occurs if: ( $k - 1$  arrivals in  $[0, y]$ ) AND (1 arrival in  $[y, y + \delta]$ ).
  - $P(N_y = k - 1) \cdot P(N_{[y, y+\delta]} = 1) \approx \left[ \frac{(\lambda y)^{k-1} e^{-\lambda y}}{(k-1)!} \right] \cdot [\lambda \delta]$
- Dividing by  $\delta$ , we get the PDF:
- $f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}$ , for  $y \geq 0$ .
- This is the **Erlang distribution** (or Gamma distribution).



# Memorylessness and the fresh-start property

- The Poisson process exhibits **memorylessness** and a **fresh-start property** analogous to the Bernoulli process.
- This is plausible due to the relationship between the two processes as the Bernoulli limit.
- These properties can be proved rigorously using the definition of the Poisson process.
- We often use intuitive reasoning based on the independence of disjoint intervals.

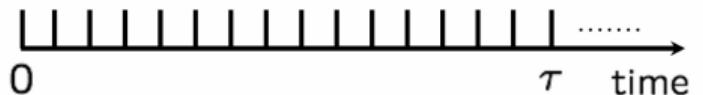
## Memorylessness and the fresh-start property (cont.)

- **Fresh-start at fixed time  $t$ :** If we start observing at time  $t$ , the future sequence of arrivals is a Poisson process with rate  $\lambda$ , independent of the history until time  $t$ .
- **Time until the next arrival:** If we have observed no arrival until time  $t$ , the remaining time  $T_1 - t$  until the next arrival is still **Exponential( $\lambda$ )**, confirming memorylessness.
- **Fresh-start at random time  $T_1$ :** The time between the first and second arrival,  $T_2 = Y_2 - Y_1$ , is:
  - **Exponential( $\lambda$ )**, independent of  $T_1$ .
- Similarly, all interarrival times  $T_k = Y_k - Y_{k-1}$ , for  $k \geq 2$ , are **i.i.d. Exponential( $\lambda$ )**.
- $Y_k = T_1 + \dots + T_k$  is the sum of  $k$  i.i.d. exponentials.
- $E[Y_k] = k/\lambda$  and  $\text{var}(Y_k) = k/\lambda^2$ .
- The property of having i.i.d. Exponential interarrival times can serve as an **equivalent definition** of the Poisson process.

# Bernoulli/Poisson relation

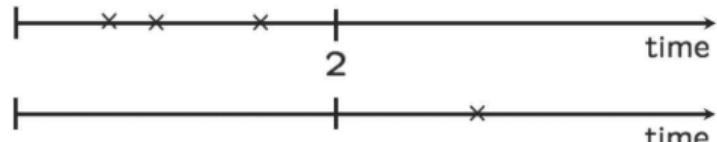
- The Poisson process is the continuous limit of the Bernoulli process.
- Discretization: Time  $\tau$  is split into  $n$  slots of length  $\delta$ , with  $p = \lambda\delta$  and  $np = \lambda\tau$ .

Property	POISSON (Continuous)	BERNOULLI (Discrete)
Times of Arrival	Continuous	Discrete
Arrival Rate	$\lambda$ /unit time	$p$ /per trial
PMF of # of Arrivals	Poisson( $\lambda\tau$ )	Binomial( $n, p$ )
Interarrival Time Distr.	Exponential( $\lambda$ )	Geometric( $p$ )
Time to $k$ -th arrival	Erlang( $k, \lambda$ )	Pascal (Negative Binomial)



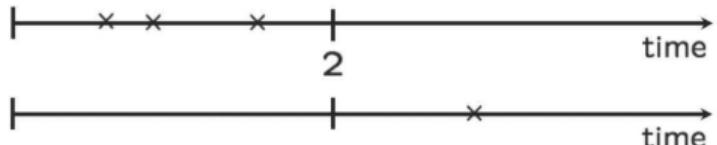
## Example: Poisson fishing (Part 1)

- Fish are caught as a Poisson process,  $\lambda = 0.6/\text{hour}$ .
- **Stopping Rule:** Fish for  $\tau = 2$  hours. If  $N_2 \geq 1$ , stop. Else, continue until the first fish is caught (i.e., until  $T_1$  occurs).
- $\lambda\tau = 0.6 \cdot 2 = 1.2$ .
- **P(fish for more than two hours):**
  - This event occurs if and only if no fish are caught in the first two hours:  $N_2 = 0$ .
  - $P(N_2 = 0) = \frac{(\lambda\tau)^0 e^{-\lambda\tau}}{0!} = e^{-1.2} \approx 0.301$ .
- **P(fish for more than two and less than five hours):**
  - This occurs if:  $(N_2 = 0)$  AND  $(2 < T_1 < 5)$ .
  - Due to memorylessness,  $P(T_1 \in (2, 5) | T_1 > 2) = P(T_1 \in (0, 3))$ .
  - $P(2 < T_1 < 5) = P(T_1 > 2) - P(T_1 > 5) = e^{-2\lambda} - e^{-5\lambda} = e^{-1.2} - e^{-3.0}$ .
  - $P(\text{more than 2 and less than 5 hours}) = P(N_2 = 0) \cdot P(T_1 \in (2, 5) | N_2 = 0)$ .
  - ... Wait, this is just  $P(2 < T_1 < 5)$  since  $N_2 = 0$  is equivalent to  $T_1 > 2$ .
  - $P(2 < T_1 < 5) = e^{-1.2} - e^{-3.0} \approx 0.301 - 0.050 = 0.251$ .



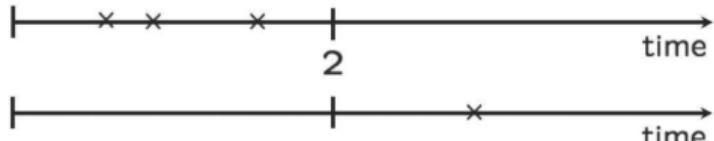
## Example: Poisson fishing (Part 2)

- $\lambda = 0.6/\text{hour}$ . Stopping Rule: If  $N_2 \geq 1$ , stop. Else, continue until  $T_1$ .
- **P(catch at least two fish):**
  - Case 1:  $N_2 \geq 2$ .  $P(N_2 \geq 2) = 1 - P(N_2 = 0) - P(N_2 = 1)$ .
  - $P(N_2 = 0) = e^{-1.2} \approx 0.301$ .
  - $P(N_2 = 1) = \lambda \tau e^{-\lambda \tau} = 1.2e^{-1.2} \approx 0.361$ .
  - $P(N_2 \geq 2) \approx 1 - 0.301 - 0.361 = 0.338$ .
  - Case 2:  $N_2 = 1$ . The rule stops after 2 hours.
  - Total  $P(\text{catch} \geq 2) = P(N_2 \geq 2) \approx 0.338$ . (The rule ensures  $\geq 1$  fish, but not  $\geq 2$ ).
- **E[future fishing time | already fished for three hours]:**
  - If 3 hours have passed, the initial rule is long over. We assume the question asks for the expected time until the next fish, given no fish has been caught in the first 3 hours.
  - Due to memorylessness of the Exponential interarrival time  $T_1$ ,
  - $$E[T_1 - 3 \mid T_1 > 3] = E[T_1] = 1/\lambda.$$
  - $E[\text{future time}] = 1/0.6 = 5/3 \approx 1.667 \text{ hours.}$



## Example: Poisson fishing (Part 3)

- $\lambda = 0.6/\text{hour}$ . Stopping Rule: Fish 2 hours. If  $N_2 \geq 1$ , stop. Else, continue until  $T_1$ .
- **E[total fishing time  $T_{\text{total}}$ ]:**
  - $T_{\text{total}} = 2$  if  $N_2 \geq 1$  (Prob  $1 - e^{-1.2}$ ).
  - $T_{\text{total}} = T_1$  if  $N_2 = 0$  (Prob  $e^{-1.2}$ ).
  - $E[T_{\text{total}}] = E[T_{\text{total}}|N_2 \geq 1]P(N_2 \geq 1) + E[T_{\text{total}}|N_2 = 0]P(N_2 = 0)$ .
  - $E[T_{\text{total}}] = 2 \cdot (1 - e^{-1.2}) + E[T_1|T_1 > 2] \cdot e^{-1.2}$ .
  - Memorylessness:  $E[T_1|T_1 > 2] = 2 + E[T_1] = 2 + 1/\lambda = 2 + 5/3 = 11/3$ .
  - $E[T_{\text{total}}] = 2(1 - e^{-1.2}) + (11/3)e^{-1.2}$ .
  - $E[T_{\text{total}}] = 2 + (11/3 - 2)e^{-1.2} = 2 + (5/3)e^{-1.2} \approx 2 + (1.667)(0.301) \approx 2.502 \text{ hours}$ .
- **E[number of fish  $N_{\text{total}}$ ]:**
  - If  $N_2 = k \geq 1$ ,  $N_{\text{total}} = k$ . If  $N_2 = 0$ ,  $N_{\text{total}} = 1$ .
  - $N_{\text{total}} = N_2 + I_{\{N_2=0\}}$ .
  - $E[N_{\text{total}}] = E[N_2] + E[I_{\{N_2=0\}}] = E[N_2] + P(N_2 = 0)$ .
  - $E[N_{\text{total}}] = \lambda\tau + e^{-\lambda\tau} = 1.2 + e^{-1.2} \approx 1.2 + 0.301 = 1.501 \text{ fish}$ .



