Lecture 1: Probability Models and Axioms

Instructor: Prof. Abolfazl Hashemi

1 Introduction to Probabilistic Modeling

The theory of probability is a mathematical framework for reasoning about and analyzing phenomena with uncertain outcomes. The process of applying this theory involves two main stages:

1. **Building a Probabilistic Model:** We first specify all the possible outcomes of an experiment and then assign a probability to each of these outcomes.
2. **Analysis and Inference:** We use the model to make predictions, calculate the probabilities of various events, and make decisions.

This lecture focuses on the first step: the fundamental structure of a probabilistic model. We will cover the sample space, the axioms of probability, and the properties that arise from these axioms.

2 The Sample Space

The first step in creating a probabilistic model is to define the **sample space**, denoted by the Greek letter Ω (Omega).

Definition (Sample Space): The sample space is the set of all possible outcomes of an experiment.

When defining a sample space, the set of outcomes must be:

- **Mutually Exclusive:** Each outcome must be distinct from all others. If one outcome occurs, no other outcome can occur at the same time.
- **Collectively Exhaustive:** The set of outcomes must include every possible result of the experiment. No possible outcome can be left out.

The **granularity** of the sample space is also important. The level of detail should be appropriate for the problem at hand. For example, when flipping a coin, the sample space could be simply {Heads, Tails}, or it could include more detail, such as the final orientation of the coin in degrees. The choice depends on what we want to analyze.

2.1 Example: Discrete Sample Space

Consider an experiment involving two successive rolls of a fair four-sided (tetrahedral) die. The faces of the die are numbered {1, 2, 3, 4}. Let the outcome of the first roll be X and the outcome of the second roll be Y. The sample space Ω is the set of all possible pairs (X, Y).

$$\Omega = \{(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4), (4,1), (4,2), (4,3), (4,4)\}$$

This sample space contains $4 \times 4 = 16$ possible outcomes. We can visualize this space as a grid or as a tree, representing the sequential nature of the experiment.

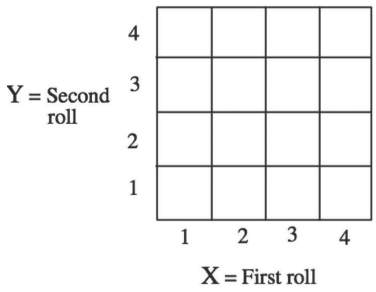


Figure 1: Grid representation of the sample space for two die rolls.

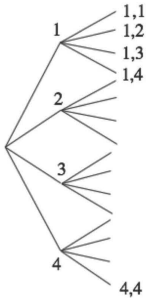


Figure 2: Tree-based sequential description of the sample space.

2.2 Example: Continuous Sample Space

Not all experiments have a finite or countably infinite number of outcomes. Consider an experiment where we throw a dart at a square board of side length 1. An outcome is the coordinate pair (x, y) where the dart lands. The sample space is the set of all possible coordinates.

$\Omega = \{(x, y) \mid 0 \leq x \leq 1, 0 \leq y \leq 1\}$

This is a continuous sample space, represented by the unit square in the Cartesian plane.

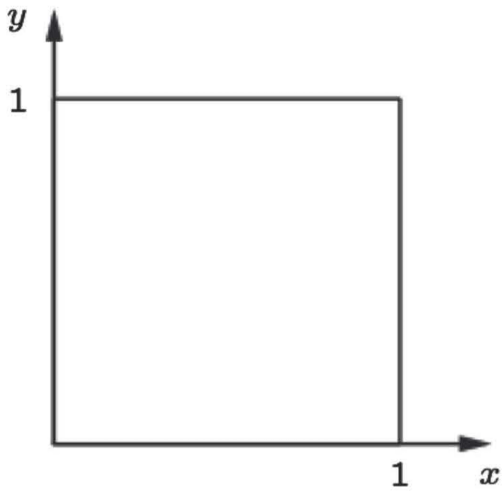


Figure 3: Continuous sample space represented by the unit square.

3 Probability Laws and Axioms

Once the sample space Ω is defined, the next step is to specify a **probability law**. This law assigns a probability $P(A)$ to every event A.

Definition (Event): An event is a subset of the sample space Ω.

The probability law must satisfy three fundamental axioms. These axioms form the bedrock of probability theory and are not derived from other principles.

3.1 The Axioms of Probability

For any events A and B:

1. **Nonnegativity:** The probability of any event is non-negative.
$$P(A) \geq 0$$
2. **Normalization:** The probability of the entire sample space is 1. This means that one of the possible outcomes must occur.
$$P(\Omega) = 1$$
3. **Additivity:** If two events A and B are disjoint (mutually exclusive), meaning they have no outcomes in common ($A \cap B = \emptyset$), then the probability of their union is the sum of their probabilities.
$$\text{If } A \cap B = \emptyset, \text{ then } P(A \cup B) = P(A) + P(B)$$

3.2 Some Consequences of the Axioms

Several fundamental properties of probability can be derived directly from these three axioms.

- **Probability of the Empty Set:** The probability of the empty set $\emptyset$ (an impossible event) is 0.

  Proof: Let $A = \Omega$ and $B = \emptyset$. Since $\Omega \cap \emptyset = \emptyset$, these events are disjoint. By the additivity axiom, $P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset)$. Since $\Omega \cup \emptyset = \Omega$, we have $P(\Omega) = P(\Omega) + P(\emptyset)$. By the normalization axiom, $1 = 1 + P(\emptyset)$, which implies $P(\emptyset) = 0$.

- **Probability of the Complement:** The probability of the complement of an event A, denoted A^c, is given by $P(A^c) = 1 - P(A)$.

  Proof: The events A and A^c are disjoint and their union is Ω. By the additivity axiom, $P(A \cup A^c) = P(A) + P(A^c)$. Since $A \cup A^c = \Omega$, we have $P(\Omega) = P(A) + P(A^c)$. By the normalization axiom, $1 = P(A) + P(A^c)$, which gives the result.

- **Upper Bound on Probability:** The probability of any event A is at most 1.

  Proof: Since $P(A^c) \geq 0$ (by nonnegativity), we have $1 - P(A) \geq 0$, which implies $P(A) \leq 1$.

- **Subset Property:** If event A is a subset of event B ($A \subset B$), then $P(A) \leq P(B)$.

  Proof: We can write B as the union of two disjoint sets: $B = A \cup (B \cap A^c)$. By additivity, $P(B) = P(A) + P(B \cap A^c)$. By nonnegativity, $P(B \cap A^c) \geq 0$, so $P(B) \geq P(A)$.

- **Probability of a Union (Inclusion-Exclusion Principle):** For any two events A and B, the probability of their union is:
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

  Proof: We can write $A \cup B$ as the union of two disjoint sets: $A \cup B = A \cup (B \cap A^c)$. By additivity, $P(A \cup B) = P(A) + P(B \cap A^c)$. We can also write $B = (A \cap B) \cup (B \cap A^c)$, which are disjoint. So, $P(B) = P(A \cap B) + P(B \cap A^c)$, or $P(B \cap A^c) = P(B) - P(A \cap B)$. Substituting this back gives the desired result.

  A direct consequence is the **Union Bound**: $P(A \cup B) \leq P(A) + P(B)$.

4 Calculating Probabilities

4.1 Discrete Uniform Law

A common probability law for finite sample spaces is the **discrete uniform law**. It assumes that every outcome in Ω is equally likely. If Ω has n elements, then the probability of any single outcome is $1/n$. For an event A containing k outcomes:

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

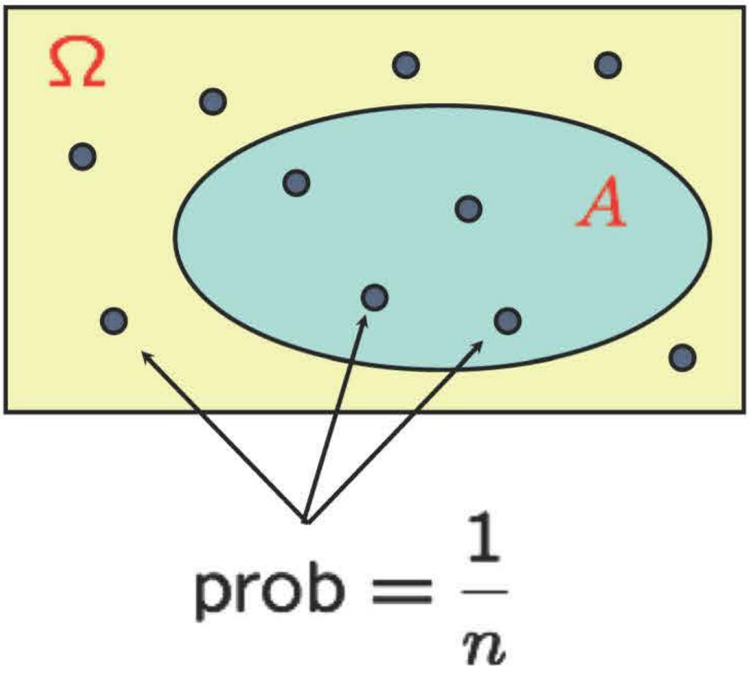


Figure 4: Illustration of the discrete uniform law.

Example: Two tetrahedral die rolls Let's use the die roll example. The sample space has 16 outcomes. Assuming the uniform law, each outcome has probability 1/16.

- What is the probability that the first roll is 1, $P(X = 1)$? The event $A = \{X = 1\}$ is the set $\{(1,1),(1,2),(1,3),(1,4)\}$. It has 4 outcomes. $P(X = 1) = 4/16 = 1/4$.
- Let $Z = \min(X, Y)$. What is the probability that $Z = 2$, $P(Z = 2)$? The event $B = \{Z = 2\}$ occurs if the outcomes are $(2,2),(2,3),(2,4),(3,2),(4,2)$. It has 5 outcomes. $P(Z = 2) = 5/16$.
- What is the probability that $Z = 4$, $P(Z = 4)$? The event $C = \{Z = 4\}$ occurs only if the outcome is $(4,4)$. It has 1 outcome. $P(Z = 4) = 1/16$.

4.2 Continuous Uniform Law

For our continuous example of a dart thrown at a unit square, we can define a uniform probability law where the probability of an event is equal to its area.

$$P(A) = \text{Area of } A$$

- What is the probability that $x + y \leq 1/2$? This event corresponds to a triangular region in the lower-left corner of the square, with vertices at $(0, 0)$, $(1/2, 0)$, and $(0, 1/2)$. The area of this triangle is $\frac{1}{2} \times$ base $\times$ height $= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = 1/8$. So, $P(x + y \leq 1/2) = 1/8$.
- What is the probability of hitting the exact point $(0.5, 0.3)$? A single point has no area. Therefore, the probability is 0. In a continuous sample space, events corresponding to single outcomes have zero probability.

5 The Countable Additivity Axiom

The additivity axiom as stated is sufficient for finite sample spaces. However, for infinite sample spaces, we need a stronger version.

The Countable Additivity Axiom: If $A_1, A_2, \ldots$ is an infinite sequence of disjoint events, then the probability of their union is the sum of their probabilities:

$$P(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i)$$

This axiom is crucial for handling infinite sample spaces.

Example: Infinite Discrete Sample Space Let the sample space be the set of positive integers, $\Omega = \{1, 2, 3, \ldots\}$. Let the probability law be $P(n) = \frac{1}{2^n}$ for each $n \in \Omega$. (First, we check that this is a valid probability law: $\sum_{n=1}^{\infty} P(n) = \sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1/2}{1-1/2} = 1$).

What is the probability that the outcome is an even number? The event is $A = \{2, 4, 6, 8, \ldots\}$. This is an infinite union of disjoint events: $A = \{2\} \cup \{4\} \cup \{6\} \cup \ldots$. Using the countable additivity axiom:

$$P(A) = P(2) + P(4) + P(6) + \cdots = \frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^6} + \ldots$$

This is a geometric series with first term $a = 1/4$ and common ratio $r = 1/4$. The sum is:

$$P(A) = \frac{a}{1-r} = \frac{1/4}{1-1/4} = \frac{1/4}{3/4} = \frac{1}{3}$$

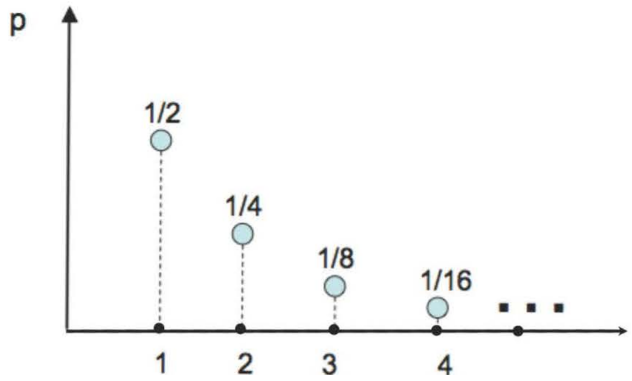


Figure 5: A discrete infinite probability distribution.

6 Conclusion: The Role of Probability Theory

Probability theory provides the tools to move from a real-world problem involving uncertainty to a mathematical model. This model allows for rigorous analysis, leading to predictions and informed decisions. The data from the real world, in turn, helps in building and refining these models through inference and statistics. The axioms are the starting point for this entire framework, providing the rules for consistent reasoning in the face of uncertainty.

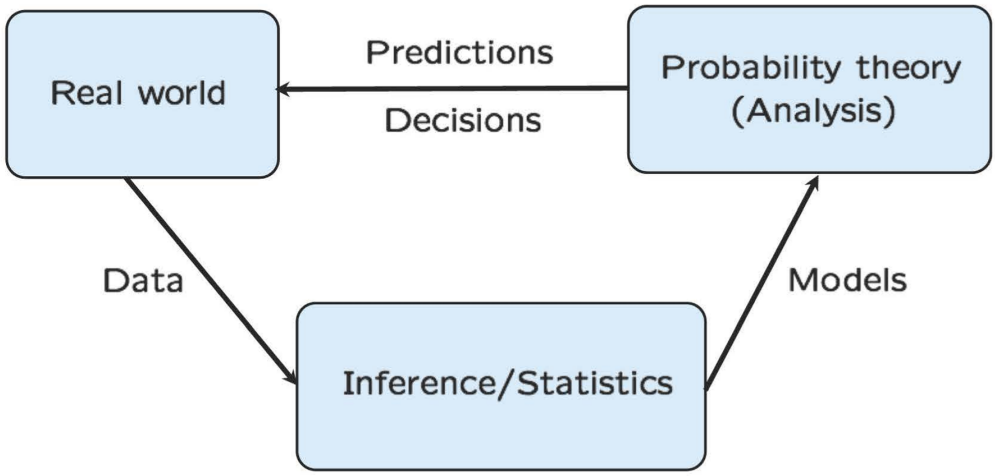


Figure 6: The cycle of modeling, analysis, and inference.

Lecture 2: Conditioning and Bayes' Rule

Instructor: Prof. Abolfazl Hashemi

1 The Concept of Conditional Probability

In probability theory, our understanding of events evolves as we receive new information. Conditional probability is the mathematical framework for updating our beliefs in light of new evidence. It allows us to answer the question: how does the probability of an event A change after we learn that another event B has occurred?

1.1 An Intuitive Example

Let's consider a sample space Ω with 12 equally likely outcomes. Let A and B be two events within this space.

In this original model, suppose event A contains 5 outcomes and event B contains 6 outcomes. The probability of each event is calculated under the discrete uniform law:

$$P(A) = \frac{\text{Number of outcomes in A}}{\text{Total outcomes}} = \frac{5}{12}$$

$$P(B) = \frac{\text{Number of outcomes in B}}{\text{Total outcomes}} = \frac{6}{12} = \frac{1}{2}$$

Now, suppose we are told that event B has definitively occurred. This information restricts the set of possible outcomes to only those within B. Our original sample space Ω is no longer relevant; the new, effective sample space is now B.

Within this new universe of 6 equally likely outcomes, we want to find the probability that A also occurred. We are interested in the outcomes that are in A *and* in B. This corresponds to the intersection $A \cap B$, which has 2 outcomes. The probability of A given this new information, which we write as $P(A|B)$, is:

$$P(A|B) = \frac{\text{Number of outcomes in } A \cap B}{\text{Number of outcomes in B}} = \frac{2}{6} = \frac{1}{3}$$

This is the "conditional probability of A given B." Our belief in A occurring has been revised from 5/12 to 1/3 based on the knowledge that B happened.

2 Formal Definition of Conditional Probability

The intuitive idea of shrinking the sample space leads to the formal definition of conditional probability.

Definition (Conditional Probability): The conditional probability of an event A given an event B with $P(B) > 0$ is defined as:

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

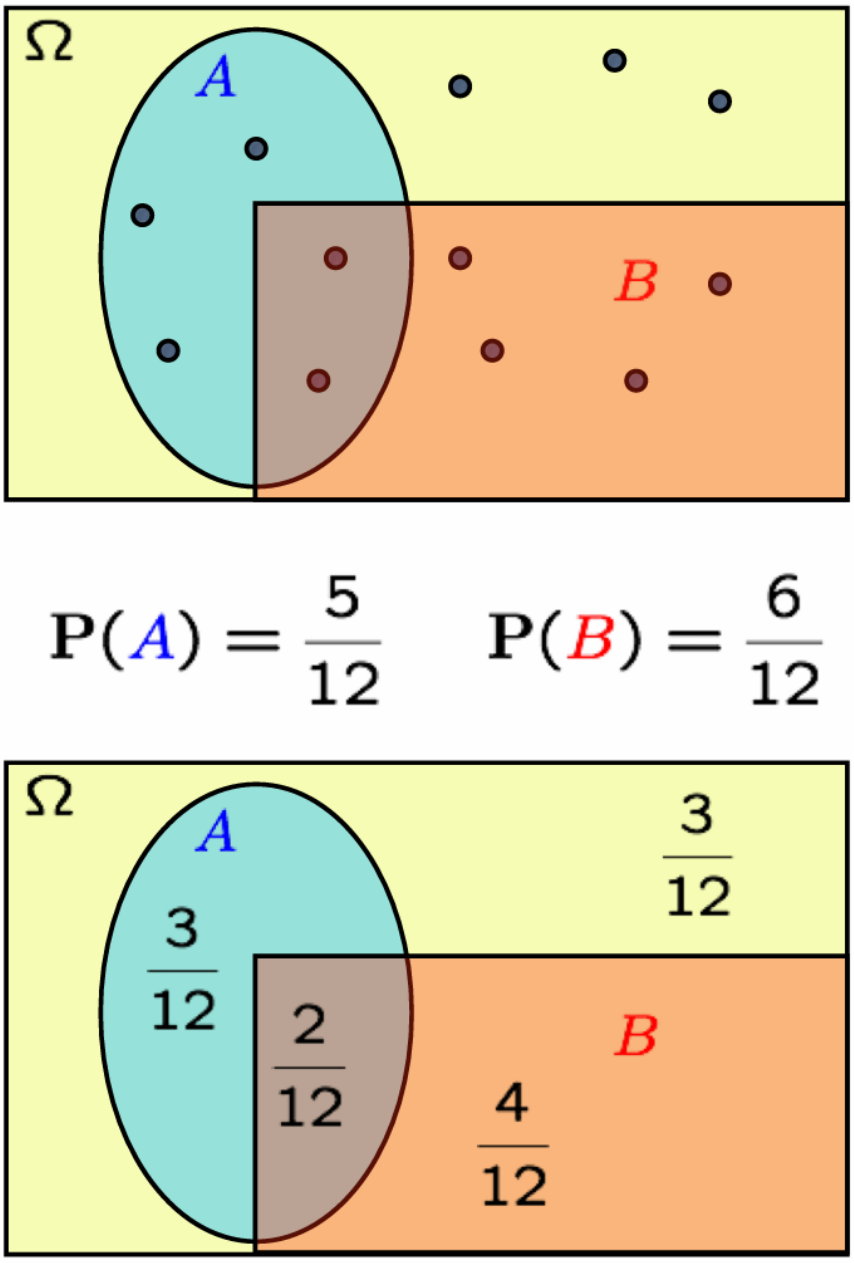


Figure 7: Original sample space with 12 outcomes.

This formula works for any probability law, not just the uniform case. It rescales the probability of the intersection, $P(A \cap B)$, by the probability of the new universe, $P(B)$. Applying this to our previous example:

$$P(A \cap B) = \frac{2}{12}, \quad P(B) = \frac{6}{12}$$

$$P(A|B) = \frac{2/12}{6/12} = \frac{2}{6} = \frac{1}{3}$$

The formal definition perfectly matches our intuition.