## **Lecture 26: The Poisson Process Part II**

---

# The sum of independent Poisson random variables

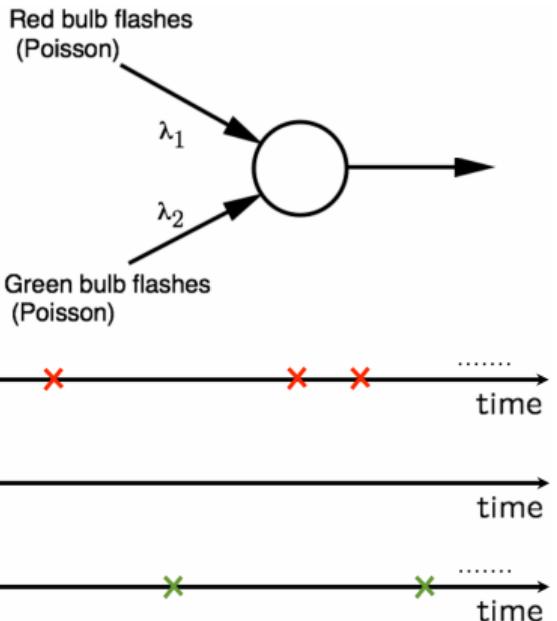
- Consider a Poisson process of rate  $\lambda = 1$ .
- Let  $M$  be the number of arrivals in an interval of length  $\tau$ .
- $M$  is a Poisson random variable with parameter  $\lambda\tau = \tau$ .
- Let  $N$  be the number of arrivals in a subsequent interval of length  $\nu$ .
- $N$  is a Poisson random variable with parameter  $\lambda\nu = \nu$ .
- Are  $M$  and  $N$  independent? Yes, by the **independent increments** property of the Poisson process.
- The sum  $M + N$ :  $M + N$  is the number of arrivals in the combined interval of length  $\tau + \nu$ .
- $M + N$  is a Poisson random variable with parameter  $\tau + \nu$ .



- **Theorem:** The sum of independent Poisson random variables, with means/parameters  $\mu$  and  $\nu$ , is Poisson with mean/parameter  $\mu + \nu$ .
- $P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}$ .

# Merging of independent Poisson processes

- Consider two independent Poisson processes with rates  $\lambda_1$  and  $\lambda_2$ .
- Example: Red bulb flashes (Poisson  $\lambda_1$ ) and Green bulb flashes (Poisson  $\lambda_2$ ).
- The **merged process** counts all flashes (Red OR Green).
- In a very small interval  $\delta$ :
  - $P(\text{Red arrival}) \approx \lambda_1\delta$ .  $P(\text{Green arrival}) \approx \lambda_2\delta$ .
  - $P(\text{Combined arrival}) \approx P(\text{Red arrival}) + P(\text{Green arrival})$  (since the probability of a simultaneous arrival is  $O(\delta^2)$ ).
  - $P(\text{Combined arrival}) \approx (\lambda_1 + \lambda_2)\delta$ .
- The merged process is a **Poisson process** with rate  $\lambda_1 + \lambda_2$ .



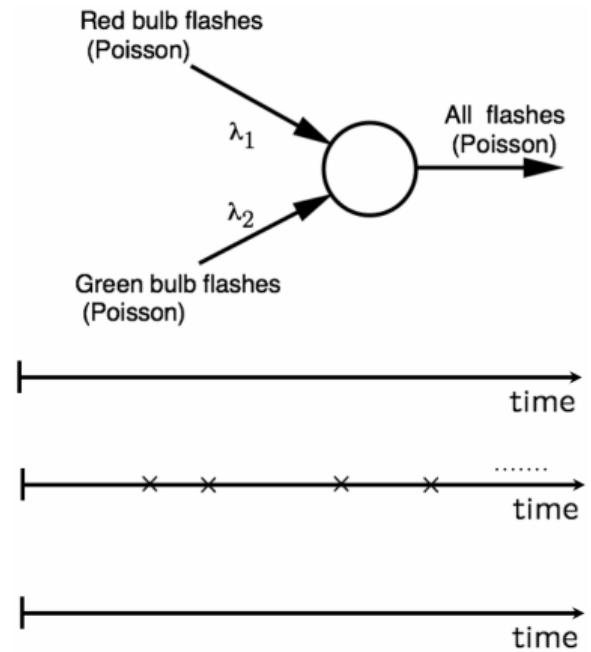
# Where is an arrival of the merged process coming from?

- If an arrival occurs at time  $t$  in the merged process (rate  $\lambda_1 + \lambda_2$ ), the probability that it originated from the Red process (rate  $\lambda_1$ ) is:

$$P(\text{Red} \mid \text{arrival at time } t) = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

- $P(k\text{-th arrival is Red}) = \frac{\lambda_1}{\lambda_1 + \lambda_2}$ .
- The decision of which source the arrival came from is **independent for different arrivals**.
- Example:**  $P(4 \text{ out of first 10 arrivals are Red})$ .

- Each arrival is independently “Red” with probability  $p = \frac{\lambda_1}{\lambda_1 + \lambda_2}$ .
- The number of Red arrivals follows a **Binomial distribution** with parameters  $n = 10$  and  $p$ .
- $P(4 \text{ Red arrivals}) = \binom{10}{4} p^4 (1 - p)^6$ .



# The time the first light bulb burns out

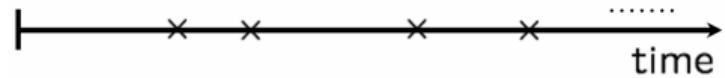
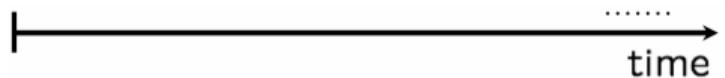
- Consider three lightbulbs with independent lifetimes  $X, Y, Z$ .
- Assume  $X \sim \text{Exponential}(\lambda_X)$ ,  $Y \sim \text{Exponential}(\lambda_Y)$ ,  $Z \sim \text{Exponential}(\lambda_Z)$ .
- The time until the first burnout is  $\min\{X, Y, Z\}$ .
- Exponential lifetimes correspond to the **first arrival** in independent Poisson processes with rates  $\lambda_X, \lambda_Y, \lambda_Z$ .
- The time  $\min\{X, Y, Z\}$  is the time until the **first arrival in the merged process**.
- The merged process is Poisson with rate  $\lambda = \lambda_X + \lambda_Y + \lambda_Z$ .
- Therefore,  $\min\{X, Y, Z\}$  is **Exponential** with rate  $\lambda_X + \lambda_Y + \lambda_Z$ .
- The expected time until the first burnout is  $E[\min\{X, Y, Z\}] = \frac{1}{\lambda_X + \lambda_Y + \lambda_Z}$ .

# The time the last light bulb burns out

- Three lightbulbs with independent lifetimes  $X, Y, Z$ .
- Find expected time until **all** burn out:  $E[\max\{X, Y, Z\}]$ .
- This calculation is generally more complex than for the minimum.
- If  $\lambda_X = \lambda_Y = \lambda_Z = \lambda$ : Let  $M = \max\{X, Y, Z\}$ .
- The CDF is  $P(M \leq t) = P(X \leq t, Y \leq t, Z \leq t) = (1 - e^{-\lambda t})^3$ .
- The expected value requires integrating  $1 - P(M \leq t)$  or  $t \cdot f_M(t)$ .
- The result for  $\lambda_X = \lambda_Y = \lambda_Z = \lambda$  is  $E[\max\{X, Y, Z\}] = \frac{1}{\lambda} + \frac{1}{2\lambda} + \frac{1}{3\lambda} = \frac{11}{6\lambda}$ .

# Splitting of a Poisson process

- Start with a Poisson process of rate  $\lambda$ .
- **Splitting Rule:** Each arrival is independently classified into Stream 1 (with probability  $q$ ) or Stream 2 (with probability  $1 - q$ ).
- The probability  $q$  is independent of the original process.
- **Result:** The resulting streams are **Poisson processes**.
- Stream 1 rate:  $\lambda_1 = \lambda q$ .
- Stream 2 rate:  $\lambda_2 = \lambda(1 - q)$ .
- **Property:** The two resulting streams (Poisson  $\lambda q$  and Poisson  $\lambda(1 - q)$ ) are **independent**.



## “Random incidence” in the Poisson process

- Consider a Poisson process (or any renewal process) that has been running forever.
- Interarrival times  $T_k$  are i.i.d. Exponential with  $E[T_k] = 1/\lambda$ .
- Example:  $\lambda = 4/\text{hour}$ .  $E[T_k] = 1/4 \text{ hour} = 15 \text{ minutes}$ .
- You show up at a “random time”  $t^*$  and measure the interarrival time  $V - U$  during which you arrived.
- Intuition suggests the observed average interarrival time should be  $E[T_k] = 15 \text{ minutes}$ .
- **Observation:** Experiments consistently show the average measured interarrival time is closer to  $\frac{1}{2} \text{ hour}$  or  $30 \text{ minutes}$ ! Why is the observed time longer?

## "Random incidence" in the Poisson process analysis

- $t^*$ : Random arrival time.
- $U$ : Last arrival time before  $t^*$ .
- $V$ : Next arrival time after  $t^*$ .
- $V - U$ : The interarrival time observed upon arrival.
- $V - U$  is **not** identically distributed as the typical interarrival time  $T_k$ .
- **Interarrival time**  $T_k$ : Exponential( $\lambda$ ),  $E[T_k] = 1/\lambda$ .
- **Observed time**  $V - U$ : Gamma(2,  $\lambda$ ) or Erlang(2,  $\lambda$ ).
- Expected length of the observed interval:  $E[V - U] = 2/\lambda$ .
- For  $\lambda = 4/\text{hour}$ ,  $E[V - U] = 2/4 = 1/2 \text{ hour} = 30 \text{ minutes}$ .

## Random incidence “paradox” is not special to the Poisson process

- The “paradox” arises because the sampling method is biased toward **longer** intervals.
- **General Example:** Interarrival times  $T$  i.i.d., equally likely to be 5 or 10 minutes.
- Expected value of  $k$ -th interarrival time:  $E[T] = \frac{1}{2}(5) + \frac{1}{2}(10) = 7.5$  minutes.
- **Bias:** The probability of arriving during a 10-minute interval is twice that of arriving during a 5-minute interval.
- $P(\text{arrive during 5-minute interval}) = \frac{5}{5+10} = \frac{1}{3}$ .
- $P(\text{arrive during 10-minute interval}) = \frac{10}{5+10} = \frac{2}{3}$ .
- Expected length of interarrival interval during which you arrive:

$$E[\text{Observed Length}] = 5 \cdot \frac{1}{3} + 10 \cdot \frac{2}{3} = \frac{5 + 20}{3} = \frac{25}{3} \approx 8.33 \text{ minutes}$$

- The expected observed length is always  $\geq E[T]$ .
- This generalizes to **renewal processes** (i.i.d. interarrival times from some general distribution).

# Different sampling methods can give different results

- The discrepancy highlights that the **sampling method matters** and leads to different expectations (or weighted averages).
- **Average family size?:**
  - Look at a “random” family (uniform probability over families): This gives the true average.
  - Look at a “random” person’s family (weighted by family size): This is biased toward larger families.
- **Average bus occupancy?:**
  - Look at a “random” bus (uniform probability over buses): Gives the true average occupancy.
  - Look at a “random” passenger’s bus (weighted by number of passengers): Biased toward full buses.
- **Average class size?:** The same principle applies. Sampling students biases the average toward larger classes.

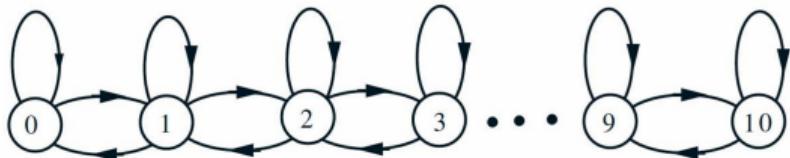
## Lecture 27: Markov Chains I

---

## Markov process definition

- A stochastic process where the future state depends only on the current state, and not on the sequence of events that preceded it (the past).
  - **Markov property/assumption:** “Given the current state, the past doesn’t matter.”
  - The state at time  $t + 1$  is a function of the state at time  $t$  and some noise:

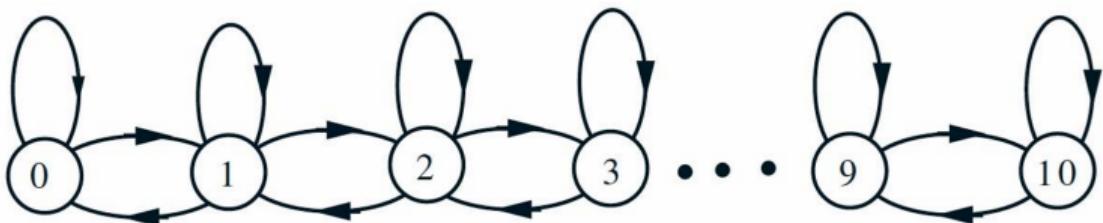
$$\text{state}(t + 1) = f(\text{state}(t), \text{noise})$$



- Time index can be discrete ( $n = 0, 1, 2, \dots$ ) or continuous ( $t \geq 0$ ).
  - State space can be finite, countable, or continuous.

# Checkout counter example

- **State**  $X_n$ : number of customers at time  $n$ .
- **Time**: Discrete time  $n = 0, 1, \dots$
- **Customer arrivals**:  $\text{Bernoulli}(p)$ . (One arrival or no arrival per time step).
- **Customer service times**:  $\text{Geometric}(q)$ . (One departure or no departure per time step).
- **Transitions**:
  - State  $i$  can transition to  $i - 1$  (departure),  $i$  (no change), or  $i + 1$  (arrival).



# Discrete-time finite state Markov chains

- **State  $X_n$ :** The state after  $n$  transitions, belonging to a finite set.
- **Initial state  $X_0$ :** Either given or random.
- **Transition probabilities:**

$$p_{ij} = P(X_1 = j | X_0 = i) = P(X_{n+1} = j | X_n = i)$$

- **Markov property/assumption** (time-homogeneous):

$$p_{ij} = P(X_{n+1} = j | X_n = i) = P(X_{n+1} = j | X_n = i, X_{n-1}, \dots, X_0)$$

- **Model specification:** Identify states, transitions, and transition probabilities.



## *n*-step transition probabilities

- ***n*-step transition probability**  $r_{ij}(n)$ : The probability of transitioning from state  $i$  to state  $j$  in exactly  $n$  steps.

$$r_{ij}(n) = P(X_n = j | X_0 = i) = P(X_{n+s} = j | X_s = i)$$

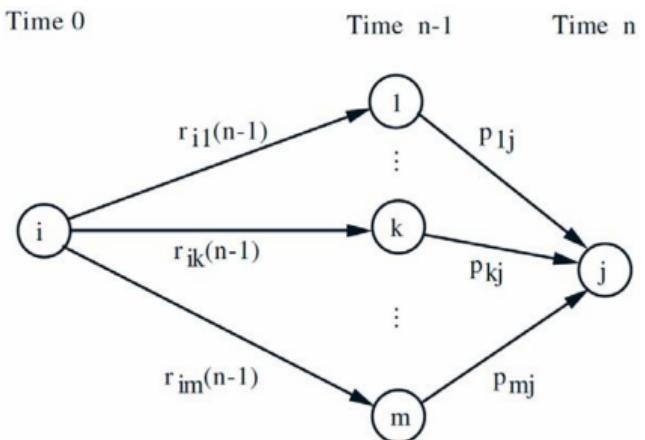
- **Key recursion (Chapman-Kolmogorov equation):**

To go from  $i$  to  $j$  in  $n$  steps, one must go from  $i$  to some intermediate state  $k$  in  $n - 1$  steps, and then take a single step from  $k$  to  $j$ .

$$r_{ij}(n) = \sum_{k=1}^m r_{ik}(n-1) p_{kj}$$

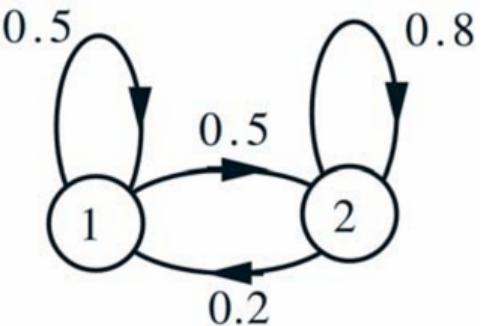
- **Random initial state:** The unconditional probability of being in state  $j$  at time  $n$  is the average over all initial states  $i$ :

$$P(X_n = j) = \sum_{i=1}^m P(X_0 = i) r_{ij}(n)$$



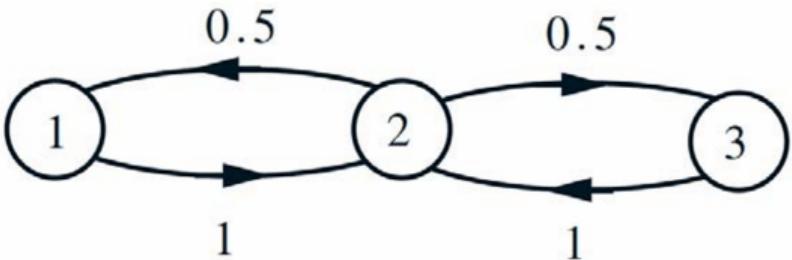
## Example: Two-state Markov Chain

- Consider a chain with states 1 and 2 and given single-step transition probabilities.
- Compute the  $n$ -step transition probabilities  $r_{ij}(n)$  for small  $n$  and discuss long-term behavior ( $n \rightarrow \infty$ ).
- Initial conditions** ( $n = 0$ ):  $r_{11}(0) = 1$ ,  $r_{12}(0) = 0$ ,  $r_{21}(0) = 0$ ,  $r_{22}(0) = 1$ .
- One step** ( $n = 1$ ):  $r_{11}(1) = 0.5$ ,  $r_{12}(1) = 0.5$ ,  $r_{21}(1) = 0.8$ ,  $r_{22}(1) = 0.2$ .
- Two steps** ( $n = 2$ ): Using the recursion  $r_{1j}(2) = r_{11}(1)p_{1j} + r_{12}(1)p_{2j}$ :
  - $r_{11}(2) = (0.5)(0.5) + (0.5)(0.8) = 0.25 + 0.40 = 0.65$ .
  - $r_{12}(2) = (0.5)(0.5) + (0.5)(0.2) = 0.25 + 0.10 = 0.35$ .



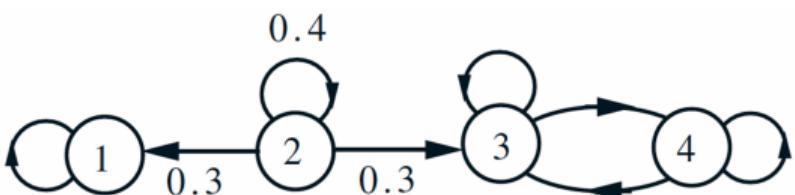
## Generic convergence questions

- **Convergence:** Does  $r_{ij}(n)$  converge to a limit as  $n \rightarrow \infty$ ?
- **Initial State Dependence:** Does the limit  $\lim_{n \rightarrow \infty} r_{ij}(n)$  depend on the initial state  $i$ ?
- **Example 1 (Non-convergence/Oscillation):** States 1, 2, 3. The process oscillates between the states.