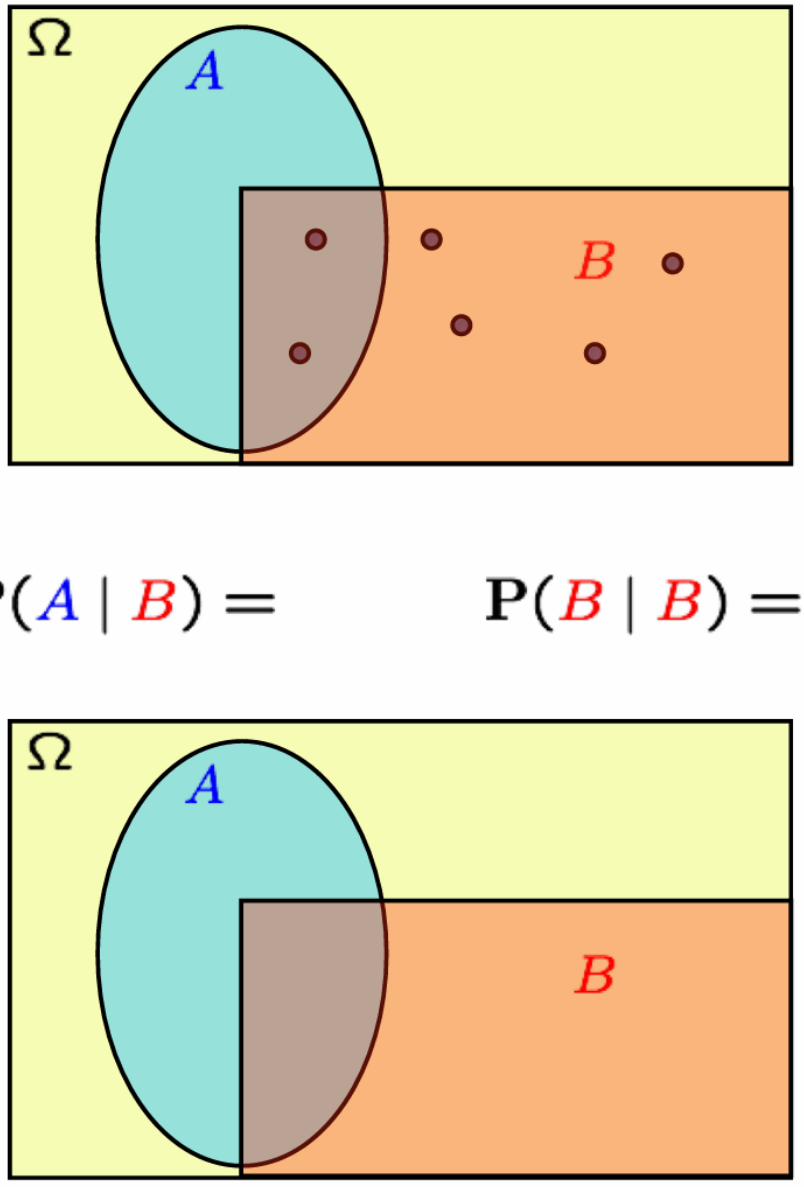


Figure 8: The revised sample space, which is now the event B.

2.1 Example: Two Rolls of a 4-Sided Die

Let's consider an experiment of rolling a 4-sided die twice, with a sample space of 16 equally likely outcomes. Let X be the first roll and Y be the second.

- Let B be the event that $\min(X, Y) = 2$. The outcomes in B are $\{(2,2), (2,3), (2,4), (3,2), (4,2)\}$. Therefore, $P(B) = 5/16$.
- Let M be the event representing the value of $\max(X, Y)$.

Let's calculate some conditional probabilities:

1. **What is $P(M = 1|B)$?** The event $M = 1$ corresponds to the outcome $\{(1,1)\}$. The

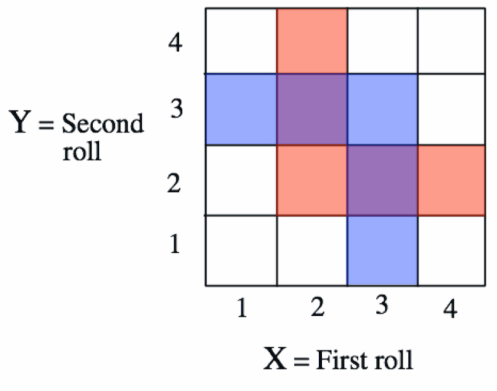


Figure 9: The event B where min(X,Y) = 2 is highlighted.

intersection of this event with B is empty. Therefore, $P(M = 1 \cap B) = 0$.

$$P(M = 1|B) = \frac{P(M = 1 \cap B)}{P(B)} = \frac{0}{5/16} = 0$$

This makes sense: if the minimum roll was 2, it is impossible for the maximum roll to have been 1.

2. **What is $P(M = 3|B)$?** The event $M = 3$ is $\{(1,3),(2,3),(3,3),(3,2),(3,1)\}$. The intersection of this event with B is the set $\{(2,3),(3,2)\}$. The probability of this intersection is $P(M = 3 \cap B) = 2/16$.

$$P(M = 3|B) = \frac{P(M = 3 \cap B)}{P(B)} = \frac{2/16}{5/16} = \frac{2}{5}$$

3 Properties of Conditional Probability and Key Tools

A conditional probability law, $P(\cdot|B)$, is a valid probability law that satisfies all the axioms of probability over the new sample space B. This means it is non-negative, normalized ($P(B|B) = 1$), and countably additive.

From the definition of conditional probability, we can derive three cornerstone tools for probabilistic analysis.

3.1 The Multiplication Rule

By rearranging the definition, we get a powerful tool for calculating the probability of an intersection.

Multiplication Rule: For any events A and B:

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

This rule is particularly useful in sequential experiments, where outcomes happen in stages. It can be extended to a sequence of multiple events:

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

This is visualized effectively with a probability tree.

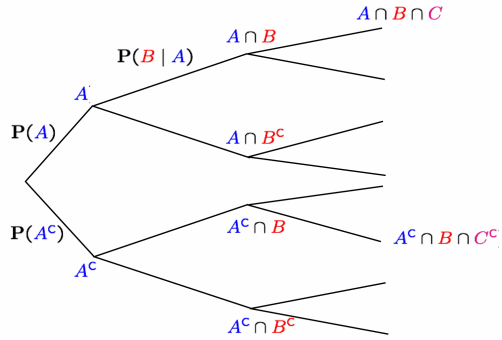


Figure 10: A probability tree representing a sequential process.

3.2 The Total Probability Theorem

This theorem provides a way to calculate the probability of an event by considering a partition of the sample space. A set of events $\{A_1, A_2, \ldots, A_n\}$ is a **partition** if the events are mutually exclusive and their union is Ω.

Total Probability Theorem: For a partition $\{A_1, \ldots, A_n\}$ of Ω:

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

Derivation: The event B can be broken into disjoint pieces: $B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$. By additivity, $P(B) = \sum P(B \cap A_i)$. Applying the multiplication rule to each term gives the theorem.

3.3 Bayes' Rule

Bayes' rule is the most important tool for inference. It allows us to "flip" a conditional probability, relating $P(A|B)$ to $P(B|A)$. It tells us how to update our beliefs about a hypothesis (A_i) given some new evidence (B).

Bayes' Rule: For a partition $\{A_1, \ldots, A_n\}$:

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

Here, $P(A_i)$ is the **prior probability** (our initial belief), $P(B|A_i)$ is the **likelihood** (from our model), and $P(A_i|B)$ is the **posterior probability** (our updated belief).

3.4 Example: Radar Detection

Let's apply these tools to a concrete problem.

- Event A: An airplane is present. $P(A) = 0.05$.
- Event B: The radar detects something.
- We are given the model's performance: $P(B|A) = 0.99$ (detection probability) and $P(B|A^c) = 0.10$ (false alarm probability).

1. **Find $P(A \cap B)$:** (The probability that a plane is present AND detected) Using the multiplication rule:
$$P(A \cap B) = P(A)P(B|A) = (0.05)(0.99) = 0.0495$$

2. **Find $P(B)$:** (The overall probability of a radar detection) The events A and A^c form a partition. Using the total probability theorem:
$$P(B) = P(A)P(B|A) + P(A^c)P(B|A^c)$$
$$P(B) = (0.05)(0.99) + (0.95)(0.10) = 0.0495 + 0.095 = 0.1445$$

3. **Find $P(A|B)$:** (The probability that a plane is actually present, given a radar detection) Using Bayes' rule:
$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{0.0495}{0.1445} \approx 0.3426$$
Even though the radar screen shows a detection, the probability of a plane being there is only about 34%. This is because the high rate of false alarms from the much more likely event of no plane being present ($P(A^c) = 0.95$) contributes significantly to the total probability of detection.

4 The Multiplication Rule

The multiplication rule is a direct and powerful consequence of the definition of conditional probability. It is the primary tool for calculating the probability of the intersection of multiple events, especially in scenarios involving sequential outcomes.

4.1 Derivation and Formula

We start with the definition of the conditional probability of event A given event B:
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$
By rearranging this equation, we can express the probability of the intersection $A \cap B$ as:
$$P(A \cap B) = P(B)P(A|B)$$
Because the intersection is symmetric ($A \cap B = B \cap A$), we can also write the rule starting from $P(B|A)$:
$$P(A \cap B) = P(A)P(B|A)$$

4.2 Generalization to Multiple Events

This rule can be extended to find the probability of the intersection of more than two events. This is often called the **chain rule** of probability. For a sequence of n events, the formula is:
$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \cdots \cap A_{n-1})$$
This formula is most intuitive when thinking about a process that unfolds in stages, such as in a probability tree. To find the probability of a specific path (a sequence of outcomes), you multiply the conditional probabilities along that path.

Example for three events: The probability of the event $A^c \cap B \cap C^c$ can be calculated as follows:

$$P(A^c \cap B \cap C^c) = P(A^c) \cdot P(B|A^c) \cdot P(C^c|A^c \cap B)$$

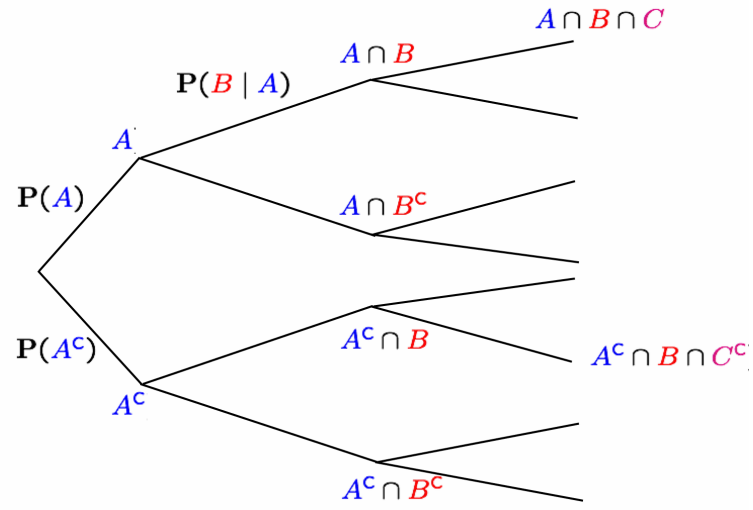


Figure 11: A probability tree illustrating the sequential calculation of intersection probabilities using the multiplication rule.

5 The Total Probability Theorem

The total probability theorem is a fundamental tool that allows us to find the probability of an event by breaking down the problem into distinct cases, or scenarios.

5.1 The Concept of a Partition

To use the theorem, we must first divide the sample space Ω into a **partition**. A set of events $\{A_1, A_2, \ldots, A_n\}$ is a partition if two conditions are met:

(a) The events are **mutually exclusive**: $A_i \cap A_j = \emptyset$ for all $i \neq j$.

(b) The events are **collectively exhaustive**: $\bigcup_{i=1}^{n} A_i = \Omega$.

In simple terms, a partition divides the sample space into non-overlapping pieces that cover all possible outcomes.

5.2 Derivation and Formula

Let $\{A_1, \ldots, A_n\}$ be a partition of Ω. Any event B can be expressed as the union of its intersections with each part of the partition:

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$$

Since the A_i events are disjoint, the pieces $(B \cap A_i)$ are also disjoint. Therefore, by the additivity axiom, we can write the probability of B as the sum of the probabilities of these pieces:

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

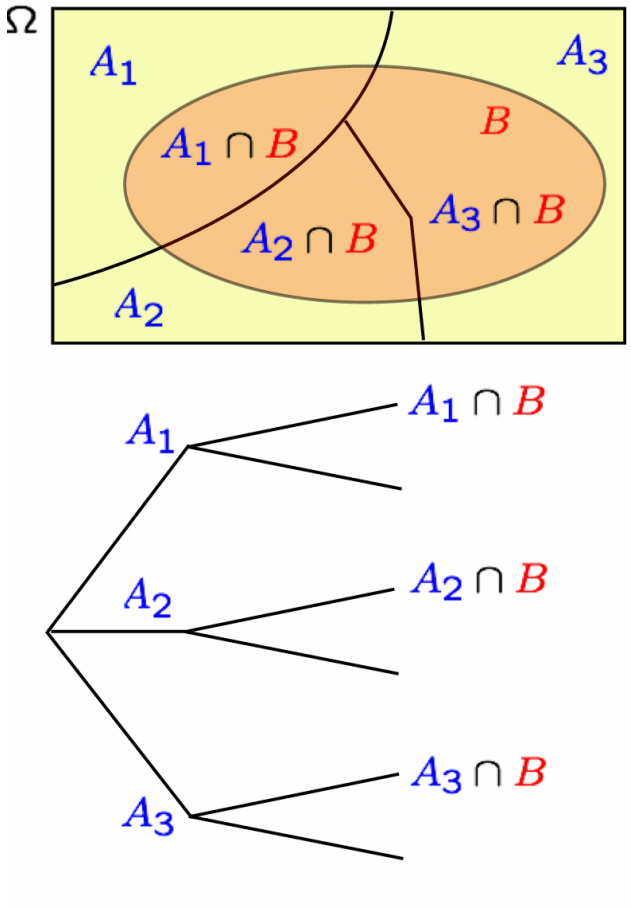


Figure 12: An event B shown with a partition of the sample space into events A_1, A_2, A_3.

Now, applying the multiplication rule to each term $P(B \cap A_i) = P(A_i)P(B|A_i)$, we arrive at the theorem.

Total Probability Theorem: For a partition $\{A_1, \ldots, A_n\}$, the probability of an event B is:

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

This formula can be interpreted as a weighted average. The total probability of event B is the weighted average of its conditional probabilities $P(B|A_i)$ across all scenarios, where the weight for each scenario is the probability of that scenario, $P(A_i)$.

6 Bayes' Rule and Statistical Inference

Bayes' rule is arguably one of the most important results in probability theory. It provides a systematic way to update our beliefs in light of new evidence. While the total probability theorem helps us compute the probability of an effect (B) given its causes (A_i), Bayes' rule allows us to infer the probability of a cause (A_i) given that we have observed an effect (B).

6.1 Derivation and Formula

We begin with the definition of conditional probability for an event A_i given an event B:

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

Using the multiplication rule, we can rewrite the numerator as $P(A_i \cap B) = P(A_i)P(B|A_i)$. Substituting this gives:

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

This is the simple form of Bayes' rule. To make it more self-contained, we can replace the denominator $P(B)$ with the expression from the total probability theorem. This gives the full form of the rule.

Bayes' Rule: For a partition $\{A_1, \ldots, A_n\}$, the conditional probability of a specific event A_i given that B has occurred is:

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

6.2 The Process of Bayesian Inference

Bayes' rule is the foundation of Bayesian inference, a major school of thought in statistics. The process involves updating our knowledge about hypotheses based on observed data.

- **Hypotheses/Causes (A_i):** These are the different states of the world or scenarios we are considering.
- **Prior Probabilities ($P(A_i)$):** This represents our initial belief about the likelihood of each hypothesis A_i *before* we observe any new evidence.
- **Evidence/Data (B):** This is the new information or data we have observed.
- **Likelihoods ($P(B|A_i)$):** This is the probability of observing the evidence B *if* hypothesis A_i were true. This is specified by our probabilistic model of the world.
- **Posterior Probabilities ($P(A_i|B)$):** This is our revised, updated belief about the likelihood of each hypothesis A_i *after* having incorporated the evidence B.

The flow of reasoning can be summarized as follows:

- **Modeling (Forward):** We model how causes lead to effects.

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

- **Inference (Backward):** We observe an effect and infer the likelihood of its potential causes.

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

This systematic approach, first described by Thomas Bayes (c. 1701-1761), is a powerful framework for reasoning under uncertainty.

Lecture 3: Independence

Instructor: Prof. Abolfazl Hashemi

1 Introduction

This lecture introduces the fundamental concept of **independence** between events. Independence is a central idea in probability that formalizes the notion of two events having no informational bearing on each other. We will explore the definition of independence for two events, extend it to collections of events, and introduce the related concept of conditional independence. Finally, we will see how these ideas are applied in practical problems like system reliability.

2 Defining Independence

2.1 Intuitive Definition

Our intuition tells us that two events, A and B, are independent if learning that one has occurred does not change the probability of the other occurring. We can express this using conditional probability:

$$P(B|A) = P(B) \quad \text{and} \quad P(A|B) = P(A)$$

While this is a very useful way to think about independence, it is not the formal definition because it is not well-defined if $P(A)$ or $P(B)$ is zero.

2.2 Formal Definition

The formal definition is more general and symmetric.

Definition (Independence): Two events A and B are **independent** if the probability of their intersection is the product of their individual probabilities:

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

If $P(A) > 0$, we can see how this relates to the intuitive definition:

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

A common point of confusion is between independent events and disjoint events. Disjoint events are not independent if their probabilities are positive. If A and B are disjoint ($A \cap B = \emptyset$), then $P(A \cap B) = 0$. For them to be independent, we would need $P(A)P(B) = 0$, which means at least one of the events must have zero probability. If both have positive probability, knowing that A occurred tells us that B certainly did not, making them highly dependent.

2.3 Independence of Complements

If events A and B are independent, then their complements are also independent. For example, A and B^c are independent.

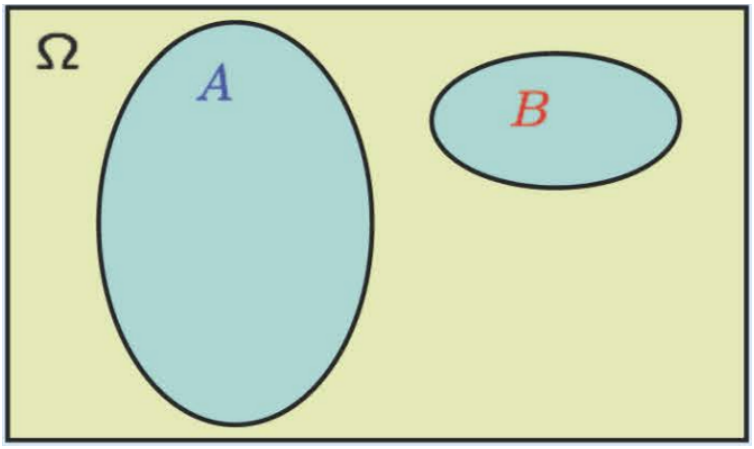


Figure 13: Disjoint events with positive probabilities are not independent.

Proof: We want to show that $P(A \cap B^c) = P(A)P(B^c)$. We can express event A as the union of two disjoint events: $A = (A \cap B) \cup (A \cap B^c)$. By the additivity axiom:

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

Rearranging the terms gives:

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

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

$$P(A \cap B^c) = P(A) - P(A)P(B) = P(A)(1 - P(B))$$

Since $1 - P(B) = P(B^c)$, we have:

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

This confirms their independence.

3 Conditional Independence

The notion of independence can be extended to a conditional probability law.

Definition (Conditional Independence): Two events A and B are **conditionally independent** given an event C if:

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

It is crucial to understand that independence does not imply conditional independence, and vice-versa. Conditioning on an event C can both create and destroy independence.

Example: Conditioning can destroy independence. Consider two independent coin tosses. Let A be the event of a head on the first toss and B be the event of a head on the second toss. Let C be the event that we get exactly one head in two tosses. Unconditionally, A and B are independent. However, if we know that C occurred, then if A is true (first toss is H), B must be false (second must be T). So $P(B|A, C) = 0$, but $P(B|C) > 0$. They are not conditionally independent.

Example: Conditioning can create independence. Consider the experiment where we choose between two biased coins, A ($P(H) = 0.9$) and B ($P(H) = 0.1$), and then toss the chosen coin twice. Let H_1 be a head on the first toss and H_2 be a head on the second. Unconditionally, these events are not independent. If we observe a head on the first toss (H_1), it becomes more likely that we chose coin A, which increases the probability of a head on the second toss. However, if we *condition* on knowing which coin was chosen (e.g., event C is "Coin A was chosen"), the tosses become independent:

$$P(H_1 \cap H_2|C) = 0.9 \times 0.9 = P(H_1|C)P(H_2|C)$$

4 Independence of a Collection of Events

The concept of independence can be generalized to more than two events.