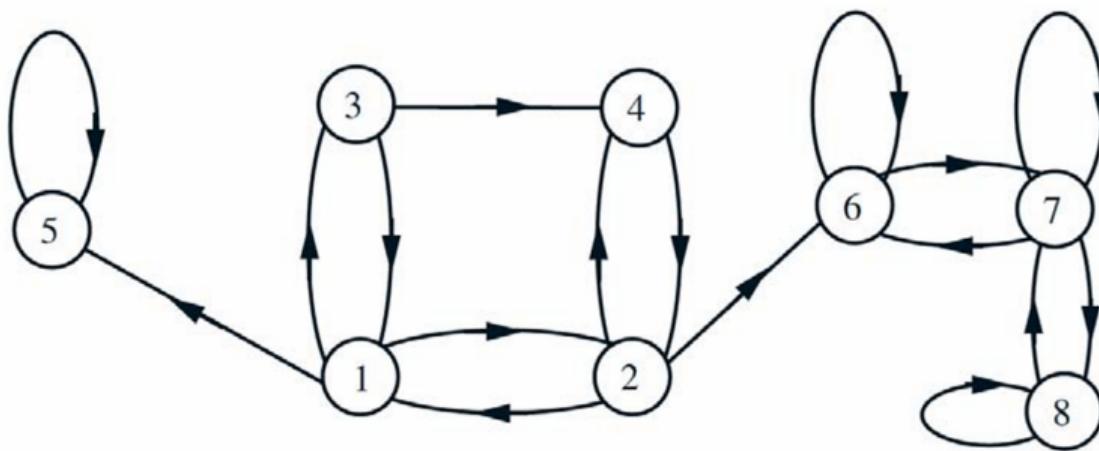
# Generic convergence questions

- **Convergence:** Does  $r_{ij}(n)$  converge to a limit as  $n \rightarrow \infty$ ?
- **Initial State Dependence:** Does the limit  $\lim_{n \rightarrow \infty} r_{ij}(n)$  depend on the initial state  $i$ ?
- **Example 2 (Multiple Recurrent Classes):** States 1, 2, 3, 4. The limit depends on whether the chain starts in  $\{1, 2\}$  or  $\{3, 4\}$ .
- $n$  odd:  $r_{22}(n) = 0$ .  $n$  even:  $r_{22}(n) = 1$ . (For a chain where  $p_{22} = 0$  and  $p_{21} = 1, p_{12} = 1$ , implying  $X_n$  alternates between 1 and 2).
- $r_{11}(n) \rightarrow 1$ .  $r_{31}(n) \rightarrow 0$ .



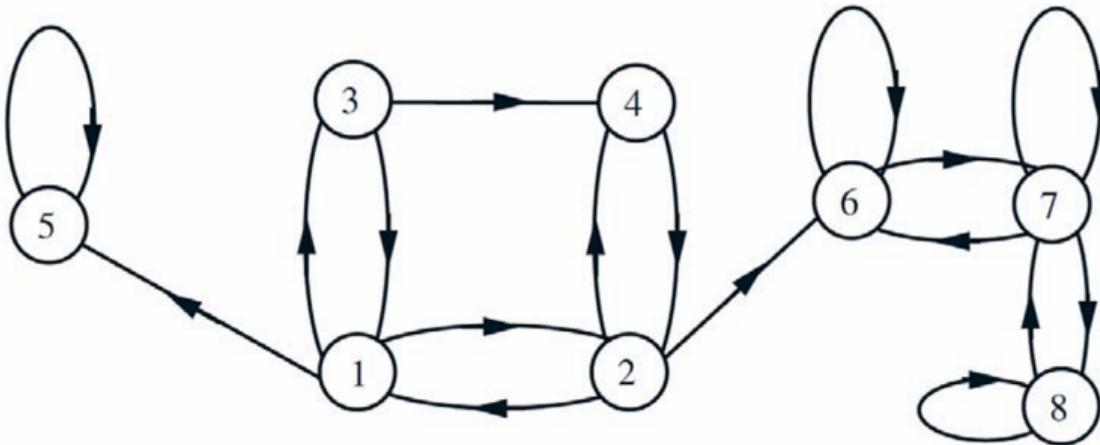
# Recurrent and transient states

- **Recurrent state  $i$ :** Starting from state  $i$  (and from wherever you can go from  $i$ ), there is a guaranteed way of returning to  $i$  with probability 1.
- **Transient state  $i$ :** If a state is not recurrent, it is called transient. Once the process leaves a transient state, it may never return.
- **Recurrent class:** A collection of recurrent states that only communicate (have non-zero transition probability) among themselves.





## Recurrent and transient states



- In the figure: States 1 and 2 are transient. State 5 is recurrent (absorbing). States  $\{3, 4\}$  form a recurrent class. States  $\{6, 7, 8\}$  form another recurrent class.
- Any path starting in a transient state (like 1 or 2) will eventually move to a recurrent class ( $3/4, 5$ , or  $6/7/8$ ) and remain there.

## Lecture 28: Markov Chains II

---



- Discrete time, discrete state space, time-homogeneous chain.
- $p_{ij}$ : One-step transition probabilities.
- Markov property: Future depends only on the present state.
- $n$ -step transition probabilities:

$$r_{ij}(n) = P(X_n = j | X_0 = i) = P(X_{n+s} = j | X_s = i)$$

- Key recursion (Chapman-Kolmogorov):

$$r_{ij}(n) = \sum_{k=1}^m r_{ik}(n-1)p_{kj}$$

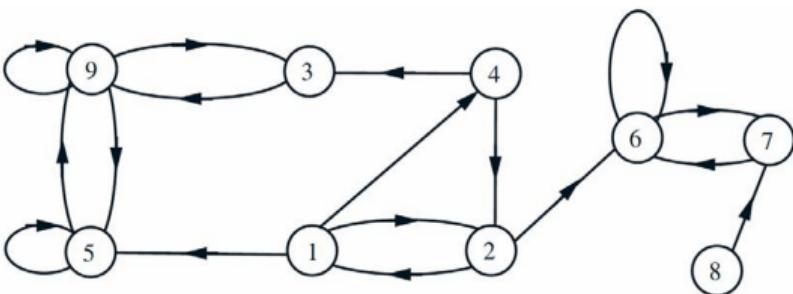
# Warm-up: Calculating Trajectory Probabilities

- **Probability of a specific trajectory (path):**

$$P(X_1 = 2, X_2 = 6, X_3 = 7 \mid X_0 = 1) = p_{12}p_{26}p_{67}$$

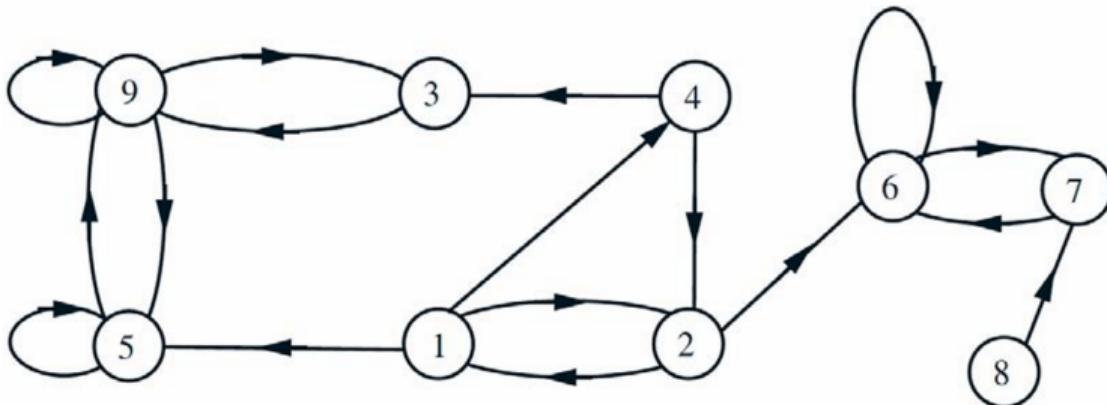
- **$n$ -step probability:**  $r_{ij}(n)$

$$P(X_4 = 7 \mid X_0 = 2) = r_{27}(4)$$



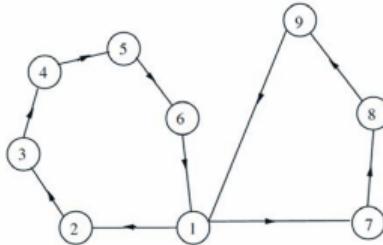
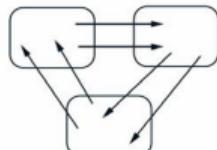
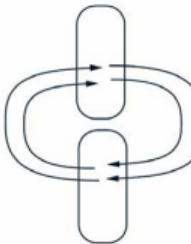
## Review: Recurrent and transient states

- **Recurrent state  $i$ :** Starting from  $i$ , the probability of returning to  $i$  is 1.
- **Transient state  $i$ :** If not recurrent, it is transient. The process may never return.
- **Recurrent class:** A collection of recurrent states that communicate only among themselves.



# Periodic states in a recurrent class

- The states in a recurrent class are **periodic** if they can be grouped into  $d > 1$  groups.
- All transitions from one group lead deterministically to the next group, forming a cycle of length  $d$ .
- Example 1 (Chain-link):** States alternate between the top and bottom groups ( $d = 2$ ).
- Example 2 (Cycle):** States  $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$  form a cycle, with period  $d = 9$  if no self-loops exist.

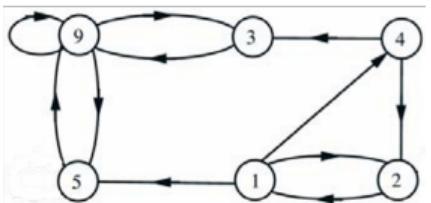
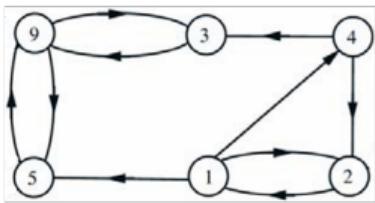
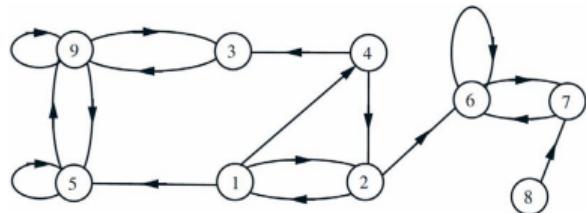


# Steady-state probabilities

- **Convergence Question:** Does the  $n$ -step probability  $r_{ij}(n)$  converge to some limit  $\pi_j$  as  $n \rightarrow \infty$ ?
- **Convergence Theorem:** Yes, if the Markov chain satisfies:
  1. Recurrent states form a **single class** (i.e., irreducible).
  2. The single recurrent class is **not periodic** (i.e., aperiodic).
- **Balance Equations:** Assuming convergence ( $\lim_{n \rightarrow \infty} r_{ij}(n) = \pi_j$ ), we take the limit of the recursion:

$$\lim_{n \rightarrow \infty} r_{ij}(n) = \sum_k \lim_{n \rightarrow \infty} r_{ik}(n-1)p_{kj} \implies \pi_j = \sum_k \pi_k p_{kj}$$

- **Normalization:** We also need  $\sum_j \pi_j = 1$  (total probability).