Definition (Mutual Independence): A collection of events $A_1, A_2, \ldots, A_n$ is said to be **mutually independent** if for any sub-collection of events $\{A_{i_1}, \ldots, A_{i_k}\}$, the probability of their intersection is the product of their probabilities:

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$$

This is a very strong condition. For three events A_1, A_2, A_3, it requires that four conditions hold:

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1)P(A_2) \\ P(A_1 \cap A_3) &= P(A_1)P(A_3) \\ P(A_2 \cap A_3) &= P(A_2)P(A_3) \\ P(A_1 \cap A_2 \cap A_3) &= P(A_1)P(A_2)P(A_3) \end{aligned}$$

The first three conditions are known as **pairwise independence**. It is possible for events to be pairwise independent but not mutually independent.

For instance, 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. $P(H_1) = 1/2$
- H_2: Second toss is H. $P(H_2) = 1/2$
- C: The two tosses had the same result (HH or TT). $P(C) = 1/2$

These events are pairwise independent:

- $P(H_1 \cap H_2) = P(\{HH\}) = 1/4 = P(H_1)P(H_2)$.
- $P(H_1 \cap C) = P(\{HH\}) = 1/4 = P(H_1)P(C)$.
- $P(H_2 \cap C) = P(\{HH\}) = 1/4 = P(H_2)P(C)$.

However, they are not independent as a collection:

$$P(H_1 \cap H_2 \cap C) = P(\{HH\}) = 1/4$$

But,

$$P(H_1)P(H_2)P(C) = (1/2)(1/2)(1/2) = 1/8$$

Since $1/4 \neq 1/8$, the events are not independent.

5 Application: System Reliability

Independence is a core assumption in modeling the reliability of systems with multiple components.

Series System: A system with components in series works only if *all* components work. If the components fail independently, with p_i being the probability that component i is working:

$$P(\text{System works}) = P(1 \text{ works AND } 2 \text{ works } ...) = p_1 p_2 \cdots p_n$$

Parallel System: A system with components in parallel works if *at least one* component works. It is easier to calculate the probability that the system fails (all components fail).

$$P(\text{System fails}) = P(1 \text{ fails AND } 2 \text{ fails } ...) = (1-p_1)(1-p_2)\cdots(1-p_n)$$

The probability that the system works is therefore:

$$P(\text{System works}) = 1 - P(\text{System fails}) = 1 - \prod_{i=1}^{n}(1-p_i)$$

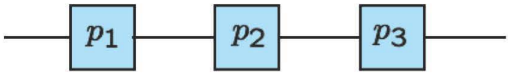


Figure 14: A parallel system.

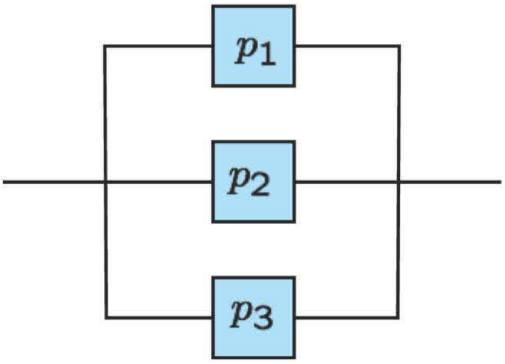


Figure 15: A series system.

6 The King's Sibling Puzzle

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

This is a classic problem where careful definition of the sample space is key. Let's assume boys (B) and girls (G) are equally likely. The initial sample space for two children, in birth order, is $\Omega = \{BB, BG, GB, GG\}$, with each outcome having probability 1/4.

The information "the king comes from a family of two children" tells us that *at least one of the children is a boy.* This reduces the sample space to:

$$\Omega' = \{BB, BG, GB\}$$

These three outcomes are now equally likely possibilities. We want to find the probability that the *other* child (the sibling) is a girl.

- In the outcome BG, the sibling is a girl.

- In the outcome GB, the sibling is a girl.
- In the outcome BB, the sibling is a boy.

Out of the 3 possible scenarios, 2 of them result in the sibling being a girl. Therefore, the probability is 2/3.

This problem highlights how conditional probability can sometimes lead to counter-intuitive results if the underlying sample space is not carefully considered.

Lecture 4: Counting

Instructor: Prof. Abolfazl Hashemi

1 Introduction to Counting in Probability

Many probabilistic models are built on a finite sample space where every outcome is equally likely. This scenario is governed by the **discrete uniform law**.

Discrete Uniform Law: If the sample space Ω consists of n equally likely outcomes, and an event A consists of k of these outcomes, then the probability of event A is:

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

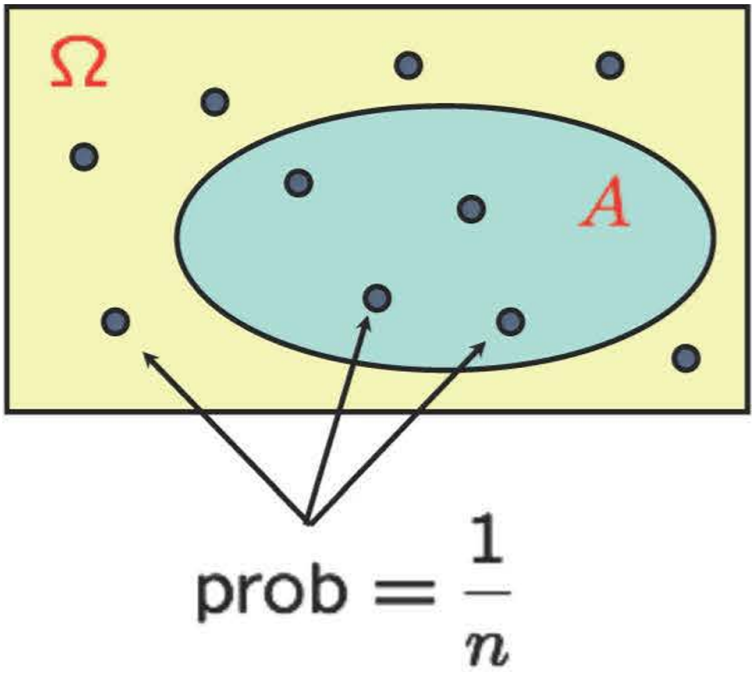


Figure 16: A sample space with n equally likely outcomes.

This simple rule transforms many probability problems into problems of counting the number of outcomes in the sample space and in the event of interest. This lecture focuses on the systematic methods of counting, known as combinatorics.

2 The Basic Counting Principle

The foundation of nearly all counting methods is the basic counting principle, which applies to processes that can be broken down into a sequence of stages.

The Basic Counting Principle: If a process consists of r sequential stages, and there are n_i possible choices at stage i (for $i = 1, \ldots, r$), then the total number of possible outcomes is the product of the number of choices at each stage:

$$\text{Total Outcomes} = n_1 \cdot n_2 \cdots n_r$$

This can be visualized as a tree, where each path from the root to a leaf represents a unique outcome.

Example: If you have 4 shirts, 3 ties, and 2 jackets, the total number of different attires you can form is a 3-stage process:

- Stage 1: Choose a shirt ($n_1 = 4$ choices)
- Stage 2: Choose a tie ($n_2 = 3$ choices)
- Stage 3: Choose a jacket ($n_3 = 2$ choices)

Total number of attires $= 4 \times 3 \times 2 = 24$.

3 Permutations and Combinations

3.1 Permutations

A permutation is an ordered arrangement of a set of distinct items.

Number of Permutations: The number of ways to order n distinct items is given by $n!$ (n-factorial):

$$n! = n \times (n-1) \times \cdots \times 2 \times 1$$

This is a direct application of the counting principle where we select items without replacement.

3.2 Combinations

A combination is an unordered selection of items from a set.

Definition: The number of ways to choose a k-element subset from an n-element set is denoted by the binomial coefficient $\binom{n}{k}$, read as "n choose k".

To find the formula for $\binom{n}{k}$, we can count the number of *ordered* sequences of k distinct items (k-permutations) in two ways:

1. **Directly:** The number of ways is $n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!}$.
2. **Indirectly:** First, choose an unordered subset of k items ($\binom{n}{k}$ ways), and then arrange these k items in order ($k!$ ways). This gives $\binom{n}{k} \times k!$.

Equating these two expressions, we get $\binom{n}{k} \times k! = \frac{n!}{(n-k)!}$, which yields the formula:

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

4 Binomial Probabilities

The binomial coefficient is essential for calculating probabilities in experiments consisting of a fixed number of independent trials, each with two possible outcomes (e.g., success/failure, heads/tails).

Consider n independent trials, each with a probability of success p.

- The probability of any *specific sequence* of k successes and $n-k$ failures is $p^k(1-p)^{n-k}$, due to the independence of the trials.
- The number of different sequences that contain exactly k successes is the number of ways to choose the k positions for the successes out of the n available trial slots, which is $\binom{n}{k}$.

Combining these, we get the binomial probability formula.

Binomial Probability Formula: The probability of obtaining exactly k successes in n independent trials is:

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

4.1 Example: A Coin Tossing Problem

Problem: Given that there were exactly 3 heads in 10 independent tosses of a fair coin, what is the probability that the first two tosses were heads?

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

Method 1: Formal Definition We use the formula $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

- $P(B)$: Using the binomial formula, $P(B) = \binom{10}{3} p^3 (1-p)^7$.
- $P(A \cap B)$: This is the event "HH followed by 1 head in the remaining 8 tosses". The probability of this is $p^2 \times [\binom{8}{1} p^1 (1-p)^7] = \binom{8}{1} p^3 (1-p)^7$.

$$P(A|B) = \frac{\binom{8}{1} p^3 (1-p)^7}{\binom{10}{3} p^3 (1-p)^7} = \frac{\binom{8}{1}}{\binom{10}{3}} = \frac{8}{120} = \frac{1}{15}$$

Method 2: Reduced Sample Space Given event B, the new sample space consists of all possible sequences with 3 heads. The number of such sequences is $\binom{10}{3}$. Since the original sequences were equally likely, these sequences in the new sample space are also equally likely. The number of favorable outcomes is the number of sequences in B that start with HH. This means the remaining 1 head must be placed in one of the last 8 positions, which can be done in $\binom{8}{1}$ ways.

$$P(A|B) = \frac{\text{Number of favorable outcomes}}{\text{Total outcomes in B}} = \frac{\binom{8}{1}}{\binom{10}{3}} = \frac{1}{15}$$

5 Partitions (Multinomial Coefficients)

We now consider partitioning a set of n distinct items into r groups of specified sizes $n_1, n_2, \ldots, n_r$, where $\sum n_i = n$.

The number of ways to do this is given by the **multinomial coefficient**:

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

5.1 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.

The total number of ways to deal the cards is the number of ways to partition 52 cards into four groups of 13:

$$|\Omega| = \binom{52}{13, 13, 13, 13} = \frac{52!}{(13!)^4}$$

To find the number of favorable outcomes, we can think of it as a two-stage process:

1. **Deal the aces:** The number of ways to give one ace to each of the four players is the number of ways to order the four aces, 4!.
2. **Deal the other 48 cards:** We need to partition the remaining 48 cards such that each player receives 12. The number of ways to do this is $\binom{48}{12,12,12,12} = \frac{48!}{(12!)^4}$.

The number of favorable outcomes is $|A| = 4! \times \frac{48!}{(12!)^4}$. The probability is:

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

6 Introduction: Beyond the Binomial

In previous discussions, we introduced the binomial distribution, which models the number of "successes" in a fixed number of independent trials, where each trial has only two possible outcomes (success or failure). However, many experiments have more than two outcomes. For example, a single roll of a die has six outcomes, and a component drawn from a bin could be one of several different types.

The **multinomial distribution** is the natural generalization of the binomial distribution to scenarios with three or more possible outcomes per trial.

6.1 The Multinomial Experiment

An experiment is considered a multinomial experiment if it satisfies the following conditions:

1. The experiment consists of a fixed number of n identical and independent trials.
2. Each trial results in exactly one of k possible outcomes, or categories.
3. The probability of outcome i, denoted p_i, is constant for every trial.
4. The probabilities of the k outcomes must sum to one: $p_1 + p_2 + \cdots + p_k = 1$.

The random variables of interest in a multinomial experiment are the counts $N_1, N_2, \ldots, N_k$, representing the number of times each of the k outcomes occurred. Note that these counts must sum to the total number of trials: $N_1 + N_2 + \cdots + N_k = n$.

6.2 Example: Building a Circuit

To make this concrete, let's consider an example. Suppose we are building a circuit and need to draw components from a large bin. The components in the bin are of three types: resistors (R), capacitors (C), and inductors (L). The proportions in the bin are known:

- 50% Resistors, so the probability of drawing a resistor is $p_R = 0.5$.
- 30% Capacitors, so the probability of drawing a capacitor is $p_C = 0.3$.
- 20% Inductors, so the probability of drawing an inductor is $p_L = 0.2$.

We draw $n = 10$ components with replacement to ensure the trials are independent and the probabilities remain constant. The question is:

What is the probability of drawing exactly 5 resistors, 3 capacitors, and 2 inductors?

6.3 Deriving the Multinomial PMF

To solve this problem, we follow a two-step process that combines the multiplication rule for independent events with the counting method for partitions.

6.3.1 Step 1: Probability of a Single Sequence

First, let's calculate the probability of one specific sequence that matches our criteria. For example, consider the sequence where we draw all 5 resistors first, then all 3 capacitors, and finally the 2 inductors:

RRRRRCCCLL

Since the trials are independent, the probability of this specific sequence is the product of the individual probabilities:

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

Any other sequence with the same composition (e.g., RCRCR...) will have the same probability, as the terms in the product are just rearranged.

6.3.2 Step 2: Count the Number of Possible Sequences

Next, we must determine how many different sequences contain exactly 5 resistors, 3 capacitors, and 2 inductors. This is a problem of partitioning a set of $n = 10$ positions into three groups of sizes $n_R = 5$, $n_C = 3$, and $n_L = 2$. The number of ways to do this is given by the **multinomial coefficient**:

$$\binom{n}{n_R, n_C, n_L} = \frac{n!}{n_R! n_C! n_L!}$$

For our example, this is:

$$\binom{10}{5, 3, 2} = \frac{10!}{5!3!2!} = \frac{3,628,800}{(120)(6)(2)} = \frac{3,628,800}{1440} = 2520$$

There are 2,520 distinct ways to arrange our 10 components.

6.3.3 The Multinomial PMF

The total probability of our event is the number of possible sequences multiplied by the probability of any single sequence.

The Multinomial PMF: Let $N_1, \ldots, N_k$ be the number of times each of k outcomes occurs in n independent trials, with respective probabilities $p_1, \ldots, p_k$. The probability of observing the specific counts $n_1, \ldots, n_k$ is:

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

where $\sum n_i = n$ and $\sum p_i = 1$.

6.4 Solving the Circuit Example

We can now solve our problem by plugging the values into the multinomial PMF.

$$\begin{aligned} P(N_R = 5, N_C = 3, N_L = 2) &= \frac{10!}{5!3!2!}(0.5)^5(0.3)^3(0.2)^2 \\ &= 2520 \cdot (0.03125) \cdot (0.027) \cdot (0.04) \\ &= 2520 \cdot (0.00003375) \\ &\approx 0.08505 \end{aligned}$$

So, there is approximately an 8.5% chance of drawing this specific mix of components.

Lecture 5: Discrete Random Variables Part I
Probability Mass Functions

Instructor: Prof. Abolfazl Hashemi

1 Introduction to Random Variables

Often in probability, we are interested not in the outcomes of an experiment themselves, but in some numerical attribute of the outcomes. A **random variable** is a way to formalize this by associating a numerical value with every possible outcome in the sample space.

1.1 Formal Definition

A random variable is a function that maps the sample space Ω to the set of real numbers $\mathbb{R}$.

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

We use an uppercase letter, like X, to denote the random variable as a function. We use a lowercase letter, like x, to denote a specific numerical value that the random variable can take. The set of all possible values for X is called its range.

A random variable is called **discrete** if its range is a finite or countably infinite set. This lecture focuses exclusively on discrete random variables.

2 The Probability Mass Function (PMF)

The **probability mass function (PMF)** is the probability law of a discrete random variable. It gives the probability for each value that the random variable can take.

Definition (PMF): The PMF of a discrete random variable X is the function $p_X(x)$ defined by:

$$p_X(x) = P(X = x)$$

The notation $P(X = x)$ is shorthand for the probability of the event consisting of all outcomes $\omega \in \Omega$ such that $X(\omega) = x$.

A PMF must satisfy two properties:

1. **Nonnegativity:** $p_X(x) \geq 0$ for all possible values x.
2. **Normalization:** The sum of the probabilities over all possible values must be 1: $\sum_x p_X(x) = 1$.

Example: PMF Calculation Consider two rolls of a fair 4-sided die. Let X be the result of the first roll and Y be the result of the second. Let $Z = X + Y$. The possible values for Z are $\{2, 3, 4, 5, 6, 7, 8\}$.

- $p_Z(2) = P(Z = 2) = P(\{(1, 1)\}) = 1/16$
- $p_Z(3) = P(Z = 3) = P(\{(1, 2), (2, 1)\}) = 2/16$
- $p_Z(4) = P(Z = 4) = P(\{(1, 3), (2, 2), (3, 1)\}) = 3/16$

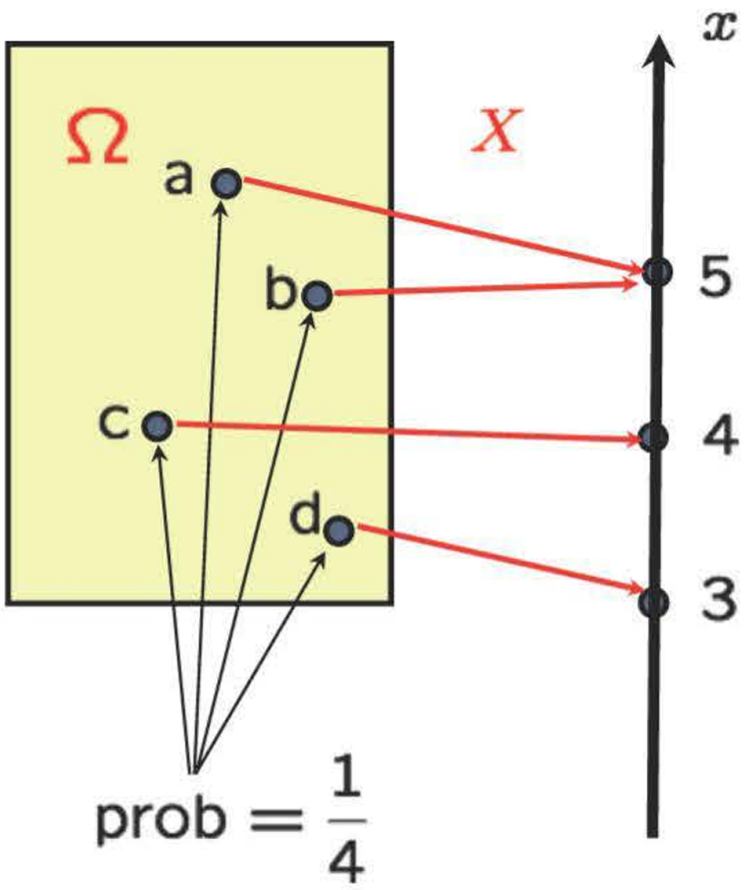


Figure 17: A random variable X mapping outcomes from Ω to numerical values.

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

The sum is $(1+2+3+4+3+2+1)/16 = 16/16 = 1$, satisfying the normalization property.

3 Common Discrete Random Variables

3.1 Bernoulli

A **Bernoulli** random variable models a single trial with two outcomes (e.g., success/failure).

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

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

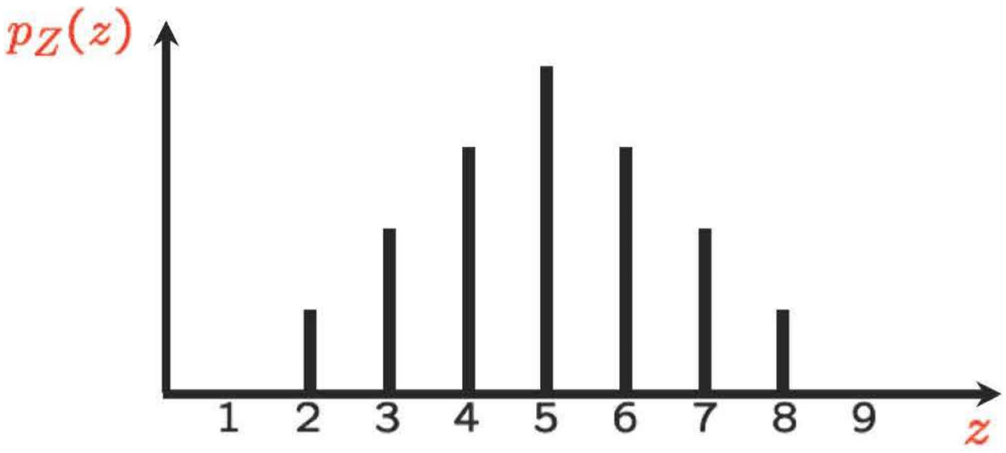


Figure 18: The PMF of Z, the sum of two 4-sided die rolls.