## Example: Steady-state calculation

- **Balance Equations:**

1.  $\pi_1 = \pi_1 p_{11} + \pi_2 p_{21} \implies \pi_1 = 0.5\pi_1 + 0.2\pi_2$
2.  $\pi_2 = \pi_1 p_{12} + \pi_2 p_{22} \implies \pi_2 = 0.5\pi_1 + 0.8\pi_2$

- **Normalization:**  $\pi_1 + \pi_2 = 1$

- **Solving for  $\pi_1$  and  $\pi_2$ :**

- From equation (1):

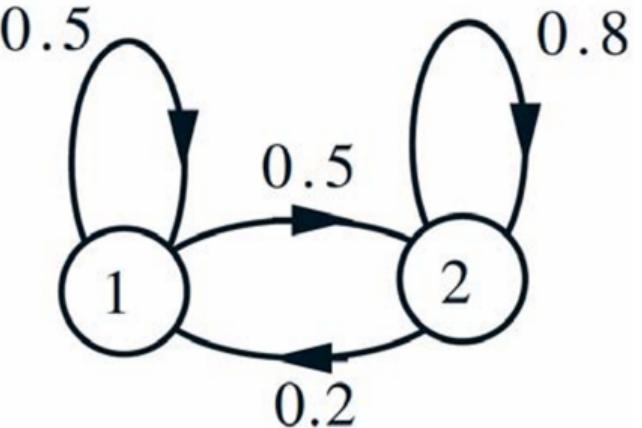
$$0.5\pi_1 = 0.2\pi_2 \implies \pi_2 = 2.5\pi_1.$$

- Substitute into normalization:

$$\pi_1 + 2.5\pi_1 = 1 \implies 3.5\pi_1 = 1.$$

- $\pi_1 = 1/3.5 = 2/7.$

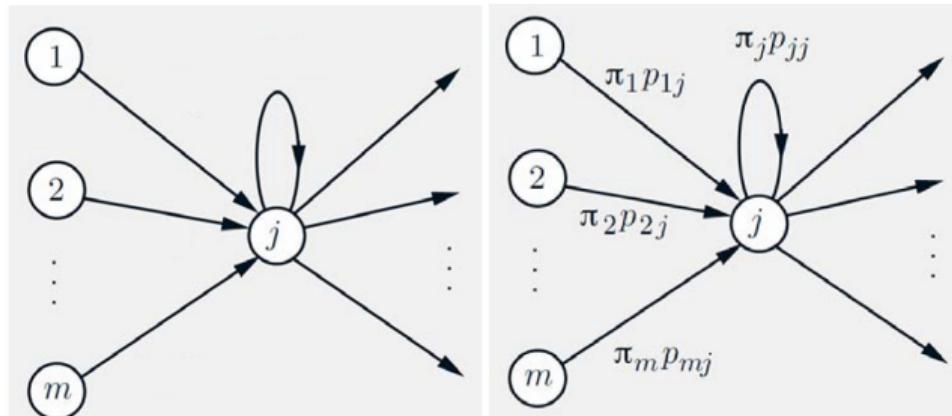
- $\pi_2 = 1 - 2/7 = 5/7.$



# Visit frequency interpretation: Balance Equations

- $\pi_j$  represents the **long run frequency** (or probability) of being in state  $j$ .
- **Frequency of transitions into  $j$ :**  $\sum_k \pi_k p_{kj}$ .
- **Frequency of transitions out of  $j$ :**  $\pi_j \sum_k p_{jk} = \pi_j \cdot 1 = \pi_j$ .
- **Balance Principle:** In steady-state, the long-run frequency of entering state  $j$  must equal the long-run frequency of leaving state  $j$ .

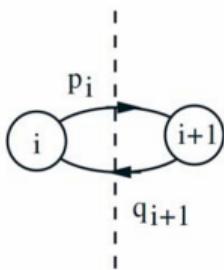
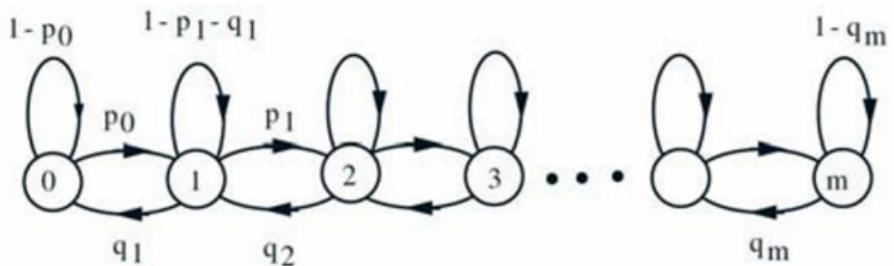
$$\sum_k \pi_k p_{kj} = \pi_j$$



# Birth-Death processes I

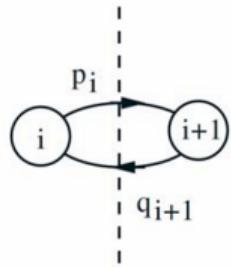
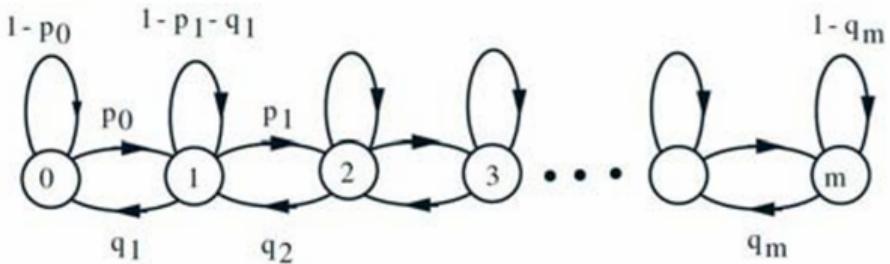
- **Definition:** A Markov chain where transitions are restricted to adjacent states (i.e., only  $\pm 1$  or 0).
- $p_i$ : Probability of moving from  $i$  to  $i + 1$  (birth).
- $q_i$ : Probability of moving from  $i$  to  $i - 1$  (death).
- The queue model (checkout counter) is a birth-death process.
- **Flow Balance:** For any two adjacent states  $i$  and  $i + 1$ , the long-run frequency of transitions from  $i \rightarrow i + 1$  must equal the frequency of transitions from  $i + 1 \rightarrow i$ .

$$\pi_i p_i = \pi_{i+1} q_{i+1} \quad \text{for } i = 0, 1, \dots, m-1$$



# Birth-Death processes II

- **Steady-State Solution:** The local balance equations ( $\pi_i p_i = \pi_{i+1} q_{i+1}$ ) can be solved recursively.
- **Special Case:**  $p_i = p$  and  $q_i = q$  (**constant rates**):
  - Let  $\rho = p/q$  (the ratio of birth to death rate).
  - $\pi_{i+1} = \pi_i \frac{p}{q} = \pi_i \rho$ .
  - This leads to a geometric probability distribution:  $\pi_i = \pi_0 \rho^i$ .
- **Infinite Case**  $m \approx \infty$  (**and**  $p < q$ ):
  - The normalization  $\sum_{i=0}^{\infty} \pi_i = 1$  requires  $p < q$  ( $\rho < 1$ ).
  - $\pi_0$  solves the geometric sum:  $\pi_0 = 1 - \rho$ .
  - Expected steady-state size:  $E[X_n] = \sum_{i=0}^{\infty} i \pi_i = \frac{\rho}{1-\rho}$ .



## Lecture 29: Markov Chains III

---

# Review of steady state behavior

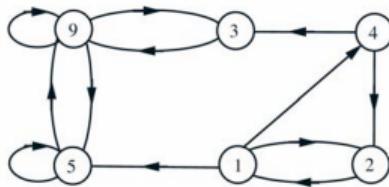
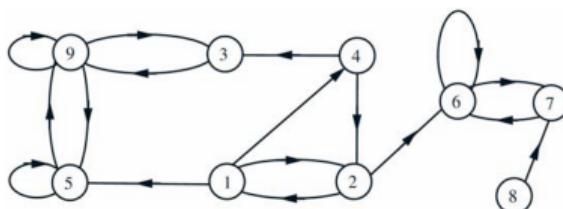
- **Convergence:** For an aperiodic Markov chain with a single class of recurrent states, the  $n$ -step probability converges:

$$\lim_{n \rightarrow \infty} r_{ij}(n) = \lim_{n \rightarrow \infty} P(X_n = j | X_0 = i) = \pi_j, \quad \forall i$$

- **Solution:** The steady-state distribution  $\{\pi_j\}$  is the unique solution to the **balance equations**:

$$\pi_j = \sum_k \pi_k p_{kj}, \quad j = 1, \dots, m$$

- **Normalization:** Together with the normalization condition  $\sum_j \pi_j = 1$ .



# On the use of steady state probabilities

- Assume  $\pi_1 = 2/7, \pi_2 = 5/7$  for the given two-state chain ( $p_{11} = 0.5, p_{21} = 0.2$ ).
- $P(X_1 = 1 \text{ and } X_{100} = 1 | X_0 = 1)$ :

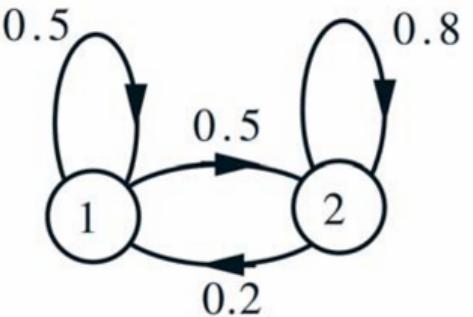
$$p_{11} \cdot P(X_{100} = 1 | X_1 = 1) = p_{11} \cdot r_{11}(99) \approx p_{11}\pi_1$$

- $P(X_{100} = 1 \text{ and } X_{101} = 2 | X_0 = 1)$ :

$$P(X_{100} = 1 | X_0 = 1) \cdot P(X_{101} = 2 | X_{100} = 1) \approx \pi_1 \cdot p_{12}$$

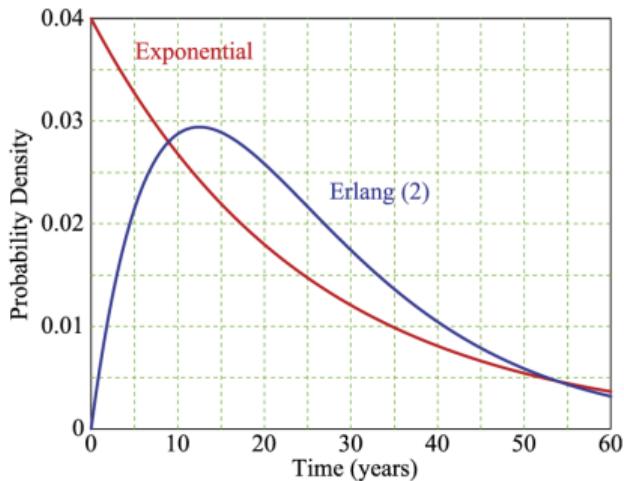
- $P(X_{100} = 1 \text{ and } X_{200} = 1 | X_0 = 1)$ :

$$P(X_{100} = 1 | X_0 = 1) \cdot P(X_{200} = 1 | X_{100} = 1) \approx \pi_1 \cdot \pi_1$$



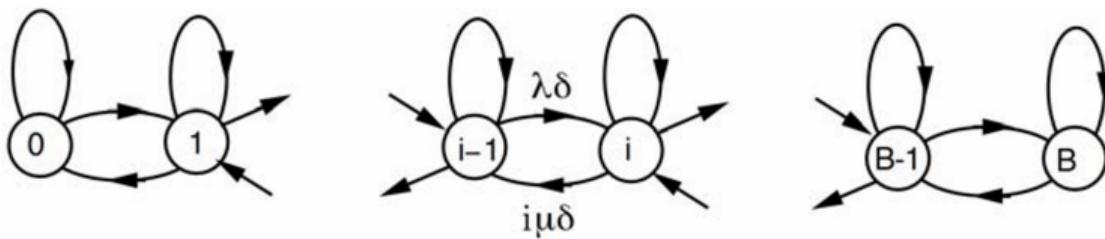
# Design of a phone system

- **Model:** Calls originate as a Poisson process (rate  $\lambda$ ).
- Each call duration is Exponential (parameter  $\mu$ ).
- Goal: Decide on the number of lines,  $B$ , to minimize blocking.
- **Discrete-time approximation** (small duration  $\delta$ ):
  - $P(\text{new call arrives}) \approx \lambda\delta$ .
  - If  $i$  active calls,  $P(\text{a departure}) \approx i\mu\delta$  (superposition of  $i$  Exponential distributions).



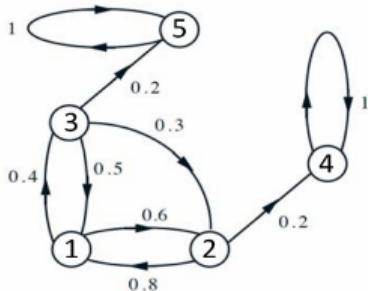
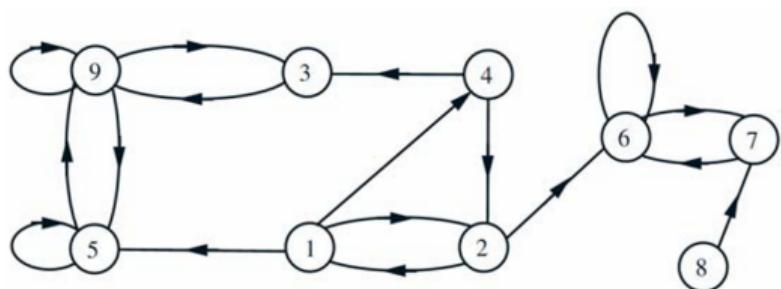
# Phone system: discrete time approximation

- **State  $i$ :** Number of active calls ( $0 \leq i \leq B$ ).
- **Birth-Death Process:** Transitions only to  $i \pm 1$ .
- **Arrival Rate:**  $\lambda_i \approx \lambda$  (for  $i < B$ ).
- **Departure Rate:**  $\mu_i \approx i\mu$  (for  $i > 0$ ).
- **Local Balance Equations** (Continuous-time rates):  $\lambda\pi_{i-1} = i\mu\pi_i$  for  $i = 1, \dots, B$
- **Steady-State Probability  $\pi_i$ :**  $\pi_i = \pi_0 \frac{\rho^i}{\mu^i i!}$  where  $\rho = \lambda/\mu$  (traffic intensity)
- $\pi_0 = 1 / \sum_{i=0}^B \frac{\rho^i}{i!}$  is the normalization constant.
- **Blocking Probability:**  $P(\text{arriving customer finds busy system}) = \pi_B$  (Erlang B Formula).



# Calculating absorption probabilities

- **Absorbing state  $k$ :** A recurrent state  $k$  with  $p_{kk} = 1$ .
- **Absorption Probability  $a_i$ :** Probability that the chain eventually settles in absorbing state 4, given it started in  $i$ .
- **Boundary Conditions:**
  - $i = 4 \implies a_4 = 1$
  - $i = 5 \implies a_5 = 0$  (5 is another absorbing state,  $p_{55} = 1$ ).
- **Governing Equation** (for transient states  $i \neq 4, 5$ ):  $a_i = \sum_j p_{ij} a_j$
- The system of linear equations yields a unique solution for  $a_i$ .



# Expected time to absorption

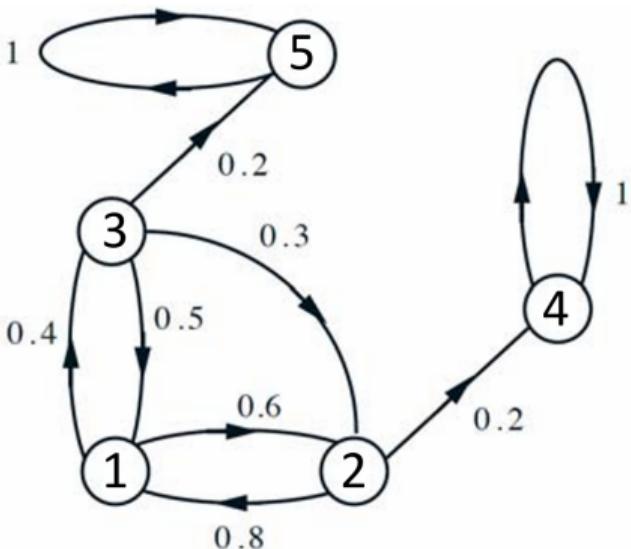
- **Expected Time**  $\mu_i$ : Expected number of transitions until reaching the absorbing state 4, given the initial state is  $i$ .

- **Boundary Conditions**:  $\mu_i = 0$  for  $i = 4, 5$  (absorbing states).

- **Governing Equation** (for transient states  $i \neq 4, 5$ ):

$$\mu_i = 1 + \sum_j p_{ij} \mu_j$$

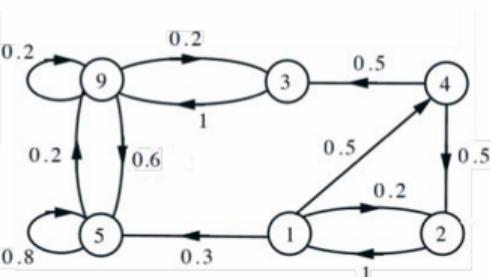
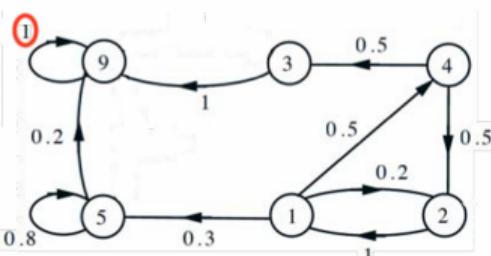
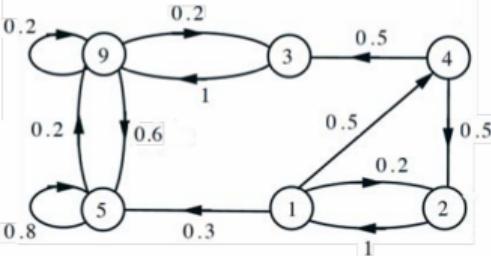
- The time  $\mu_i$  must include the first step (the 1) plus the expected future time, weighted by the transition probabilities  $p_{ij}$ .



# Mean first passage and recurrence times

- **Chain with one recurrent class** (non-absorbing states).
- **Mean first passage time  $t_i$  from  $i$  to  $s$  ( $i \neq s$ ):**  
Expected number of transitions until reaching state  $s$ , starting from  $i$ .
- **Governing Equations:**
  - $t_s = 0$  (Base case).
  - $t_i = 1 + \sum_j p_{ij} t_j$ , for all  $i \neq s$ .
- **Mean recurrence time  $t_s^*$  of  $s$ :** Expected number of transitions until returning to state  $s$ , starting from  $s$ .
- **Governing Equation:**

$$t_s^* = 1 + \sum_j p_{sj} t_j$$
- **Relation to Steady State:** For an irreducible, aperiodic chain, the steady-state probability is  $\pi_s = 1/t_s^*$ .