3.2 Discrete Uniform

A **discrete uniform** random variable models a choice from a range of integers $\{a, a+1, \ldots, b\}$, where each choice is equally likely. Its PMF is:

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

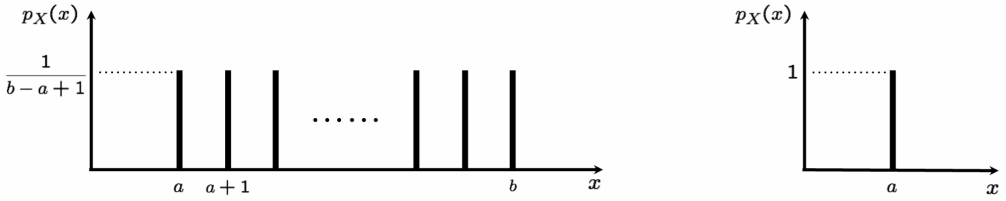


Figure 19: The PMF of Discrete Uniform

3.3 Binomial

A **binomial** random variable models the number of successes in a fixed number of independent trials. It is defined by two parameters: n (the number of independent Bernoulli trials) and p (the probability of success on each trial). The PMF for $X \sim \text{Binomial}(n, p)$ is:

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

3.4 Geometric

A **geometric** random variable models the number of independent Bernoulli trials required to achieve the first success. It is defined by the parameter p (probability of success). The PMF for $X \sim \text{Geometric}(p)$ is:

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

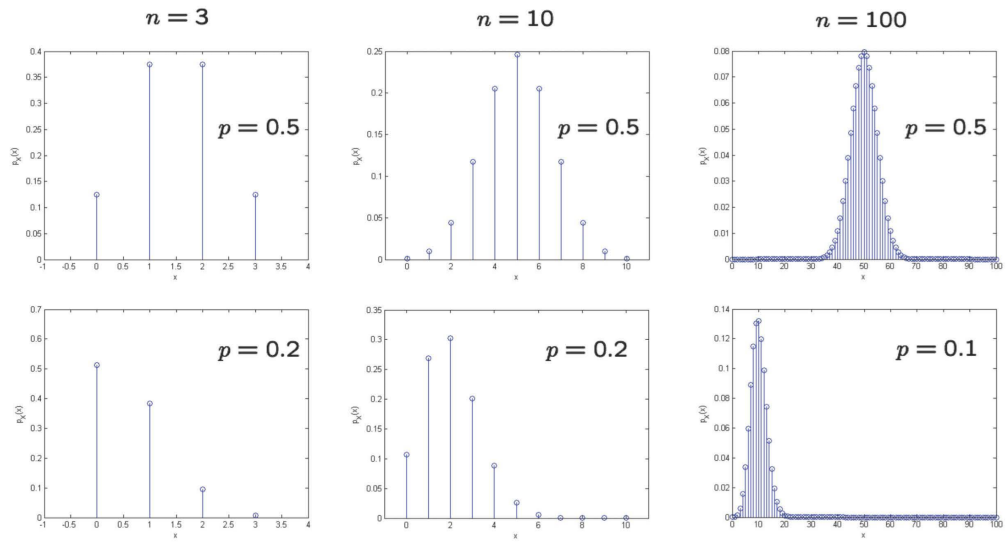


Figure 20: The PMF of Binomial variables with different parameters

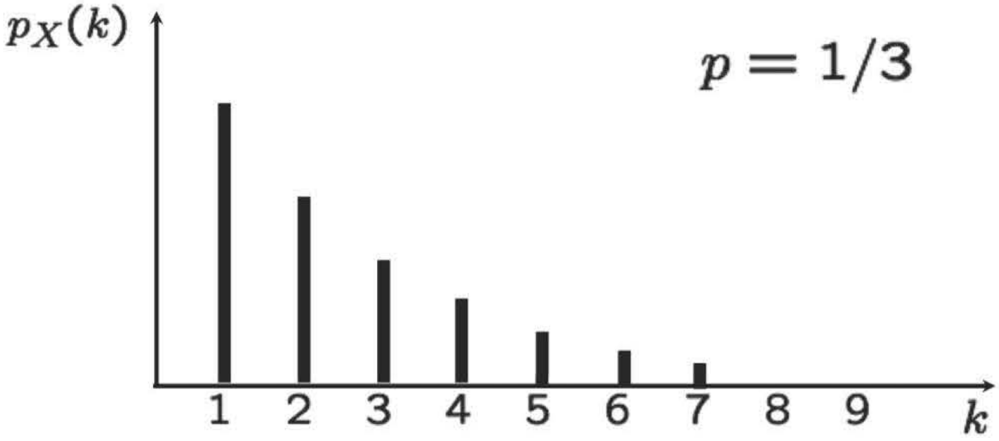


Figure 21: The PMF of Geometric variable

3.5 The Poisson Random Variable

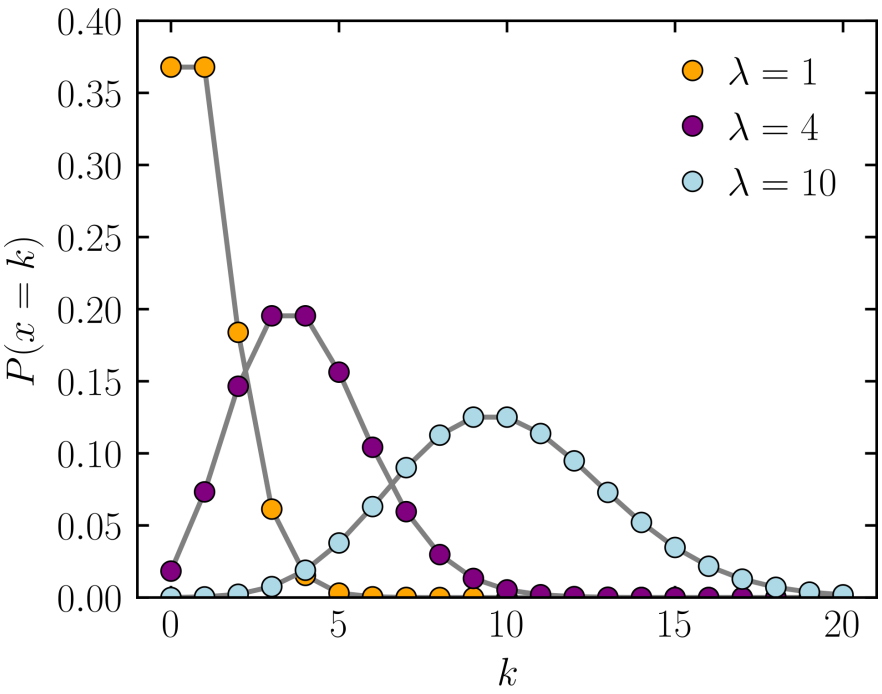
The Poisson distribution is a fundamental model for counting the number of events that occur over a fixed interval of time or space. Unlike the Binomial distribution, which counts successes in a fixed number of trials (n), the Poisson distribution operates over a continuous interval where there isn't a clear concept of "n trials." It is particularly useful for modeling events that are relatively rare.

Definition A discrete random variable X is said to follow a **Poisson distribution** with parameter λ (lambda), where $\lambda > 0$, if its Probability Mass Function (PMF) is given by:

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

We denote this as $X \sim \text{Pois}(\lambda)$. The parameter λ represents the **average rate** or mean number of events in the given interval.

Examples The Poisson distribution is used across many fields to model count data:



- **Telecommunications:** The number of phone calls arriving at a call center in one minute. If the center receives an average of 3 calls per minute, we can model this with $\lambda = 3$.
- **Biology:** The number of mutations on a particular strand of DNA after exposure to radiation.
- **Quality Control:** The number of defects found in a square meter of fabric.
- **Physics:** The number of radioactive particles detected by a Geiger counter in a 10-second interval.

Relation to the Binomial Distribution The Poisson distribution can be viewed as a limiting case of the Binomial distribution, a relationship often called the **law of rare events**.

Consider a Binomial random variable $Y \sim \text{Bin}(n, p)$. If the number of trials n is very large and the probability of success p is very small, the Binomial PMF can be approximated by the Poisson PMF. Specifically, if we let $n \to \infty$ and $p \to 0$ such that the mean np remains constant at a value λ, then the Binomial distribution converges to the Poisson distribution with parameter λ.

$$\lim_{n\to\infty, p\to 0, np=\lambda} \binom{n}{k} p^k (1-p)^{n-k} = \frac{e^{-\lambda}\lambda^k}{k!}$$

This makes the Poisson distribution an excellent tool for approximating Binomial probabilities when dealing with a large number of trials and a low probability of success, such as calculating the probability of finding 5 defective chips in a batch of 10,000 where the defect rate is 0.01%.

Lecture 6: Discrete Random Variables Part II
Expectation; Variance; Conditioning

Instructor: Prof. Abolfazl Hashemi

1 Expectation

The **expectation** (or expected value, or mean) of a random variable is a weighted average of its possible values, where the weights are the probabilities given by the PMF. It provides a summary of the central tendency of the distribution.

Definition (Expectation): The expected value of a discrete random variable X is:

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

The expectation can be thought of as the long-run average of the outcomes of many independent repetitions of the underlying experiment.

1.1 Expectations of Common Random Variables

- **Bernoulli(p):**
$$E[X] = (1 \cdot p) + (0 \cdot (1-p)) = p$$
For an indicator variable I_A, this means $E[I_A] = P(A)$.
- **Discrete Uniform on $\{0, 1, \ldots, n\}$:**
$$E[X] = \sum_{k=0}^{n} k \cdot \frac{1}{n+1} = \frac{1}{n+1} \sum_{k=0}^{n} k = \frac{1}{n+1} \frac{n(n+1)}{2} = \frac{n}{2}$$

1.2 The Expected Value Rule

To find the expectation of a function of a random variable, $Y = g(X)$, one can first find the PMF of Y and then apply the definition of expectation. A more direct method is the **expected value rule**.

Expected Value Rule: For a random variable $Y = g(X)$:

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

This rule allows us to compute the expectation of $g(X)$ using the PMF of X, without needing to find the PMF of Y. It is important to note that, in general, $E[g(X)] \neq g(E[X])$.

1.3 Linearity of Expectation

A crucial property of expectation is linearity. For any random variable X and constants a and b, let $Y = aX + b$.

Linearity Property:

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

Figure 22: The expected value rule maps values of X to Y=g(X) and computes a weighted average.

Proof: Using the expected value rule with $g(X) = aX + b$:

$$\begin{aligned} E[aX+b] &= \sum_x (ax+b)p_X(x) \\ &= \sum_x ax \cdot p_X(x) + \sum_x b \cdot p_X(x) \\ &= a\left(\sum_x x \cdot p_X(x)\right) + b\left(\sum_x p_X(x)\right) \\ &= aE[X] + b \cdot 1 \\ &= aE[X] + b \end{aligned}$$

2 Variance

While the expectation provides a measure of the central tendency of a random variable, it does not describe the spread or dispersion of its distribution. The **variance** is the most common measure of this spread.

2.1 Definition

Let X be a random variable with mean $\mu = E[X]$. The variance of X, denoted by $\text{var}(X)$, is defined as the expected value of the squared deviation of X from its mean.

Definition (Variance):

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

Using the expected value rule, this can be written as a sum over the PMF of X:

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

The **standard deviation**, $\sigma_X = \sqrt{\text{var}(X)}$, is often used as it has the same units as the random variable itself.

2.2 Properties of Variance

A more convenient formula for computation is derived as follows:

$$\begin{aligned}\text{var}(X) &= E[(X-\mu)^2] \\ &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - E[2\mu X] + E[\mu^2] \quad \text{(by linearity of expectation)} \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - \mu^2\end{aligned}$$

Computational Formula:

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

Another key property concerns linear transformations. For constants a and b:

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

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

Linear Transformation Property:

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

Note that adding a constant b shifts the distribution but does not change its spread, so the variance remains unchanged.

2.3 Variance of Common Random Variables

- **Bernoulli(p):** We know $E[X] = p$. We find $E[X^2] = 1^2 \cdot p + 0^2 \cdot (1-p) = p$.

$$\text{var}(X) = E[X^2] - (E[X])^2 = p - p^2 = p(1-p)$$

- **Discrete Uniform on $\{0, 1, \ldots, n\}$:** We know $E[X] = n/2$.

$$E[X^2] = \sum_{k=0}^{n} k^2 \frac{1}{n+1} = \frac{1}{n+1}\frac{n(n+1)(2n+1)}{6} = \frac{n(2n+1)}{6}$$

$$\text{var}(X) = \frac{n(2n+1)}{6} - \left(\frac{n}{2}\right)^2 = \frac{2n^2+n}{6} - \frac{n^2}{4} = \frac{n^2+2n}{12} = \frac{n(n+2)}{12}$$

3 Conditioning a Random Variable on an Event

We can define the distribution and expectation of a random variable conditioned on the occurrence of an event A.

Conditional PMF: The conditional PMF of X given an event A (with $P(A) > 0$) is:

$$p_{X|A}(x) = P(X = x|A) = \frac{P(\{X = x\} \cap A)}{P(A)}$$

This is a valid PMF and sums to 1.

Conditional Expectation: The conditional expectation of X given A is the expectation calculated using the conditional PMF:

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

Similarly, the conditional variance is $\text{var}(X|A) = E[X^2|A] - (E[X|A])^2$.

3.1 The Total Expectation Theorem

This theorem relates the overall expectation to conditional expectations over a partition.

Total Expectation Theorem: If $A_1, \ldots, A_n$ is a partition of the sample space:

$$E[X] = \sum_{i=1}^{n} P(A_i)E[X|A_i]$$

This is proven by starting with the definition of $E[X]$ and substituting the total probability theorem for the PMF, $p_X(x) = \sum_i P(A_i)p_{X|A_i}(x)$, then rearranging the sums.

4 The Geometric Random Variable Revisited

4.1 Memorylessness

The geometric distribution has a unique property among discrete distributions.

Memorylessness: If $X \sim \text{Geometric}(p)$, then for any positive integers n and k:

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

This means the process "forgets" its past. Given that there have been n failures, the probability of having at least k more failures is the same as the original probability of having at least k failures.

4.2 Mean of the Geometric

We can elegantly derive the mean using the total expectation theorem and memorylessness. Let $X \sim \text{Geometric}(p)$. We partition on the outcome of the first trial: $A_1 = \{\text{Success on 1st trial}\}$ and $A_2 = \{\text{Failure on 1st trial}\}$.

$$E[X] = P(A_1)E[X|A_1] + P(A_2)E[X|A_2]$$

- $P(A_1) = p$. If the first trial is a success, the process stops, so $X = 1$. Thus, $E[X|A_1] = 1$.
- $P(A_2) = 1 - p$. If the first trial is a failure, one trial has been used. By memorylessness, the *remaining* number of trials until success follows the same geometric distribution, with mean $E[X]$. The total number of trials is therefore $1 + E[X]$. So, $E[X|A_2] = 1 + E[X]$.

Substituting these into the equation:

$$E[X] = p \cdot (1) + (1 - p) \cdot (1 + E[X])$$

$$E[X] = p + 1 - p + (1 - p)E[X] \implies E[X] = 1 + (1 - p)E[X]$$

$$E[X] - (1 - p)E[X] = 1 \implies p \cdot E[X] = 1 \implies E[X] = \frac{1}{p}$$

4.3 The Poisson Random Variable

A key and convenient property of the Poisson distribution is that its expected value and variance are both equal to the parameter λ.

$$E[X] = \lambda \quad \text{and} \quad \text{Var}(X) = \lambda$$

This means that the average number of occurrences we expect to see is λ, and the spread (variance) of the observed counts is also λ.

Lecture 7: Discrete Random Variables Part III
Multiple Random Variables; Conditioning on a Random Variable; Independence of r.v.'s

Instructor: Prof. Abolfazl Hashemi

1 Multiple Random Variables

We often need to model multiple numerical quantities from a single experiment.

1.1 Joint and Marginal PMFs

The **joint PMF** of two random variables X and Y specifies their probabilistic relationship:

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

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

$$p_X(x) = \sum_y p_{X,Y}(x,y) \quad \text{and} \quad p_Y(y) = \sum_x p_{X,Y}(x,y)$$

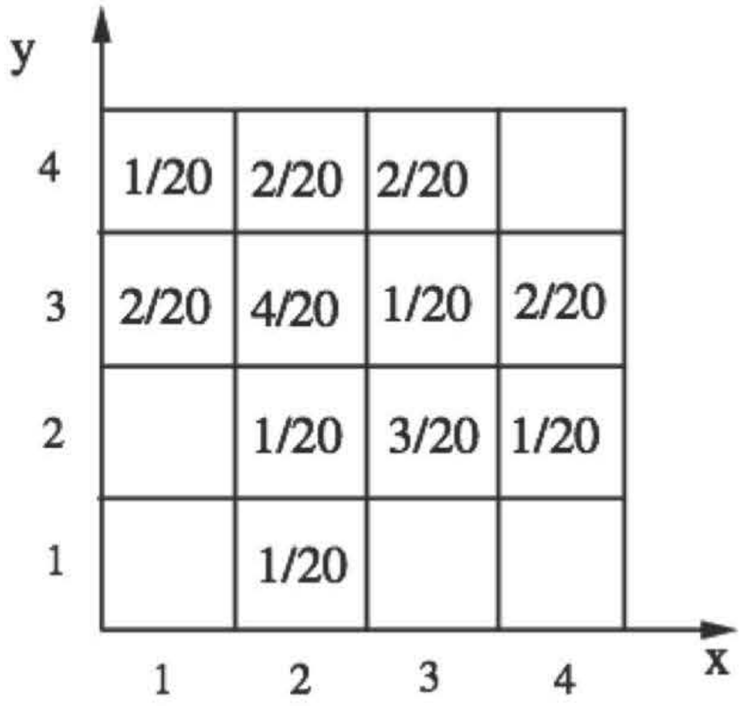


Figure 23: A joint PMF table. Summing across a row gives a marginal probability for Y; summing down a column gives a marginal for X.

1.2 Linearity of Expectation

The expected value rule extends to functions of multiple variables: $E[g(X,Y)] = \sum_x \sum_y g(x,y) p_{X,Y}(x,y)$. The most important result from this is the linearity of expectations.

Linearity of Expectation: For any random variables $X_1, \ldots, X_n$ and constants $a_1, \ldots, a_n$:

$$E[a_1X_1 + \cdots + a_nX_n] = a_1E[X_1] + \cdots + a_nE[X_n]$$

Crucially, this property holds **regardless of whether the random variables are independent**.

1.3 Mean of the Binomial

Linearity provides a simple way to find the mean of a binomial random variable, $X \sim \text{Binomial}(n, p)$. We can think of X as the sum of n independent Bernoulli indicator variables, $X = X_1 + \cdots + X_n$, where $X_i = 1$ if the i-th trial is a success. By linearity:

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

Since each X_i is a Bernoulli(p) trial, $E[X_i] = p$. Therefore:

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

2 Conditioning a Random Variable on Another

In previous lectures, we discussed conditioning a random variable on an *event*. We now extend this idea to conditioning on the value of another random variable. This is a powerful tool for breaking down complex problems into simpler, more manageable parts.

2.1 Conditional PMF

The conditional PMF of a random variable X given that another random variable Y has taken on a specific value y is a direct application of the definition of conditional probability.

Definition (Conditional PMF): The conditional PMF of X given $Y = y$ is defined as:

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

This is defined for any y for which the marginal PMF $p_Y(y)$ is positive.

For a fixed value of y, the function $p_{X|Y}(x|y)$ is a valid PMF for X, meaning it is non-negative and sums to one over all possible values of x.

2.2 Conditional Expectation

Once we have a conditional PMF, we can define a conditional expectation in the natural way.

Definition (Conditional Expectation): The conditional expectation of X given $Y = y$ is the expected value of X under the conditional PMF $p_{X|Y}(x|y)$:

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

The expected value rule also has a conditional version: $E[g(X)|Y = y] = \sum_x g(x)p_{X|Y}(x|y)$.

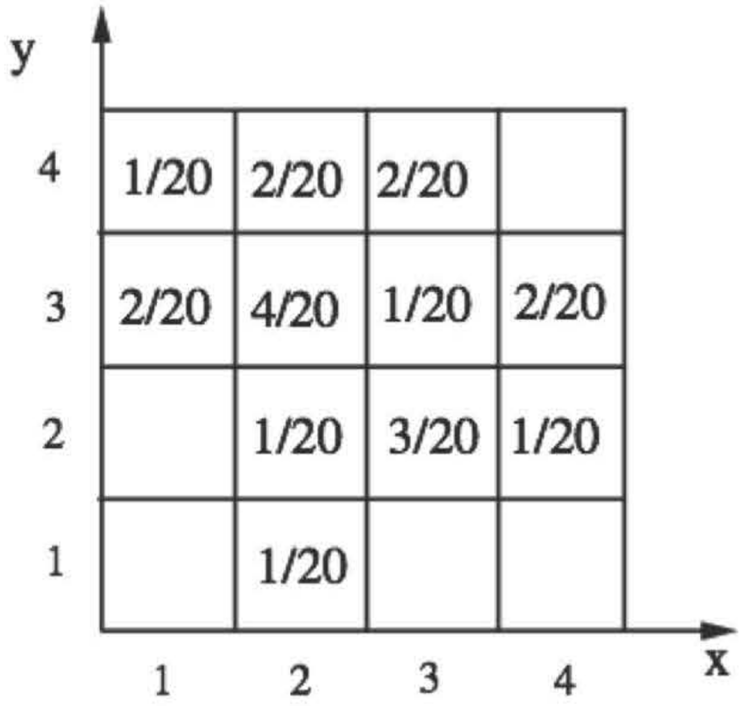


Figure 24: A joint PMF table. The conditional PMF $p_{X|Y}(x|y)$ for a given row y is found by taking the probabilities in that row and dividing them by the row's sum (the marginal probability $p_Y(y)$).

2.3 The Total Expectation Theorem

This theorem is the random variable version of the law of total probability. It states that the overall expectation of a random variable can be found by taking a weighted average of its conditional expectations.

Total Expectation Theorem: The expected value of a random variable X can be expressed as:

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

This is often written in the compact form $E[X] = E[E[X|Y]]$.

3 Independence of Random Variables

The concept of independence extends from events to random variables. Intuitively, two random variables are independent if knowing the value of one provides no information about the value of the other.

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

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

An equivalent condition is that for all x and y, $P(X = x|Y = y) = P(X = x)$.

3.1 Properties of Independent Random Variables

Independence is a very powerful property that simplifies the calculation of expectations and variances.

Expectation of a Product: For any random variables, $E[X+Y] = E[X]+E[Y]$. However, the same is not true for products.

If X and Y are independent, 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 $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$.

Variance of a Sum: For any two random variables, $\text{var}(X+Y) = \text{var}(X)+\text{var}(Y)+2(E[XY]-E[X]E[Y])$. The term $E[XY]-E[X]E[Y]$ is the covariance, and it is zero when the variables are independent.

If X and Y are independent, then:

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

This property is crucial for many calculations. Note that for independent variables, $\text{var}(X-Y) = \text{var}(X) + (-1)^2\text{var}(Y) = \text{var}(X) + \text{var}(Y)$.

4 Application: Variance of the Binomial

We can use the properties of independence to easily find the variance of a binomial random variable $X \sim \text{Binomial}(n,p)$. We represent X as the sum of n independent Bernoulli indicator variables, $X = X_1 + \cdots + X_n$, where $X_i = 1$ if the i-th trial is a success. Because the trials are independent, the random variables X_i are independent. Therefore, we can sum their variances:

$$\text{var}(X) = \text{var}(X_1 + \cdots + X_n) = \text{var}(X_1) + \cdots + \text{var}(X_n)$$

The variance of a single Bernoulli(p) variable is $p(1-p)$. Thus:

$$\text{var}(X) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

5 Application: The Hat Problem

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

Solution: The PMF of X is very complicated. A much simpler approach is to use indicator variables. Let X_i be an indicator variable for the event that person i gets their own hat back.

$$X_i = \begin{cases} 1, & \text{if person } i \text{ gets their own hat} \\ 0, & \text{otherwise} \end{cases}$$

The total number of people who get their own hat back is the sum of these indicators:

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

Mean: We use the linearity of expectation, which holds even though the X_i are not independent.

$$E[X] = E[X_1 + \cdots + X_n] = \sum_{i=1}^{n} E[X_i]$$

The expectation of an indicator is the probability of the event it indicates: $E[X_i] = P(X_i = 1)$. The probability that person i gets their own hat is $1/n$.

$$E[X] = \sum_{i=1}^{n} \frac{1}{n} = n \cdot \frac{1}{n} = 1$$

On average, exactly one person gets their own hat back, regardless of the number of people!

Variance: We use the formula $\text{var}(X) = E[X^2] - (E[X])^2 = E[X^2] - 1$.

$$X^2 = \left(\sum_{i=1}^{n} X_i\right)^2 = \sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i X_j$$

By linearity of expectation:

$$E[X^2] = \sum_{i=1}^{n} E[X_i^2] + \sum_{i \neq j} E[X_i X_j]$$

We calculate the two types of terms:

- For an indicator, $X_i^2 = X_i$, so $E[X_i^2] = E[X_i] = 1/n$.
- For $i \neq j$, the product $X_i X_j$ is also an indicator variable for the event that both person i and person j get their own hats. $E[X_i X_j] = P(X_i = 1 \text{ and } X_j = 1) = P(X_i = 1)P(X_j = 1 | X_i = 1) = \frac{1}{n} \cdot \frac{1}{n-1}$.

There are n terms of the first type and $n(n-1)$ terms of the second type.

$$E[X^2] = n \cdot \left(\frac{1}{n}\right) + n(n-1) \cdot \left(\frac{1}{n(n-1)}\right) = 1 + 1 = 2$$

Finally, the variance is:

$$\text{var}(X) = E[X^2] - (E[X])^2 = 2 - 1^2 = 1$$

The variance is also 1, regardless of the number of people.

Lecture 8: Continuous Random Variables Part I
Probability Density Functions

Instructor: Prof. Abolfazl Hashemi

1 Introduction to Continuous Random Variables

Until now, we have focused on discrete random variables, which can take on a finite or countably infinite number of values. We now turn our attention to **continuous random variables**, which can take on any value within a continuous range, such as an interval on the real number line. Examples include the height of a person, the temperature of a room, or the time until an event occurs.

Because a continuous random variable can take on an uncountably infinite number of values, we can no longer assign a positive probability to each individual value. Instead, we describe its probabilistic behavior using a **Probability Density Function (PDF)**.

2 The Probability Density Function (PDF)

The PDF, denoted $f_X(x)$, is the continuous analogue of the Probability Mass Function (PMF). While a PMF gives direct probabilities, a PDF gives a probability *density*.

Definition (PDF): For a continuous random variable X, the probability of X falling within an interval $[a, b]$ is the area under the PDF curve over that interval.

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

A valid PDF must satisfy two properties:

1. **Nonnegativity:** $f_X(x) \geq 0$ for all x.
2. **Normalization:** The total area under the PDF curve must be 1.

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

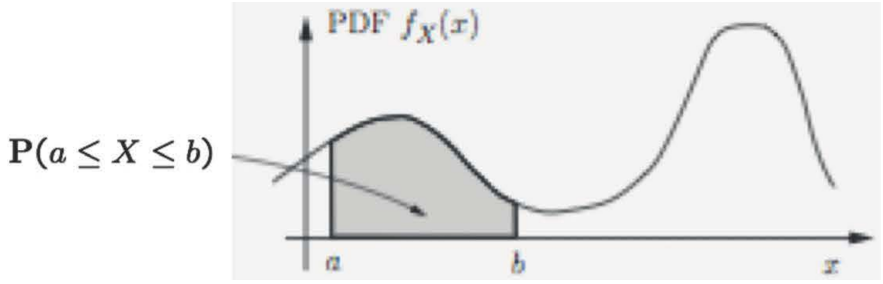


Figure 25: Probability as the area under the PDF curve.

2.1 Interpretation of the PDF

The value of the PDF at a point, $f_X(x)$, is not a probability. Instead, it tells us the relative likelihood of the random variable being near x. For a very small interval of width δ:

$$P(x \leq X \leq x + \delta) = \int_x^{x+\delta} f_X(t)dt \approx f_X(x) \cdot \delta$$

A direct and crucial consequence is that the probability of a continuous random variable taking on any single value is zero:

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

3 Expectation and Variance

The definitions of expectation and variance for continuous random variables are analogous to their discrete counterparts, with sums replaced by integrals.

Expectation: The expected value, or mean, of a continuous random variable X is:

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

Expected Value Rule: For a function $g(X)$:

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

Variance: The variance of X with mean $\mu = E[X]$ is:

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

The properties of linearity of expectation, $E[aX+b] = aE[X]+b$, and variance, $\text{var}(aX+b) = a^2\text{var}(X)$, hold just as they did in the discrete case. The computational formula $\text{var}(X) = E[X^2] - (E[X])^2$ is also still valid and extremely useful.

4 Common Continuous Distributions

4.1 The Continuous Uniform Distribution

This distribution models complete uncertainty over a fixed interval $[a, b]$.

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

- **Mean:** $E[X] = \int_a^b x \frac{1}{b-a} dx = \frac{a+b}{2}$.
- **Variance:** $\text{var}(X) = \int_a^b (x - \frac{a+b}{2})^2 \frac{1}{b-a} dx = \frac{(b-a)^2}{12}$.

4.2 The Exponential Distribution

The exponential distribution is often used to model waiting times until an event occurs, like the lifetime of a device or the time until the next phone call arrives. It is parameterized by a rate parameter $\lambda > 0$.

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

- **Mean:** $E[X] = \frac{1}{\lambda}$. A higher rate λ means a shorter average waiting time.
- **Variance:** $\text{var}(X) = \frac{1}{\lambda^2}$.

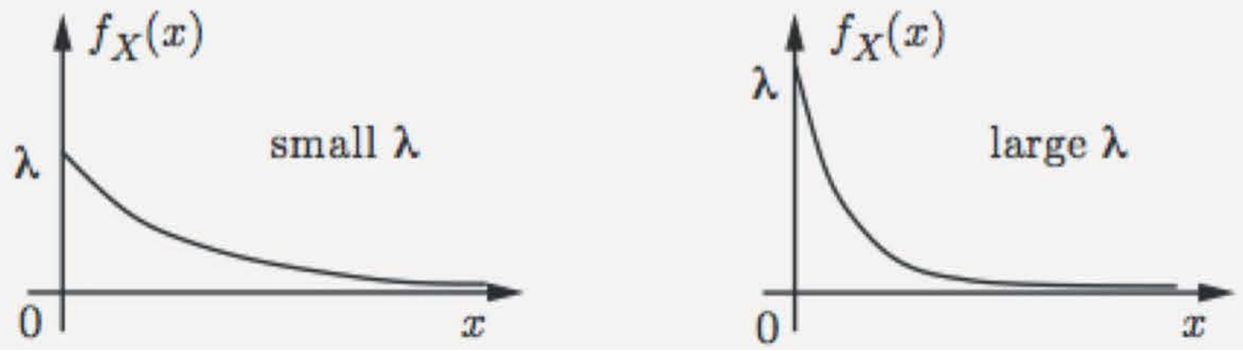


Figure 26: The exponential PDF for a small λ (long average wait) and a large λ (short average wait).

5 The Cumulative Distribution Function (CDF)

The CDF is a universal way to describe a random variable's distribution that works for both discrete and continuous cases.

Definition (CDF): The Cumulative Distribution Function of a random variable X is:

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

For a continuous random variable, the CDF is the integral of the PDF:

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

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

All CDFs are non-decreasing and satisfy $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.

6 The Normal (Gaussian) Distribution

The normal distribution is the most important in all of probability and statistics. It arises naturally in many contexts due to the Central Limit Theorem and has convenient mathematical properties.

Standard Normal $N(0,1)$**:** A normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. Its PDF is:

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

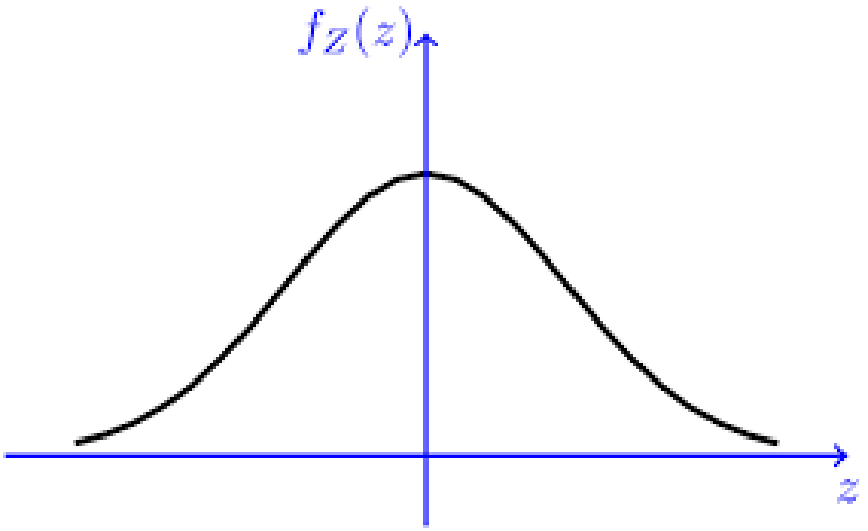


Figure 27: The PDF of a Normal Random Variable.

General Normal $N(\mu, \sigma^2)$: A normal distribution with mean μ and variance σ^2. Its PDF is:

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

6.1 Properties of Normal Random Variables

A key property is that any linear transformation of a normal random variable is also normal. If $X \sim N(\mu, \sigma^2)$, then $Y = aX+b$ is also normal with mean $E[Y] = a\mu+b$ and variance $\text{var}(Y) = a^2\sigma^2$. So, $Y \sim N(a\mu + b, a^2\sigma^2)$.

This allows us to **standardize** any normal random variable. If $X \sim N(\mu, \sigma^2)$, then the variable:

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

has a standard normal distribution, $Z \sim N(0, 1)$.

6.2 Calculating Normal Probabilities

The integral of the normal PDF cannot be expressed in terms of elementary functions. We rely on a table for the standard normal CDF, denoted $\Phi(z) = P(Z \leq z)$.

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

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

The table typically provides values for $z \geq 0$. For negative values, we use the symmetry of the "bell curve":

$$\Phi(-z) = 1 - \Phi(z)$$

Lecture 9: Continuous Random Variables Part II Conditioning on an Event; Multiple Continuous r.v.'s

Instructor: Prof. Abolfazl Hashemi

1 Conditioning a Continuous Random Variable on an Event

The concepts of conditioning extend from discrete to continuous random variables. We can define a conditional PDF and a conditional expectation given that an event A has occurred.

1.1 Conditional PDF

For a continuous random variable X and an event A with $P(A) > 0$, the **conditional PDF** of X given A, denoted $f_{X|A}(x)$, is defined such that the probability of X falling in a set B given A is the integral of the conditional PDF over B:

$$P(X \in B|A) = \int_B f_{X|A}(x)dx$$

A particularly common case is conditioning on the event that X falls within a certain subset S of the real line. In this case, the conditional PDF is zero outside of S, and a rescaled version of the original PDF inside S.

Conditional PDF given $X \in S$:

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

where $P(X \in S) = \int_S f_X(t)dt$.

1.2 Conditional Expectation

The conditional expectation is the expected value computed using the conditional PDF.

Conditional Expectation:

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

The expected value rule also applies: $E[g(X)|A] = \int_{-\infty}^{\infty} g(x) f_{X|A}(x)dx$.

1.3 Memorylessness of the Exponential PDF

The exponential distribution has a unique property among continuous distributions. Let $T \sim$ Exponential(λ) be the lifetime of a device. The probability that the device survives past time t is $P(T > t) = e^{-\lambda t}$.

Memorylessness Property: Suppose we know the device is still working at time t. The distribution of the *remaining* lifetime, $X = T - t$, is the same as the original distribution of T.

$$P(X > x|T > t) = P(T > x)$$

Proof:

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

This implies that a used "exponential" device is probabilistically as good as a new one.

2 Total Probability and Total Expectation Theorems

These theorems allow us to compute unconditional probabilities and expectations from conditional ones, using a partition of the sample space $A_1, \ldots, A_n$.

Total Probability Theorem for PDFs: The unconditional PDF is a weighted average of the conditional PDFs.

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

Total Expectation Theorem: The unconditional expectation is a weighted average of the conditional expectations.

$$E[X] = \sum_{i=1}^{n} P(A_i) E[X|A_i]$$

Example: Bill goes to the supermarket. With probability 1/3, he goes "early" at a time uniformly distributed in $[0, 2]$. With probability 2/3, he goes "late" at a time uniformly distributed in $[6, 8]$. Let X be the time he goes. Let A_1 = "early" and A_2 = "late". The conditional PDFs are: $f_{X|A_1}(x) = 1/2$ for $x \in [0, 2]$, and $f_{X|A_2}(x) = 1/2$ for $x \in [6, 8]$. The overall PDF is:

$$f_X(x) = P(A_1) f_{X|A_1}(x) + P(A_2) f_{X|A_2}(x) = \begin{cases} (1/3)(1/2) = 1/6, & 0 \le x \le 2 \\ (2/3)(1/2) = 1/3, & 6 \le x \le 8 \\ 0, & \text{otherwise} \end{cases}$$

The overall expectation is: $E[X] = P(A_1)E[X|A_1] + P(A_2)E[X|A_2] = (1/3)(1) + (2/3)(7) = 1/3 + 14/3 = 5$.

3 Multiple Continuous Random Variables

3.1 Joint PDF

Two random variables are **jointly continuous** if their probabilistic behavior is described by a joint PDF, $f_{X,Y}(x, y)$.

Joint PDF: Probability is the volume under the joint PDF surface.

$$P((X, Y) \in B) = \iint_B f_{X,Y}(x, y) dx dy$$

The joint PDF must be non-negative and integrate to 1 over the entire plane.

3.2 From Joint to Marginal PDFs

We can obtain the individual (marginal) PDF of one variable by integrating the joint PDF over all possible values of the other variable.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Example: Uniform PDF on a set S If (X,Y) is uniform over a region S, then $f_{X,Y}(x,y) = 1/\text{Area(S)}$ if $(x,y) \in S$, and 0 otherwise. The marginal PDF $f_X(x)$ is found by integrating this constant with respect to y over the vertical slice of S at that x. The result is the length of this slice divided by the total area of S.

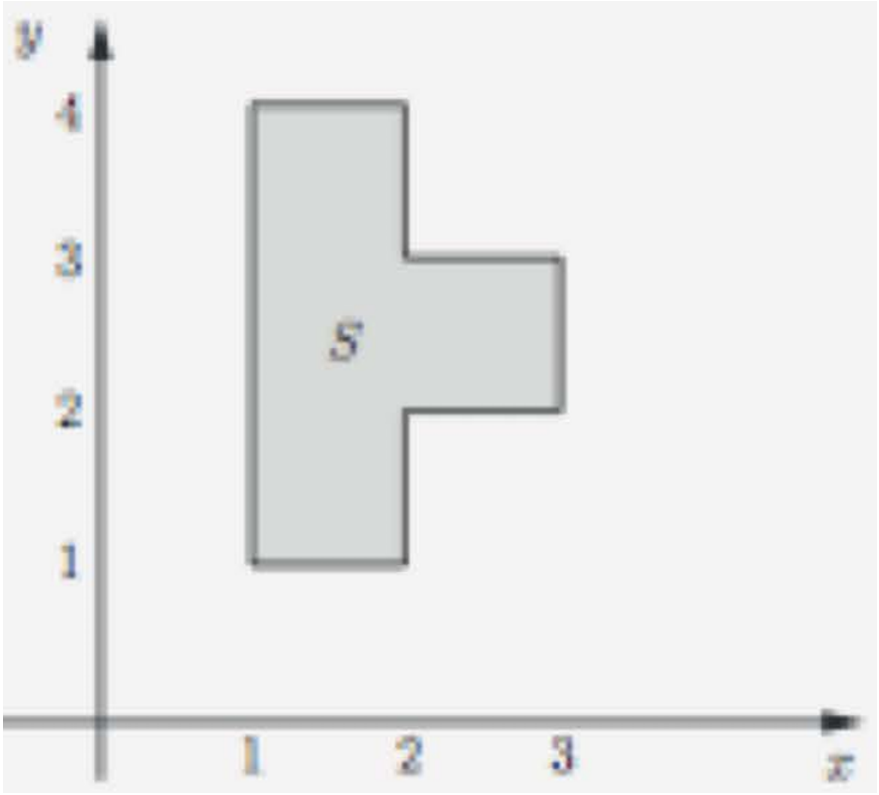


Figure 28: A uniform joint PDF over a non-rectangular region S results in non-uniform marginal PDFs.

3.3 Expectations and Joint CDFs

The tools for working with multiple random variables extend to the continuous case.

Expected Value Rule: For $Z = g(X,Y)$:

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

The linearity of expectation, $E[X+Y] = E[X] + E[Y]$, holds for all continuous random variables.

Joint CDF: The joint CDF is defined as $F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$. It is related to the joint PDF by:

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

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

Instructor: Prof. Abolfazl Hashemi

1 Conditioning a Continuous RV on Another Continuous RV

We have previously discussed conditioning a random variable on an event. We now extend this to the case where we condition a continuous random variable X on the value of another continuous random variable Y.

1.1 Conditional PDF

The conditional PDF is the primary tool for this type of analysis. It is defined in a way that is perfectly analogous to the discrete case.

Definition (Conditional PDF): The conditional PDF of a random variable X given that $Y = y$ is defined as:

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

This is defined for any y for which the marginal PDF $f_Y(y)$ is positive. For a fixed value of y, the function $f_{X|Y}(x|y)$ is a valid PDF for X, meaning it is non-negative and integrates to 1.