## Gambler's example

- **Process:** Gambler starts with  $i$  dollars, bets \$1 in a fair game ( $p = 0.5$  win,  $1 - p = 0.5$  lose) until wealth is 0 (lose) or  $n$  (win).
- **States:**  $i \in \{0, 1, \dots, n\}$ . 0 and  $n$  are absorbing states.
- **Absorption Probability**  $a_i$  ( $\rightarrow n$ ):
  - $i = 0 \implies a_0 = 0$ .
  - $i = n \implies a_n = 1$ .
  - $0 < i < n \implies a_i = 0.5a_{i-1} + 0.5a_{i+1}$ .
  - Solution (fair game):  $a_i = i/n$ .
- **Expected Wealth at the end:**  $0 \cdot (1 - a_i) + n \cdot a_i = n \cdot (i/n) = i$ . (Expected initial wealth equals expected final wealth).
- **Expected time**  $\mu_i$  **in the game** ( $\rightarrow 0$  or  $n$ ):
  - $\mu_0 = 0, \mu_n = 0$ .
  - $\mu_i = 1 + 0.5\mu_{i-1} + 0.5\mu_{i+1}$ .
  - Solution (fair game):  $\mu_i = i(n - i)$ .

## **Lecture 30: Stationarity and Ergodicity**

---

# Review: Random Processes

- A **Random Process**  $X(t)$  is a collection of random variables indexed by time  $t$ .
- For a fixed time  $t$ ,  $X(t)$  is a single random variable.
- A single realization of  $X(t)$  is a **sample function** or a time series.
- The process is fully characterized by its **joint distributions** for all time instants  $t_1, t_2, \dots, t_n$ :

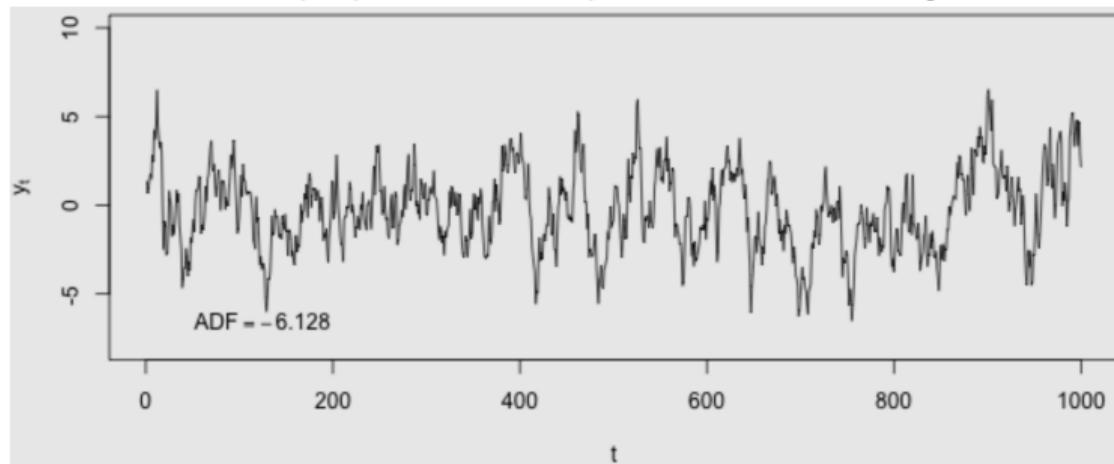
$$F_{X(t_1), \dots, X(t_n)}(x_1, \dots, x_n) = P(X(t_1) \leq x_1, \dots, X(t_n) \leq x_n)$$

# Strict-Sense Stationarity (SSS)

- A random process  $X(t)$  is **Strict-Sense Stationary (SSS)** if its finite-dimensional joint distributions are **invariant to a shift in time**.
- For any  $n$ , any time instants  $t_1, \dots, t_n$ , and any time shift  $\tau$ :

$$F_{X(t_1), \dots, X(t_n)}(x_1, \dots, x_n) = F_{X(t_1 + \tau), \dots, X(t_n + \tau)}(x_1, \dots, x_n)$$

- **Interpretation:** The statistical properties of the process do not change over time.



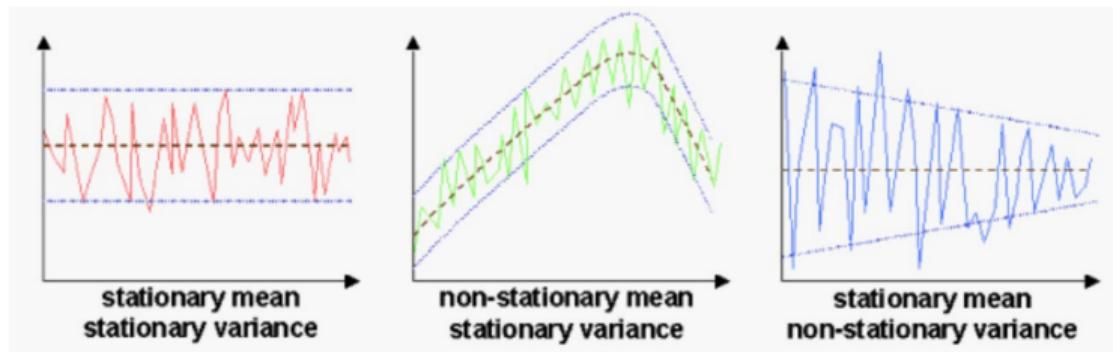
# Implications of SSS

- **First-order distribution:**  $F_{X(t)}(x) = F_{X(t+\tau)}(x)$ .
  - The mean  $E[X(t)]$  must be a constant, independent of  $t$ .

$$E[X(t)] = \mu_x \quad (\text{constant})$$

- **Second-order distribution:**  $F_{X(t_1), X(t_2)}(x_1, x_2) = F_{X(t_1+\tau), X(t_2+\tau)}(x_1, x_2)$ .
  - The **autocorrelation function**  $R_X(t_1, t_2) = E[X(t_1)X(t_2)]$  must only depend on the time difference  $\tau = t_2 - t_1$ .

$$R_X(t_1, t_2) = R_X(t_2 - t_1) = R_X(\tau)$$

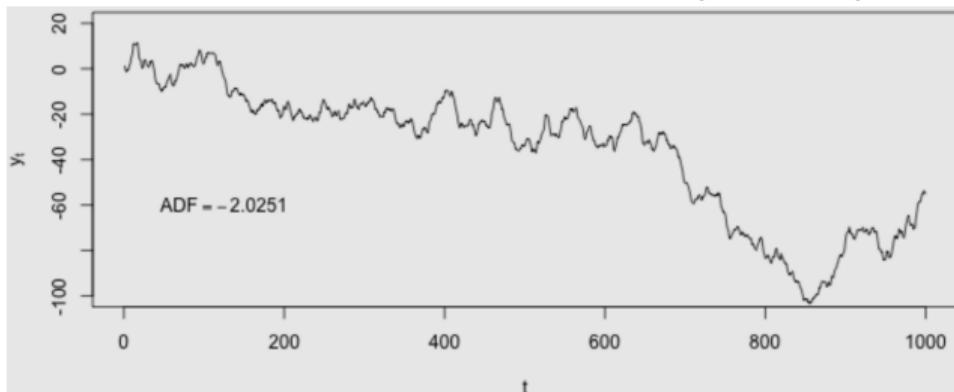


## Example: Process that is NOT SSS

- **Process:**  $X(t) = A \sin(\omega_0 t)$ , where  $A$  is a random variable  $A \sim U[0, 1]$ .
- **Goal:** Show that  $X(t)$  is not SSS by examining the mean  $E[X(t)]$ .
- **Mean Calculation:** Since  $A$  and  $\sin(\omega_0 t)$  are independent:

$$E[X(t)] = E[A]E[\sin(\omega_0 t)], \quad E[A] = \int_0^1 a \cdot 1 da = 1/2, \quad E[X(t)] = (1/2) \sin(\omega_0 t)$$

- **Conclusion:** The mean  $E[X(t)]$  depends on time  $t$ , specifically it oscillates.
- Since the mean is not constant, the process cannot be SSS (nor WSS).



# Wide-Sense Stationarity (WSS)

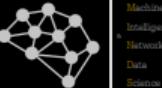
- A random process  $X(t)$  is **Wide-Sense Stationary (WSS)** if it satisfies two conditions on its first two moments:
  1. The **mean** is constant and independent of time:

$$E[X(t)] = \mu_x \quad (\text{constant})$$

- 2. The **autocorrelation function** depends only on the time difference  $\tau = t_2 - t_1$ :

$$R_X(t_1, t_2) = R_X(t_2 - t_1) = R_X(\tau)$$

- WSS is a less stringent condition than SSS.
- **Relationship:** SSS  $\implies$  WSS (if moments exist). WSS  $\not\implies$  SSS.



## Example: WSS Check

- **Process:**  $X(t) = A \cos(\omega_0 t) + B \sin(\omega_0 t)$ , where  $A$  and  $B$  are independent  $N(0, \sigma^2)$  random variables.
- **Goal:** Verify the two WSS conditions.
- **Condition 1: Constant Mean**

$$E[X(t)] = E[A] \cos(\omega_0 t) + E[B] \sin(\omega_0 t)$$

$$E[X(t)] = 0 \cdot \cos(\omega_0 t) + 0 \cdot \sin(\omega_0 t) = 0 \quad (\text{constant})$$

## Example: WSS Check (Cont.)

- **Process:**  $X(t) = A \cos(\omega_0 t) + B \sin(\omega_0 t)$ .  $A, B \sim N(0, \sigma^2)$ , independent.
- **Condition 2: Autocorrelation**  $R_X(t_1, t_2)$

$$R_X(t_1, t_2) = E[X(t_1)X(t_2)]$$

$$R_X(t_1, t_2) = E[(A \cos(\omega_0 t_1) + B \sin(\omega_0 t_1)) \cdot (A \cos(\omega_0 t_2) + B \sin(\omega_0 t_2))]$$

$$R_X(t_1, t_2) = E[A^2] \cos(\omega_0 t_1) \cos(\omega_0 t_2) + E[B^2] \sin(\omega_0 t_1) \sin(\omega_0 t_2)$$

- Since  $E[A^2] = E[B^2] = \sigma^2$  (variance of zero-mean RVs).
- $R_X(t_1, t_2) = \sigma^2 [\cos(\omega_0 t_1) \cos(\omega_0 t_2) + \sin(\omega_0 t_1) \sin(\omega_0 t_2)]$
- Using  $\cos(a - b) = \cos a \cos b + \sin a \sin b$ :  $R_X(t_1, t_2) = \sigma^2 \cos(\omega_0(t_2 - t_1))$
- **Conclusion:**  $R_X(t_1, t_2) = \sigma^2 \cos(\omega_0 \tau)$ , which depends only on  $\tau = t_2 - t_1$ . The process is WSS.