The conditional PDF can be visualized as a "slice" of the 3D surface of the joint PDF at a specific value of y, which is then rescaled to have a total area of 1.

From this definition, we also get a continuous version of the multiplication rule:

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

1.2 Total Probability and Expectation Theorems

The law of total probability and the total expectation theorem can be expressed by replacing the sums from the discrete case with integrals.

Total Probability Theorem for PDFs:

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

Total Expectation Theorem:

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

This is often written in the compact form $E[X] = E[E[X|Y]]$, where the outer expectation is with respect to the distribution of Y.

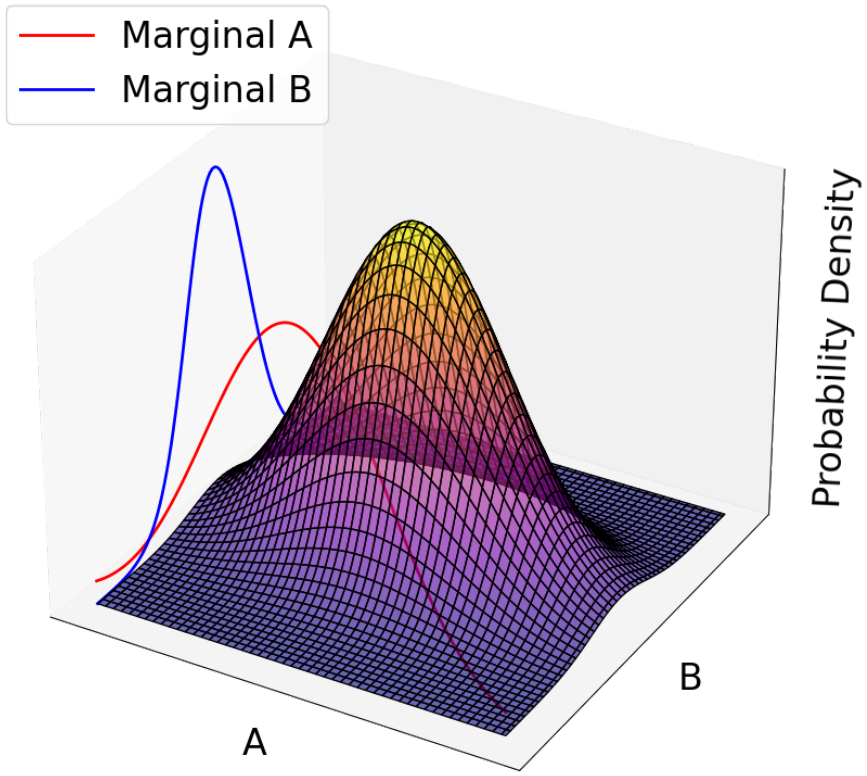


Figure 29: A slice of the joint PDF surface at a fixed y gives the shape of the conditional PDF $f_{X|Y}(x|y)$.

2 Independence of Continuous Random Variables

Two continuous random variables are independent if their joint PDF factors into the product of their marginal PDFs.

Definition (Independence):

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

This is equivalent to the conditional PDF being equal to the marginal PDF: $f_{X|Y}(x|y) = f_X(x)$. If two normal random variables are independent, their joint PDF forms a symmetric or elliptical bell shape.

3 A Comprehensive Example: Stick-Breaking

Problem: We break a stick of length ℓ twice. The first break occurs at a position X chosen uniformly on $[0, \ell]$. The second break occurs at a position Y chosen uniformly on the remaining piece, $[0, X]$. We want to find the marginal PDF of Y and the expected value of Y.

Solution:

1. **Define the PDFs:** The marginal PDF of X is uniform: $f_X(x) = 1/\ell$ for $0 \le x \le \ell$. The conditional PDF of Y given $X = x$ is uniform: $f_{Y|X}(y|x) = 1/x$ for $0 \le y \le x$.

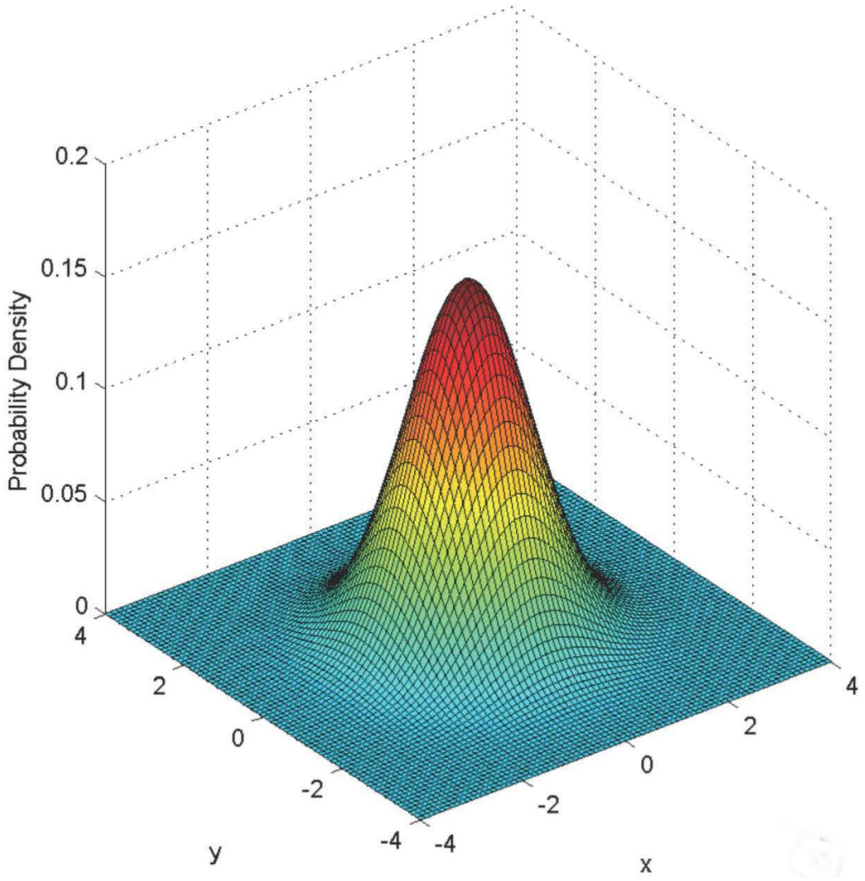


Figure 30: The joint PDF of two independent standard normal random variables.

2. **Find the Joint PDF:** Using the multiplication rule, the joint PDF is defined over the triangular region $0 \leq y \leq x \leq \ell$:

$$f_{X,Y}(x,y) = f_X(x) f_{Y|X}(y|x) = \frac{1}{\ell} \cdot \frac{1}{x} = \frac{1}{\ell x}$$

3. **Find the Marginal PDF of Y:** We integrate the joint PDF over all possible values of x. For a given y, x can range from y to ℓ.

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

This is valid for $0 < y \leq \ell$.

4. **Find the Expectation of Y:** We use the total expectation theorem, which is much simpler than integrating $y \cdot f_Y(y)$. First, find the conditional expectation of Y given $X = x$. Since Y is uniform on $[0, x]$, its mean is the midpoint:

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

Now, apply the total expectation theorem:

$$E[Y] = E[E[Y|X]] = \int_0^{\ell} E[Y|X = x] f_X(x) dx = \int_0^{\ell} \frac{x}{2} \cdot \frac{1}{\ell} dx$$

$$= \frac{1}{2\ell}\left[\frac{x^2}{2}\right]_0^{\ell} = \frac{1}{2\ell}\frac{\ell^2}{2} = \frac{\ell}{4}$$

4 Bayes' Rule: A Theme with Variations

Bayes' rule is a universal inference engine that can be adapted to any combination of discrete and continuous random variables. The core idea is always the same: Posterior $\propto$ Prior $\times$ Likelihood.

4.1 Case 1: Discrete Unknown, Continuous Measurement

Let K be a discrete random variable we want to infer, and Y be a continuous measurement.

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

Example: We send a signal $K \in \{-1, 1\}$ with $p_K(1) = p_K(-1) = 1/2$. We receive a noisy signal $Y = K + W$, where $W \sim N(0, 1)$. Find the posterior probability $p_{K|Y}(1|y)$.

- **Prior:** $p_K(1) = 1/2$.
- **Likelihood:** If $K = 1$, then $Y = 1 + W \sim N(1, 1)$. So, $f_{Y|K}(y|1) = \frac{1}{\sqrt{2\pi}}e^{-(y-1)^2/2}$. If $K = -1$, then $Y = -1 + W \sim N(-1, 1)$. So, $f_{Y|K}(y|-1) = \frac{1}{\sqrt{2\pi}}e^{-(y+1)^2/2}$.
- **Evidence (Denominator):** $f_Y(y) = \frac{1}{2}f_{Y|K}(y|1) + \frac{1}{2}f_{Y|K}(y|-1)$.
- **Posterior:**

$$p_{K|Y}(1|y) = \frac{\frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-(y-1)^2/2}}{\frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-(y-1)^2/2} + \frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-(y+1)^2/2}} = \frac{e^{-(y^2-2y+1)/2}}{e^{-(y^2-2y+1)/2} + e^{-(y^2+2y+1)/2}} = \frac{e^y}{e^y + e^{-y}}$$

4.2 Case 2: Continuous Unknown, Discrete Measurement

Let Y be a continuous random variable we want to infer, and K be a discrete measurement.

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

Example: An unknown quantity Y is modeled as uniform on $[0, 1]$. We perform a Bernoulli trial K with success probability equal to Y. We observe that the trial is a success ($K = 1$). Find the posterior PDF of Y.

- **Prior:** $f_Y(y) = 1$ for $y \in [0, 1]$.
- **Likelihood:** $p_{K|Y}(1|y) = P(K = 1|Y = y) = y$.
- **Evidence (Denominator):** $p_K(1) = \int_0^1 p_{K|Y}(1|y)f_Y(y)dy = \int_0^1 y \cdot 1 dy = [\frac{y^2}{2}]_0^1 = 1/2$.
- **Posterior:**

$$f_{Y|K}(y|1) = \frac{f_Y(y)p_{K|Y}(1|y)}{p_K(1)} = \frac{1 \cdot y}{1/2} = 2y, \quad \text{for } y \in [0, 1]$$

Observing a success makes us believe that higher values of Y are more likely, shifting the prior uniform distribution to a triangular posterior distribution.

Lecture 11: Derived Distributions

Instructor: Prof. Abolfazl Hashemi

1 Introduction

In many applications of probability, we begin with a random variable whose distribution is known and are interested in a new random variable which is a function of the original one. For instance, if X represents a random current in a circuit, we might be interested in the distribution of the power, $Y = RX^2$, where R is a constant resistance. The process of finding the probability distribution (PMF or PDF) of a function of one or more random variables is the central topic of this lecture. We will develop systematic methods to find these **derived distributions**.

We will address two main questions:

1. Given the distribution of a random variable X and a function g, how can we find the distribution of $Y = g(X)$?
2. Given the joint distribution of two random variables, X and Y, and a function g, how can we find the distribution of $Z = g(X, Y)$?

2 Functions of a Single Discrete Random Variable

When dealing with discrete random variables, the process of finding a derived distribution is a straightforward accounting exercise. If we know the PMF of X, we can find the PMF of $Y = g(X)$ by identifying all possible values of X that map to a specific value of Y and summing their probabilities.

The Discrete Method: The PMF of Y is found by summing the probabilities of all values of X that are mapped to y by the function g.

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

This formula effectively groups the probability mass from the original distribution of X onto the new set of values taken by Y.

Example: Linear Transformation of a Discrete RV

Consider a random variable X with the PMF given by $p_X(-1) = 2/6, p_X(1) = 1/6, p_X(2) = 3/6$. Let us find the PMFs of two new random variables, $Z = 2X$ and $Y = 2X + 3$.

Finding the PMF of $Z = 2X$: We map each value of X to its corresponding value of Z:

- If $X = -1$, then $Z = 2(-1) = -2$. The probability of this event is $p_X(-1) = 2/6$.
- If $X = 1$, then $Z = 2(1) = 2$. The probability is $p_X(1) = 1/6$.
- If $X = 2$, then $Z = 2(2) = 4$. The probability is $p_X(2) = 3/6$.

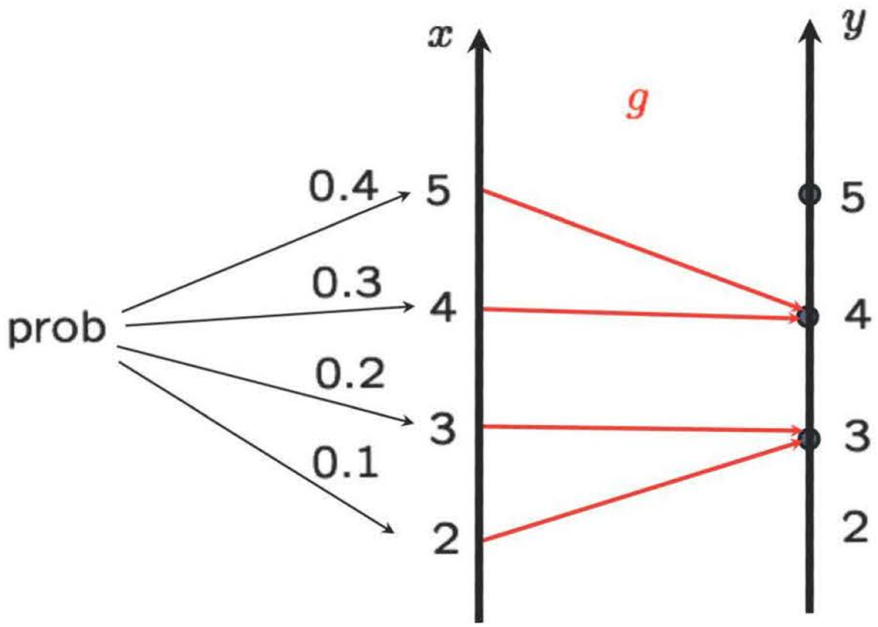


Figure 31: To find the probability $p_Y(y)$, we sum the probabilities of all x values that are mapped to y by the function g.

The resulting PMF for Z is:

$$p_Z(z) = \begin{cases} 2/6, & z = -2 \\ 1/6, & z = 2 \\ 3/6, & z = 4 \\ 0, & \text{otherwise} \end{cases}$$

Finding the PMF of $Y = 2X + 3$: Similarly, we map each value of X to its corresponding value of Y:

- If $X = -1$, then $Y = 2(-1) + 3 = 1$. The probability is $p_X(-1) = 2/6$.
- If $X = 1$, then $Y = 2(1) + 3 = 5$. The probability is $p_X(1) = 1/6$.
- If $X = 2$, then $Y = 2(2) + 3 = 7$. The probability is $p_X(2) = 3/6$.

The resulting PMF for Y is:

$$p_Y(y) = \begin{cases} 2/6, & y = 1 \\ 1/6, & y = 5 \\ 3/6, & y = 7 \\ 0, & \text{otherwise} \end{cases}$$

In the special case of a linear transformation $Y = aX + b$ with $a \neq 0$, the mapping is one-to-one, so $p_Y(y) = p_X\left(\frac{y-b}{a}\right)$.

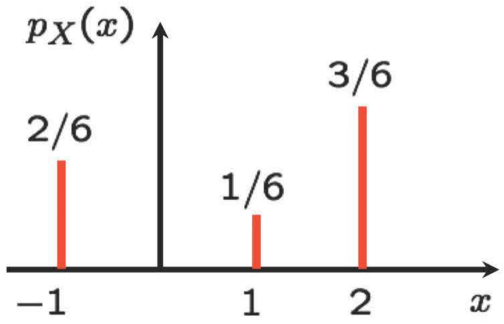


Figure 32: The PMF of the original random variable X.

3 Functions of a Single Continuous Random Variable

For continuous random variables, we cannot sum individual probabilities as they are zero. Instead, we must work with probability densities. The most general and reliable method for finding the PDF of $Y = g(X)$ involves the Cumulative Distribution Function (CDF).

The Two-Step CDF Method

Step 1: Find the CDF of Y. We calculate $F_Y(y) = P(Y \leq y)$. The core of this step is to express the event $\{Y \leq y\}$ in terms of an equivalent event for X.

$$F_Y(y) = P(g(X) \leq y) = \int_{\{x|g(x)\leq y\}} f_X(x)dx$$

Step 2: Differentiate the CDF. The PDF of Y is the derivative of its CDF.

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

The Linear Case: $Y = aX + b$

Let X be a continuous random variable and $Y = aX + b$ with $a \neq 0$. Let's apply 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 gives:

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

If $a < 0$, the inequality flips: $F_Y(y) = P(X \geq \frac{y-b}{a}) = 1 - F_X(\frac{y-b}{a})$. Differentiating gives $f_Y(y) = -f_X(\frac{y-b}{a}) \cdot \frac{1}{a} = \frac{1}{|a|} f_X(\frac{y-b}{a})$. Both cases are captured by the general formula:

General Formula for Linear Transformation:

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

The PDF of Y is a scaled (vertically by $1/|a|$) and transformed (horizontally) version of the PDF of X.

Example: Linear Function of a Normal RV. If $X \sim \mathcal{N}(\mu, \sigma^2)$, we can show that $Y = aX + b$ is also normal.

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

Using the formula, we substitute $x = (y-b)/a$:

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

This is the PDF of a normal random variable with mean $a\mu + b$ and variance $a^2\sigma^2$. Thus, $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

The Monotonic Function Method

If g is a strictly monotonic and differentiable function, there is a direct formula that bypasses the CDF calculation. Let $y = g(x)$ and $x = h(y)$ be the inverse function.

Formula for Monotonic g:

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

Example: Treadmill Time. You set treadmill speed $X \sim U[5, 10]$. Find the PDF of the time Y to run 10km. The relation is $Y = 10/X$. The PDF of X is $f_X(x) = 1/5$ for $x \in [5, 10]$. The function $g(x) = 10/x$ is monotonic decreasing. The range of Y is $[1, 2]$. The inverse is $x = h(y) = 10/y$, so $|h'(y)| = |-10/y^2| = 10/y^2$.

$$f_Y(y) = f_X\left(\frac{10}{y}\right) \cdot \frac{10}{y^2} = \frac{1}{5} \cdot \frac{10}{y^2} = \frac{2}{y^2}, \quad \text{for } y \in [1, 2]$$

Non-Monotonic Functions

If $g(x)$ is not monotonic, we must use the CDF method.

Example: $Y = X^2$. For $y > 0$, the event $Y \leq y$ corresponds to $X^2 \leq y$, or $-\sqrt{y} \leq X \leq \sqrt{y}$. The CDF is $F_Y(y) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$. Differentiating gives the PDF: $f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}$.

4 Functions of Multiple Random Variables

The CDF method extends to functions of multiple variables, such as $Z = g(X, Y)$.

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

1. **Find the CDF of Z:** $F_Z(z) = P(Z \leq z) = \iint_{\{(x,y)|g(x,y)\leq z\}} f_{X,Y}(x, y)\, dx\, dy$.
2. **Differentiate the CDF:** $f_Z(z) = \frac{dF_Z}{dz}(z)$.

Example: Let X, Y be independent $U[0,1]$. Find the PDF of $Z = Y/X$. The joint PDF is $f_{X,Y}(x,y) = 1$ on the unit square. We find $F_Z(z) = P(Y \leq zX)$.

- **Case 1:** $0 < z \leq 1$**.** The region is a triangle with area $z/2$. So $F_Z(z) = z/2$.
- **Case 2:** $z > 1$**.** The region is the unit square minus a triangle of area $1/(2z)$. So $F_Z(z) = 1 - 1/(2z)$.

Differentiating the piecewise CDF gives the PDF:

$$f_Z(z) = \begin{cases} 1/2, & 0 < z \leq 1 \\ 1/(2z^2), & z > 1 \\ 0, & \text{otherwise} \end{cases}$$

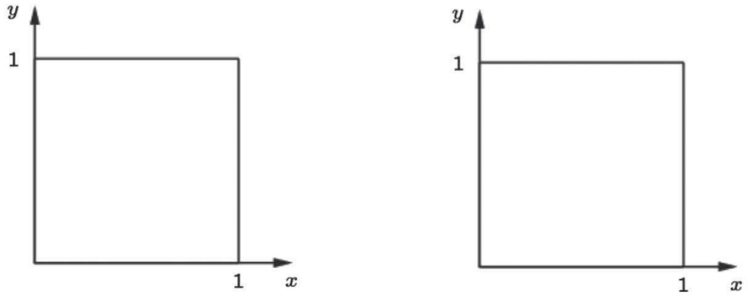


Figure 33: The region of integration for the CDF of $Z = Y/X$.

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

Instructor: Prof. Abolfazl Hashemi

1 Introduction

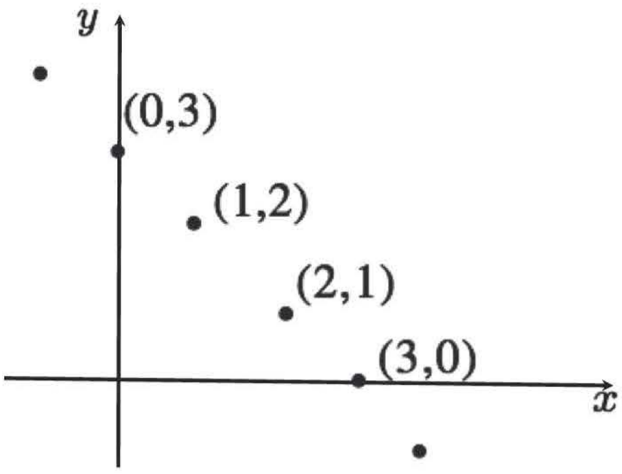
This lecture explores two fundamental topics in the study of multiple random variables. First, we will develop a systematic method for finding the probability distribution of a sum of two independent random variables, a process known as convolution. This is a powerful tool with many applications, including a key result regarding the sum of normal random variables. Second, we will move beyond the case of independence and introduce a measure for the relationship and dependency between two random variables: covariance and its normalized version, the correlation coefficient. Understanding covariance will allow us to derive a general formula for the variance of a sum of random variables.

2 The Distribution of a Sum of Independent Random Variables

2.1 The Discrete Case: Convolution

Let X and Y be independent discrete random variables with known PMFs, $p_X(x)$ and $p_Y(y)$. Our goal is to find the PMF of their sum, $Z = X + Y$.

The PMF of Z is defined as $p_Z(z) = P(Z = z) = P(X + Y = z)$. The event $\{X + Y = z\}$ can be broken down into a union of disjoint events of the form $\{X = x \text{ and } Y = z - x\}$ for all possible values of x. For example, the event $\{Z = 3\}$ corresponds to the union of events like $\{X = 0, Y = 3\}$, $\{X = 1, Y = 2\}$, $\{X = 2, Y = 1\}$, etc.



By the additivity of probability for disjoint events, we can write:

$$p_Z(z) = P(X + Y = z) = \sum_x P(X = x, Y = z - x)$$

Since X and Y are independent, the joint probability $P(X = x, Y = z - x)$ factors into the product of the marginal probabilities, $P(X = x)P(Y = z - x)$. This leads to the main result.

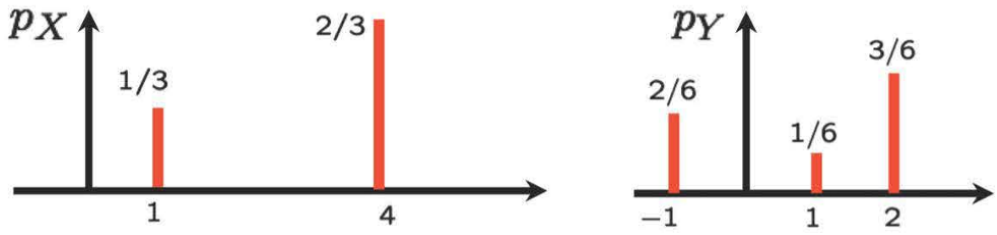
Discrete Convolution Formula:

$$p_Z(z) = \sum_x p_X(x) p_Y(z - x)$$

This operation is called the **convolution** of the PMFs of X and Y.

The convolution formula can be interpreted graphically as a "flip, shift, multiply, and sum" procedure. To compute $p_Z(z)$ for a specific value z:

1. Take the PMF of Y and flip it horizontally around the y-axis.
2. Shift this flipped PMF to the right by the value z.
3. Place the resulting PMF, $p_Y(z - x)$, directly below the PMF of X, $p_X(x)$.
4. For each x, multiply the corresponding values $p_X(x)$ and $p_Y(z - x)$.
5. Sum all of these products to get the final value of $p_Z(z)$.



2.2 The Continuous Case: Convolution

The logic for finding the PDF of the sum of two independent continuous random variables is perfectly analogous. Let $Z = X + Y$. We can derive the PDF of Z, $f_Z(z)$, using a conditional argument. By the law of total probability:

$$f_Z(z) = \int_{-\infty}^{\infty} f_{Z|X}(z|x) f_X(x) dx$$

Given the event $\{X = x\}$, the random variable Z becomes $Z = x + Y$. This is a simple linear transformation of Y. The PDF of $x + Y$ is simply the PDF of Y shifted by x, so $f_{Z|X}(z|x) = f_Y(z - x)$. Substituting this into the integral gives the convolution formula for continuous variables.

Continuous Convolution Formula:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx$$

The Sum of Independent Normal Random Variables. A cornerstone result in probability theory, which can be proven using the convolution integral, is that the sum of independent normal random variables is also normal.

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 with a mean that is the sum of the means and a variance that is the sum of the variances.

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

3 Covariance and Correlation

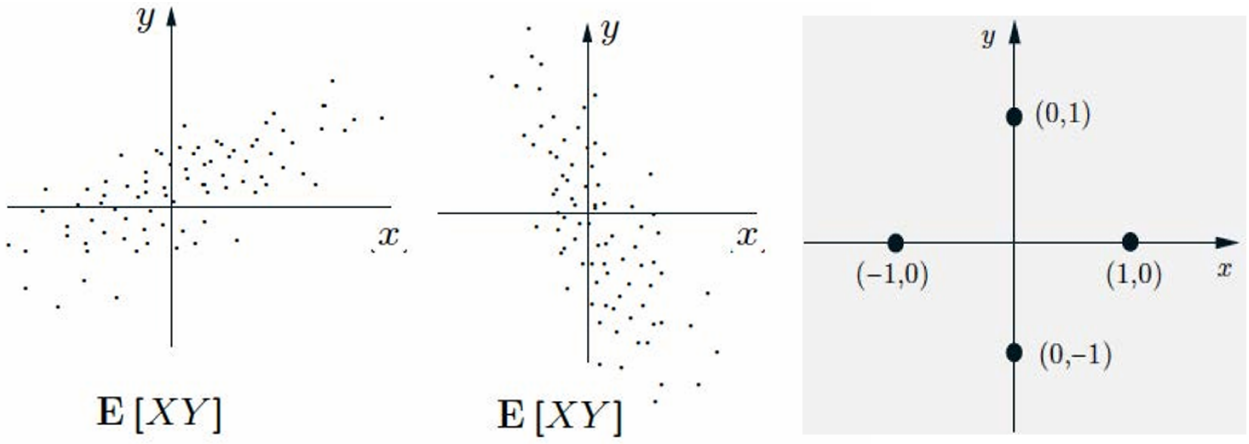
3.1 Covariance

When random variables are not independent, we need a way to measure their relationship. The **covariance** measures the degree to which two variables tend to move together relative to their means.

Definition (Covariance): The covariance of two random variables X and Y is:

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

- If $\text{cov}(X, Y) > 0$, X and Y tend to be on the same side of their respective means (e.g., when X is large, Y tends to be large). This is called positive correlation.
- If $\text{cov}(X, Y) < 0$, X and Y tend to be on opposite sides of their means. This is called negative correlation.
- If X and Y are independent, their covariance is zero. However, the converse is not true: zero covariance does not imply independence.



3.2 Properties of Covariance

- $\text{cov}(X, X) = E[(X - E[X])^2] = \text{var}(X)$.
- **Computational Formula:** $\text{cov}(X, Y) = E[XY] - E[X]E[Y]$.
- **Bilinearity:** $\text{cov}(aX + b, Y) = a \cdot \text{cov}(X, Y)$ and $\text{cov}(X, Y + Z) = \text{cov}(X, Y) + \text{cov}(X, Z)$.

3.3 Variance of a Sum

Covariance is the key to finding the variance of a sum of dependent random variables.

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

For a sum of n random variables, this generalizes to:

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

3.4 The Correlation Coefficient

The magnitude of the covariance depends on the units of X and Y. To get a standardized measure of the linear relationship, we use the **correlation coefficient**.

Definition (Correlation Coefficient):

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

- The correlation coefficient is always between -1 and 1: $-1 \leq \rho(X, Y) \leq 1$.
- $|\rho| = 1$ if and only if Y is a linear function of X (i.e., $Y = aX + b$).
- A high correlation does not imply causation; it often reflects an underlying common factor. For example, if $X = Z + V$ and $Y = Z + W$ where Z, V, W are independent, X and Y will be correlated because of the shared common factor Z.

Lecture 13: Conditional expectation and variance revisited; Sum of a random number of independent R.V.'s

Instructor: Prof. Abolfazl Hashemi

1 Introduction

This lecture delves into a more abstract and powerful view of conditional expectation and conditional variance. Instead of treating $E[X|Y = y]$ as a number that depends on a specific value y, we will begin to treat the conditional expectation $E[X|Y]$ as a random variable in its own right—a function of the random variable Y. This perspective leads to two powerful analytical tools: the law of iterated expectations and the law of total variance. We will then apply these tools to a common and important problem: finding the mean and variance of a sum of a random number of independent random variables.

2 Conditional Expectation as a Random Variable

Recall the definition of the conditional expectation of X given that the random variable Y takes a specific value y:

$$E[X|Y = y] = \sum_x x \cdot p_{X|Y}(x|y) \quad \text{(or an integral in the continuous case)}$$

This expression, for a fixed y, is a number. Let us define a function $g(y)$ that maps each possible value y to this number:

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

We can now think about what happens when we evaluate this function at the random variable Y itself. The result, $g(Y)$, is a new random variable. For any given outcome of our experiment, a value y for Y is realized, and the random variable $g(Y)$ takes on the corresponding value $g(y) = E[X|Y = y]$.

Definition: The **conditional expectation** $E[X|Y]$ is defined as the random variable $g(Y)$. It is a function of the random variable Y, and as such, it has its own distribution, mean, and variance.

2.1 The Law of Iterated Expectations

A fundamental property of this new random variable is that its expected value is simply the original, unconditional expected value of X. This is a more general statement of the total expectation theorem.

The Law of Iterated Expectations:

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

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

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

Substituting the definition of $g(y)$:

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

This is precisely the formula for the total expectation theorem, which we have already shown is equal to $E[X]$. The same logic applies in the continuous case with integrals replacing sums.

2.2 Example: Stick-Breaking

Consider a stick of length ℓ. We break it at a point Y, chosen uniformly on $[0, \ell]$. We then take the left piece of length Y and break it again at a point X, chosen uniformly on $[0, Y]$.

- The PDF of Y is $f_Y(y) = 1/\ell$ for $y \in [0, \ell]$.
- The conditional PDF of X given Y=y is $f_{X|Y}(x|y) = 1/y$ for $x \in [0, y]$.

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

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

From this, we define the conditional expectation as a random variable:

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

We can now find the overall mean of X using the law of iterated expectations:

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

3 Conditional Variance and the Law of Total Variance

We can extend this abstract view to the conditional variance.

Definition: The conditional variance $\text{var}(X|Y)$ is the random variable that takes the value $\text{var}(X|Y = y) = E[(X - E[X|Y=y])^2|Y=y]$ when $Y = y$.

For example, if $X \sim U[0, Y]$, then $\text{var}(X|Y=y) = y^2/12$. The random variable is thus $\text{var}(X|Y) = Y^2/12$.

3.1 The Law of Total Variance

This law provides a way to decompose the total variance of a random variable into two parts: the average of the conditional variances, and the variance of the conditional means.

The Law of Total Variance:

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

The term $E[\text{var}(X|Y)]$ can be thought of as the "average variability within groups," where groups are defined by the value of Y. The term $\text{var}(E[X|Y])$ can be thought of as the "variability between group means."

Proof: We start with the two terms on the right-hand side.

$$\begin{aligned} E[\text{var}(X|Y)] &= E[E[X^2|Y] - (E[X|Y])^2] \quad \text{(by def. of conditional variance)} \\ &= E[E[X^2|Y]] - E[(E[X|Y])^2] \quad \text{(by linearity of expectation)} \\ &= E[X^2] - E[(E[X|Y])^2] \quad \text{(by law of iterated expectations on } X^2) \end{aligned}$$

$$\begin{aligned} \text{var}(E[X|Y]) &= E[(E[X|Y])^2] - (E[E[X|Y]])^2 \quad \text{(by def. of variance)} \\ &= E[(E[X|Y])^2] - (E[X])^2 \quad \text{(by law of iterated expectations)} \end{aligned}$$

Adding these two expressions together, the $E[(E[X|Y])^2]$ terms cancel, leaving:

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

4 Application: Sum of a Random Number of Random Variables

Let $Y = X_1 + X_2 + \cdots + X_N$, where N is a non-negative integer random variable, and the X_i are i.i.d. random variables, also independent of N. Let $E[X_i] = E[X]$ and $\text{var}(X_i) = \text{var}(X)$.

4.1 Mean of the Sum

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

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

First, we find the inner conditional expectation. Given $N = n$, Y is a sum of a fixed number n of random variables.

$$E[Y|N = n] = E[X_1 + \cdots + X_n|N = n] = E[X_1 + \cdots + X_n] = nE[X]$$

The random variable $E[Y|N]$ is therefore $N \cdot E[X]$. Now we take the outer expectation:

$$E[Y] = E[N \cdot E[X]] = E[N]E[X]$$

4.2 Variance of the Sum

We use the law of total variance: $\text{var}(Y) = E[\text{var}(Y|N)] + \text{var}(E[Y|N])$.

1. **First Term:** $E[\textbf{var}(Y|N)]$. Given $N = n$, and since the X_i are independent, the variance of the sum is the sum of the variances:

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

   The random variable $\text{var}(Y|N)$ is therefore $N \cdot \text{var}(X)$. Its expectation is:

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

2. **Second Term:** $\textbf{var}(E[Y|N])$. We already found that the random variable $E[Y|N]$ is $N \cdot E[X]$. Its variance is:

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

Adding the two terms gives the final result:

Variance of a Random Sum:

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

Lecture 14: Bi-variate and Multivariate Normal

Instructor: Prof. Abolfazl Hashemi

1 The Bivariate Normal Distribution

1.1 Definition of Jointly Normal Random Variables

The bivariate normal distribution is a fundamental probability model for the joint behavior of two continuous random variables that are often correlated. It is widely used in statistics, econometrics, signal processing, and many other fields due to its elegant analytical properties.

Definition (Jointly Normal Random Variables) 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. That is, there exist scalars a, b, c, d such that:

$$X = aU + bV$$

$$Y = cU + dV$$

A critical consequence of this definition is that any linear combination of jointly normal random variables is itself a normal random variable. For example, if $Z = s_1X + s_2Y$, then by substitution, $Z = (s_1a + s_2c)U + (s_1b + s_2d)V$. Since this is a sum of the independent normal random variables $(s_1a + s_2c)U$ and $(s_1b + s_2d)V$, Z is also normal.

1.2 Key Property: Zero Correlation Implies Independence

For general random variables, we know that independence implies zero correlation, but the converse is not true. However, the bivariate normal distribution has a special and exceptionally useful property in this regard.

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 dramatically simplifies the analysis of jointly normal variables, as checking for independence reduces to calculating a single covariance value.

1.3 The Conditional Distribution of X Given Y

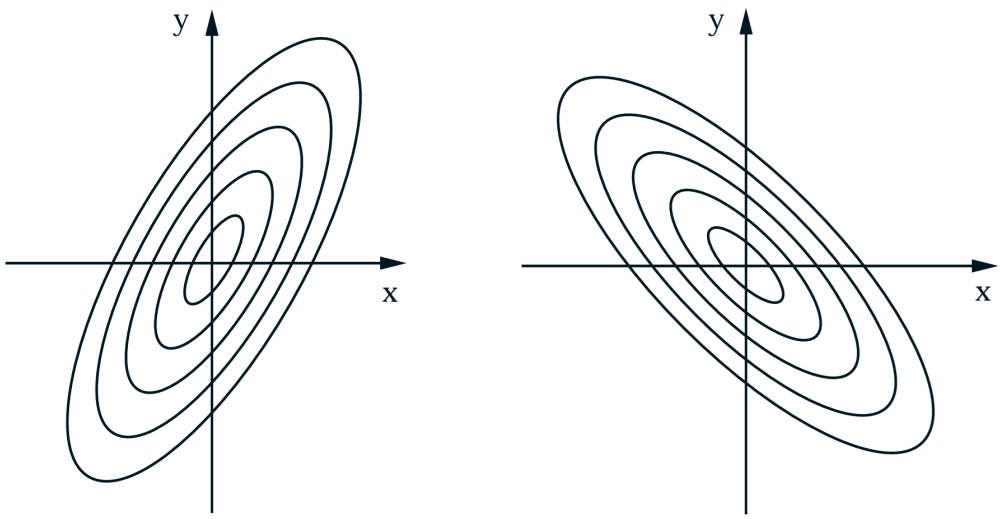
A powerful feature of the bivariate normal distribution is that the conditional distributions are also normal and easy to characterize. To derive these, we first decompose the random variable X into two components: a part that is predictable from Y and an error term that is independent of Y.

For simplicity, let's first assume X and Y have zero means. We define the linear least squares estimator of X given Y, denoted $\hat{X}$, and the corresponding estimation error, $\tilde{X}$:

$$\hat{X} = \rho\frac{\sigma_X}{\sigma_Y}Y \qquad \text{and} \qquad \tilde{X} = X - \hat{X}$$

Since $\hat{X}$ and $\tilde{X}$ are linear combinations of X and Y, they are also jointly normal with X and Y. We can show that the error $\tilde{X}$ is uncorrelated with Y:

$$E[Y\tilde{X}] = E\left[Y\left(X - \rho\frac{\sigma_X}{\sigma_Y}Y\right)\right] = E[XY] - \rho\frac{\sigma_X}{\sigma_Y}E[Y^2]$$



Using $E[XY] = \text{cov}(X, Y) = \rho\sigma_X\sigma_Y$ and $E[Y^2] = \text{var}(Y) = \sigma_Y^2$, this becomes:

$$E[Y\tilde{X}] = \rho\sigma_X\sigma_Y - \rho\frac{\sigma_X}{\sigma_Y}\sigma_Y^2 = 0$$

Since they are jointly normal and uncorrelated, $\tilde{X}$ and Y are independent. This leads to a crucial result for the conditional expectation.

Generalizing to the non-zero mean case, we arrive at the following properties.

Properties of the 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:** $\text{var}(X|Y = y) = (1 - \rho^2)\sigma_X^2$

Notice that the conditional variance does not depend on the value y of the conditioning variable.

1.4 The Form of the Bivariate Normal PDF

By applying the multiplication rule, $f_{X,Y}(x, y) = f_Y(y)f_{X|Y}(x|y)$, and substituting the formulas for the normal PDF of Y and the conditional normal PDF of X given Y, we can derive the full joint PDF. After significant algebra, the result is as follows.

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 joint distribution is fully characterized by the five parameters: the two means (μ_X, μ_Y), the two variances (σ_X^2, σ_Y^2), and the correlation coefficient (ρ).

The contours of this PDF, where $f_{X,Y}(x, y)$ is constant, are ellipses centered at the mean.

2 Examples

2.1 Example 1: Linear Combinations

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 a new random variable $Y = 2X - 3Z$. Find the PDF of Y and the conditional PDF of X given Y.

Solution:

1. **Find the PDF of Y.** Since Y is a linear combination of jointly normal variables, Y is itself normal. Its mean is $E[Y] = 2E[X] - 3E[Z] = 0$. Its variance is:

$$\begin{aligned}\sigma_Y^2 &= \text{var}(2X - 3Z) = 4\text{var}(X) + 9\text{var}(Z) - 12\text{cov}(X, Z) \\ &= 4(4) + 9(17/9) - 12(2) = 16 + 17 - 24 = 9.\end{aligned}$$

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

2. **Find the conditional PDF of X given Y.** The variables X and Y are jointly normal. We have $\sigma_X = 2$ and $\sigma_Y = 3$. We need their correlation coefficient.

$$\text{cov}(X, Y) = E[XY] = E[X(2X - 3Z)] = 2E[X^2] - 3E[XZ] = 2(4) - 3(2) = 2$$

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

   Now we use the formulas for the conditional distribution:

   - Conditional Mean: $E[X|Y = y] = 0 + \frac{1}{3}\frac{2}{3}(y - 0) = \frac{2}{9}y$.
   - Conditional Variance: $\text{var}(X|Y = y) = (1 - \rho^2)\sigma_X^2 = (1 - 1/9) \cdot 4 = \frac{32}{9}$.

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

2.2 Example 2: A Cautionary Note on Marginal vs. Joint Normality

It is important to remember that while two jointly normal random variables must have marginals that are normal, the converse is not true. It is possible to construct two random variables, X and Y, that are each marginally normal but are not jointly normal.

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: $E[XY] = E[X(ZX)] = E[Z]E[X^2] = 0 \cdot 1 = 0$.
- However, X and Y are clearly dependent. For example, if we know $X = 2$, then Y must be either 2 or -2; its value is constrained.
- Since they are dependent but uncorrelated, they cannot be jointly normal.

3 The Multivariate Normal Distribution

The concepts of the bivariate normal distribution generalize to the case of more than two random variables.