# WSS and Gaussian Processes

- **Gaussian Process:** A random process where all finite-dimensional joint distributions are multivariate Gaussian.
- The distribution of a Gaussian process is completely defined by its mean and autocorrelation functions.
- **Crucial Exception:** For a Gaussian process, the WSS condition (on the mean and autocorrelation) is sufficient to imply SSS.

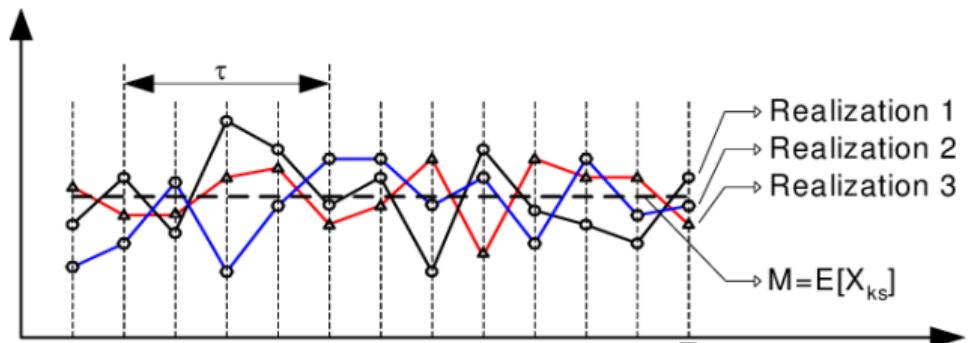
$$\text{WSS} \xrightleftharpoons{\text{Gaussian}} \text{SSS}$$

# Ergodicity: The Time vs. Ensemble Average

- Stationarity deals with the **ensemble** (average across realizations at fixed time).
- **Ergodicity** allows us to replace the ensemble average with the **time average** from a single, long realization.
- A process  $X(t)$  is **Ergodic in the Mean** if the time average converges to the ensemble mean:

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T X(t) dt = E[X(t)] \quad (\text{in probability or a.s.})$$

- Ergodicity in the Mean requires the process to be WSS.



## Example: Process that is NOT Ergodic

- **Process:**  $X(t) = A$ , where  $A$  is a random variable  $A \sim U[0, 1]$ .
- **Goal:** Show that this WSS process is NOT Ergodic in the Mean.
- **WSS Check:**
  - Mean:  $E[X(t)] = E[A] = 1/2$ . (Constant).
  - Autocorrelation:  $R_X(t_1, t_2) = E[X(t_1)X(t_2)] = E[A^2] = \text{var}(A) + (E[A])^2 = 1/12 + 1/4 = 1/3$ .  
(Depends only on  $\tau = 0$ , so it's a constant  $R_X(\tau) = 1/3$ ).
  - The process is WSS.
- **Time Average Calculation:**

$$\text{Time Average} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T Adt = \lim_{T \rightarrow \infty} \frac{1}{T} [At]_0^T = A$$

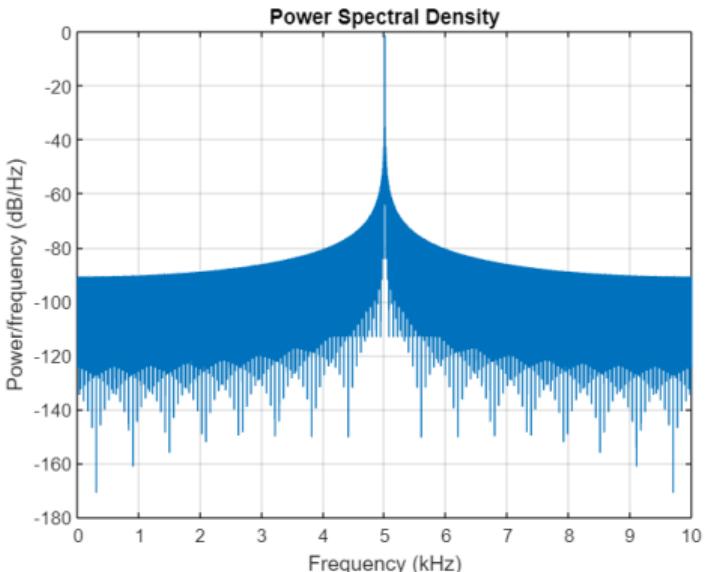
- **Conclusion:** The time average  $A$  is a random variable, while the ensemble mean is a constant  $1/2$ . They are not equal, so the process is NOT Ergodic.

# Application: Signal Processing (WSS)

- **Assumption:** Most fundamental results in linear filtering, spectral analysis, and optimal estimation assume input signals are **WSS**.
- **Example: Power Spectral Density (PSD):** For a WSS process, the PSD  $S_X(\omega)$  is the Fourier Transform of the autocorrelation function  $R_X(\tau)$ :

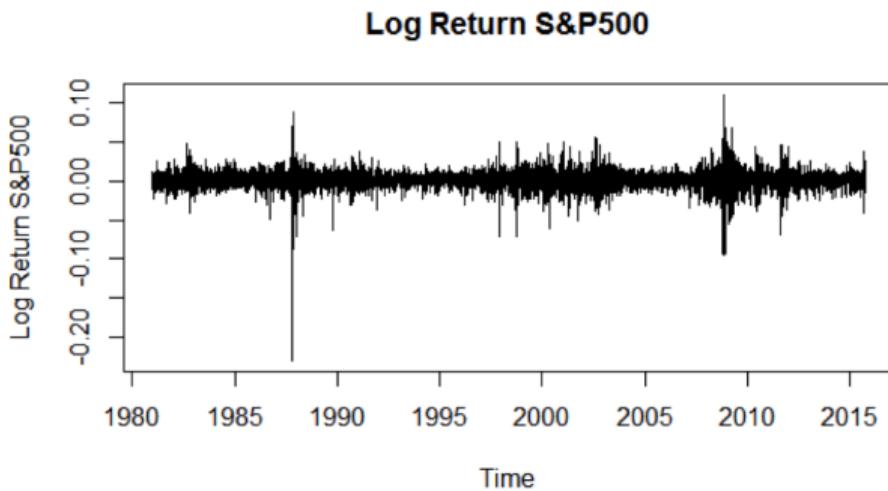
$$S_X(\omega) = \mathcal{F}\{R_X(\tau)\} = \int_{-\infty}^{\infty} R_X(\tau) e^{-j\omega\tau} d\tau$$

- **Wiener-Khinchin Theorem:**  
This relation holds exclusively for WSS processes.



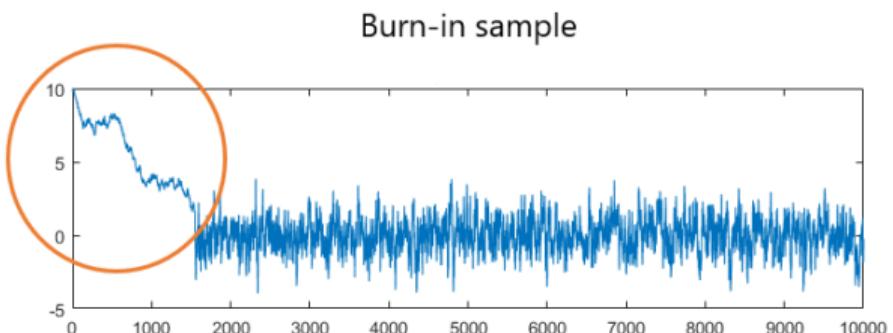
# Application: Financial Engineering (Non-Stationarity)

- **Stock Returns:** Logarithmic returns  $R(t)$  are often assumed to be WSS (or at least approximately WSS) to simplify modeling (e.g., in the Black-Scholes model).
- **Real-World Data:** Financial time series are highly **non-stationary**.
- **Modeling:** Non-stationary models like GARCH (Generalized Autoregressive Conditional Heteroskedasticity) are necessary to capture time-varying volatility.



# Application: Machine Learning (Ergodicity)

- **Monte Carlo Methods / MCMC:** Markov Chain Monte Carlo (MCMC) algorithms use time-evolution to sample a complex distribution.
- The underlying Markov Chain must be **ergodic** (irreducible and aperiodic) to ensure:
  1. The chain converges to the target stationary distribution  $\pi_j$ .
  2. Time averages computed from a single long chain realization approximate the expected values under  $\pi_j$ .
- Ergodicity is necessary for the outputs of MCMC simulations to be reliable estimates of expectations.



# Summary and Conditions for Ergodicity

- **SSS:** Joint distributions independent of time shift  $\tau$ .
- **WSS:** Mean is constant, Autocorrelation  $R_X(t_1, t_2)$  depends only on  $\tau = t_2 - t_1$ .
- $\text{SSS} \implies \text{WSS}$ .  $\text{WSS} \xrightleftharpoons{\text{Gaussian}} \text{SSS}$ .
- **Ergodicity in the Mean:** Time Average  $\rightarrow$  Ensemble Mean.
- **Sufficient Condition:** A WSS process is Ergodic in the Mean if its autocorrelation satisfies:

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T |R_X(\tau)| d\tau = 0$$