Definition (Multivariate Normal): A set of random variables $X_1, \ldots, X_n$ are said to be **jointly normal** if they are all linear functions of a set of independent normal random variables $U_1, \ldots, U_n$.

The key properties extend naturally:

- Zero correlation between any pair of the variables implies their independence.
- The conditional expectation of one variable, given some of the others, is a linear function of the conditioning variables.
- The conditional PDF of a subset of the variables, given the others, is also multivariate normal.
- The joint PDF has the form $f(\mathbf{x}) = c \cdot \exp(-q(\mathbf{x}))$, where $q(\mathbf{x})$ is a quadratic function of the variables $x_1, \ldots, x_n$.

Multivariate normal models are exceptionally common in many fields of science and engineering.

Lecture 15: Transforms and Moment Generating Functions (MGFs)

Instructor: Prof. Abolfazl Hashemi

1 Introduction to Transforms

In our study of random variables, we have so far characterized their distributions using the PMF or the PDF. In this lecture, we introduce an alternative representation of a probability law: the **transform**, also known as the **moment generating function (MGF)**. While not always as intuitive as a PMF or PDF, the transform is a powerful mathematical tool that is particularly convenient for certain types of manipulations, especially for finding moments and for analyzing sums of independent random variables.

Definition (Transform/MGF): The transform associated with a random variable X is a function $M_X(s)$ of a scalar parameter s, defined by the expectation:

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

This definition applies to any random variable. The specific calculation depends on whether the variable is discrete or continuous.

- For a **discrete** random variable, the transform is given by the sum:

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

- For a **continuous** random variable, the transform is given by the integral:

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

It is important to note that the transform $M_X(s)$ is only defined for those values of s for which the corresponding sum or integral is finite.

2 Calculating Transforms for Common Distributions

2.1 The Poisson Transform

Let X be a Poisson random variable with parameter λ. Its PMF is $p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k = 0, 1, 2, \ldots$. The transform is calculated as follows:

$$\begin{aligned} M_X(s) = E[e^{sX}] &= \sum_{k=0}^{\infty} e^{sk} \cdot e^{-\lambda}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=0}^{\infty} \frac{(e^s\lambda)^k}{k!} \end{aligned}$$

We recognize the sum as the Taylor series expansion for e^z, where $z = \lambda e^s$. Therefore:

$$M_X(s) = e^{-\lambda}e^{\lambda e^s} = e^{\lambda(e^s-1)}$$

2.2 The Exponential Transform

Let X be an exponential random variable with parameter λ. Its PDF is $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. The transform is calculated as:

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

This integral converges to a finite value only if the exponent $(s - \lambda)$ is negative, which requires $s < \lambda$. Under this condition, the upper limit evaluates to 0, and we get:

$$M_X(s) = \lambda \left(0 - \frac{e^0}{s-\lambda} \right) = \frac{-\lambda}{s-\lambda} = \frac{\lambda}{\lambda - s}, \quad \text{for } s < \lambda$$

3 Key Properties of Transforms

3.1 Moment Generation

The name "moment generating function" arises from the fact that the moments of X (i.e., $E[X], E[X^2], E[X^3], \dots$) can be easily generated from the derivatives of its transform. By differentiating the definition of the transform with respect to s, we find:

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

If we evaluate this derivative at $s = 0$, we get:

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

By repeatedly differentiating, we can find all the moments.

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

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

3.2 Uniqueness and Inversion

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

Inversion Property: The transform $M_X(s)$ uniquely determines the CDF of X, assuming $M_X(s)$ is finite for all s in some interval $[-a, a]$ where $a > 0$.

This means that if two random variables have the same transform, they must have the same distribution. This property allows us to identify the distribution of a random variable by calculating its transform and then recognizing it from a table of known transform-distribution pairs. This "pattern matching" approach is the primary method for inverting transforms.

3.3 Sums of Independent Random Variables

One of the most powerful applications of transforms is in analyzing sums of independent random variables, a task that would otherwise require a potentially difficult convolution. Let X and Y be independent, and let $Z = X + Y$.

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

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

Example: Sum of Independent Poissons. Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ be independent. Let $Z = X + Y$. The transform of the sum is:

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

We immediately recognize this as the transform of a Poisson random variable with parameter $\lambda+\mu$. By the uniqueness property, we conclude that $Z \sim \text{Poisson}(\lambda + \mu)$.

4 Advanced Application: Sum of a Random Number of RVs

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

We find the transform of Y using the law of iterated expectations:

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

Given the event $N = n$, Y is the sum of n independent random variables, so its conditional transform is $E[e^{sY}|N = n] = (M_X(s))^n$. The random variable $E[e^{sY}|N]$ is therefore $(M_X(s))^N$. Taking the outer expectation:

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

We can recognize this expression. It is the PMF of N, but with e^s replaced by the transform $M_X(s)$.

Formula for Random Sums:

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

Lecture 16: Introduction to Bayesian Inference

Instructor: Prof. Abolfazl Hashemi

1 The Big Picture of Inference

1.1 Outline

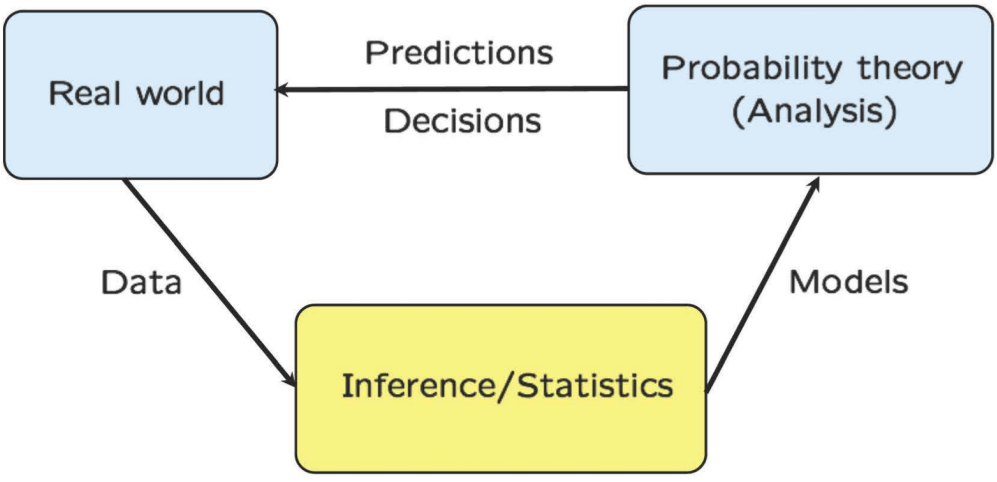
This lecture introduces the framework of Bayesian inference. We will begin by discussing the "big picture" of how inference relates to probability theory and the real world, including its motivations and a wide range of modern applications. We will also define the main problem types encountered in this field, such as hypothesis testing and estimation.

Following this conceptual overview, we will establish the general Bayesian framework. This will involve:

- Reviewing the four versions of Bayes' rule (discrete/continuous combinations) as the engine for updating our beliefs.
- Defining the output of this process, the posterior distribution.
- Discussing how to summarize this posterior distribution using point estimates, specifically the Maximum a Posteriori (MAP) estimate and the Least Mean Squares (LMS) estimate.
- Introducing the performance measures used to evaluate these estimates, such as the probability of error for hypothesis testing and the mean squared error for estimation.
- Finally, we will work through several key examples.

1.2 Inference: The Big Picture

Probability theory and statistical inference are two sides of the same coin. The relationship between them can be visualized as a loop.



- **Probability Theory (Analysis):** This is a deductive process. We start with a set of axioms and a fully specified probabilistic **Model**. We then use mathematical analysis to derive the

properties of this model and make **Predictions** about outcomes or **Decisions** based on those predictions, which apply to the **Real World**.

- **Inference/Statistics:** This is an inductive process. We begin with **Data** from the **Real World**. Our goal is to use this data to learn about the underlying process that generated it. We use statistical methods to build, select, or refine a probabilistic **Model**.

In essence, probability theory moves from model to data, while inference moves from data to model.

1.3 Inference Then and Now

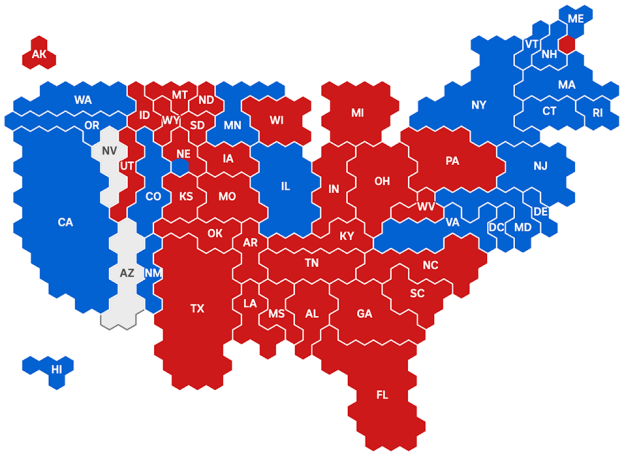
The fundamental questions of inference have existed for a long time, but the scale and complexity of the problems have changed dramatically.

- **Then:** Inference was characterized by small, sparse data sets. A typical problem might involve observing that "10 patients were treated: 3 died" while "10 patients were not treated: 5 died," and trying to draw a conclusion. The limited data required simple models and led to conclusions with high uncertainty.
- **Now:** The modern world is defined by "Big Data," "Big Models," and "Big Computers." We have access to massive datasets from complex systems and the computational power to build and analyze equally complex probabilistic models. This allows for unprecedented accuracy and new types of applications.

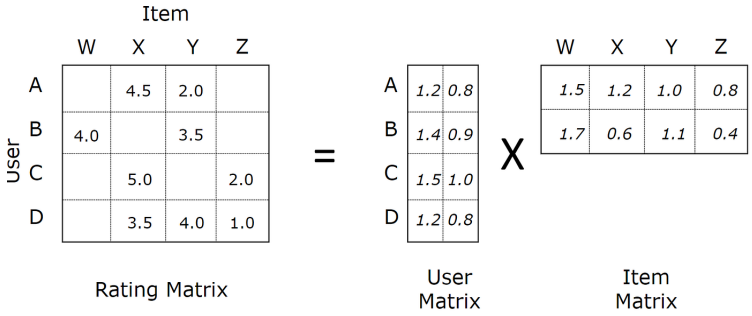
2 A Sample of Application Domains

The methods of statistical inference are fundamental to nearly every quantitative field.

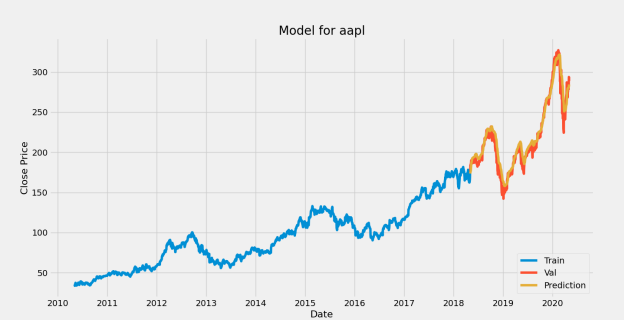
- **Design and Interpretation of Experiments:** This is the classical application, most notably in political polling and election forecasting. An inference model takes polling data (data) to build a model of voter preferences (model), which is then used to predict an election outcome (prediction).



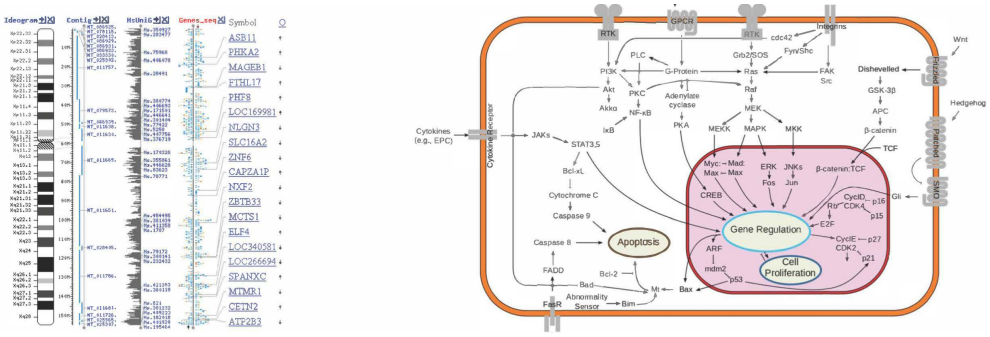
- **Marketing and Recommendation Systems:** Modern online platforms use inference to model user behavior. In a recommendation system (like the Netflix competition), the system observes a sparse matrix of user ratings for movies and must infer the missing ratings to recommend new movies to a user.



- **Finance:** Inference is used to model the behavior of financial markets, asset prices, and volatility based on historical time-series data.



- **Life Sciences:** Fields like genomics, systems biology, and neuroscience rely on inference to build models from enormous and complex experimental datasets, such as mapping gene expression or understanding neural pathways.



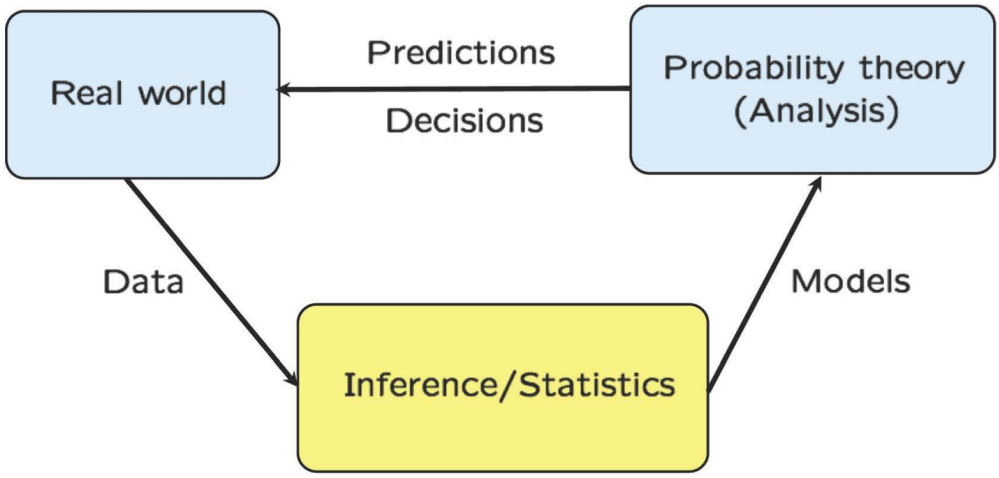
- **Physical and Environmental Sciences:** Inference is used for modeling and monitoring global climate, pollution, interpreting data from physics experiments, and processing astronomical data.
- **Signal Processing (ECE):** This field is built on inference. Examples include:
  - Communication systems (extracting a signal from noise)
  - Speech and image processing and understanding
  - Tracking objects with radar or vision
  - Positioning systems like GPS

– Detection of abnormal events

2.1 Model Building vs. Variable Estimation

Inference problems can often be divided into two main categories, which we can illustrate with a simple signal-plus-noise model: $X = aS + W$. Here, X is our observation, S is the true signal, a is a parameter of the system (like channel attenuation), and W is random noise.

- **Model Building:** In this problem, we know the signal S (e.g., we sent a known test signal) and we observe X. Our goal is to infer the unknown parameter a, which defines the model of our system.
- **Variable Estimation:** In this problem, we know the system parameter a and we observe X. Our goal is to infer the value of the original, unobserved signal S.



2.2 Hypothesis Testing vs. Estimation

We can also classify inference problems by the nature of the unknown quantity.

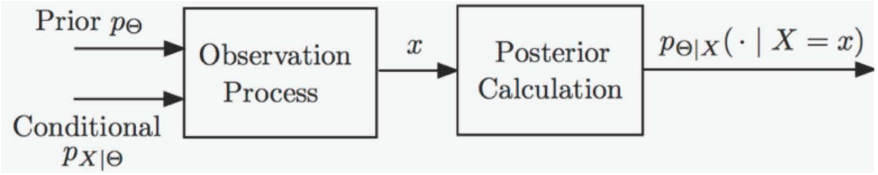
- **Hypothesis Testing:** The unknown quantity, Θ, is discrete and takes one of a few possible values. For example, $\Theta \in \{\text{airplane, bird}\}$ or $\Theta \in \{\text{disease, no disease}\}$. The goal is to make a decision and select the correct hypothesis, aiming to minimize the probability of making an incorrect decision.
- **Estimation:** The unknown, Θ, is a continuous numerical value (or a vector of values). For example, Θ could be the precise location of the airplane, the temperature of a system, or the bias of a coin. The goal is to produce an estimate, $\hat{\theta}$, that is "close" to the true unknown value Θ.

3 The Bayesian Inference Framework

The Bayesian approach to inference is a unified framework that treats all unknown quantities as random variables.

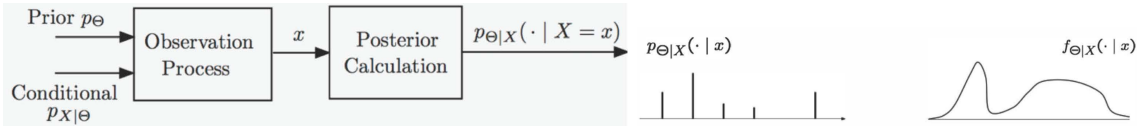
The framework consists of three main components:

1. **The Prior Distribution (p_Θ or f_Θ):** This is a probability distribution (PMF or PDF) that represents our beliefs about the unknown quantity Θ *before* we have seen any data. This prior can come from physical constraints (e.g., Θ must be in $[0, 1]$), symmetry, results from earlier studies, or even a subjective belief.
2. **The Observation Model ($p_{X|\Theta}$ or $f_{X|\Theta}$):** This is a conditional distribution (PMF or PDF) that describes the data-generating process. It tells us the probability of observing the data X given that the unknown parameter Θ has a specific value. This is also called the "likelihood" of the data given the parameter.
3. **The Posterior Distribution ($p_{\Theta|X}$ or $f_{\Theta|X}$):** After we make an observation $X = x$, we use Bayes' rule to update our beliefs. The result is the posterior distribution, which represents our new, refined belief about Θ *after* incorporating the evidence from the data.



3.1 The Output of Bayesian Inference

The complete answer to a Bayesian inference problem is the full posterior distribution. This PMF or PDF encapsulates all the information we have about Θ.



From this complete answer, we can derive simpler, more actionable summaries. The most common summaries are **point estimates** (a single "best guess") and **error analyses** (a measure of our confidence in that guess).

3.2 Point Estimates

An **estimator** is a rule, or function $g(X)$, that maps an observation X to a guess $\hat{\Theta}$. The resulting number, $\hat{\theta} = g(x)$, is the **estimate**. There are two major Bayesian estimators:

1. **Maximum a Posteriori (MAP) Estimate:** This is the value of θ that maximizes the posterior distribution. It is the "peak" of the posterior, or the most likely value of Θ given the data.
$$\hat{\theta}_{MAP} = \arg\max_\theta p_{\Theta|X}(\theta|x) \quad \text{or} \quad \arg\max_\theta f_{\Theta|X}(\theta|x)$$
2. **Conditional Expectation / Least Mean Squares (LMS) Estimate:** This is the expected value (or mean) of the posterior distribution.
$$\hat{\theta}_{LMS} = E[\Theta|X = x]$$
As we will see later, this estimate is the one that minimizes the mean squared error.

4 The Four Cases of Bayesian Inference

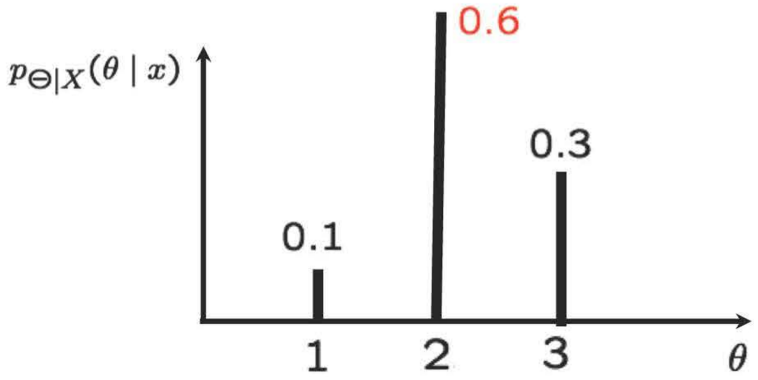
4.1 Case 1: Discrete Θ, Discrete X (Hypothesis Testing)

This is the classic discrete hypothesis testing problem.

- **Bayes' Rule:** $p_{\Theta|X}(\theta|x) = \frac{p_\Theta(\theta)p_{X|\Theta}(x|\theta)}{p_X(x)}$
- **Evidence:** $p_X(x) = \sum_{\theta'} p_\Theta(\theta')p_{X|\Theta}(x|\theta')$

In the example shown in the figure, Θ can be 1, 2, or 3. Given the observation x, the posterior probabilities are $p_{\Theta|X}(1|x) = 0.1$, $p_{\Theta|X}(2|x) = 0.6$, and $p_{\Theta|X}(3|x) = 0.3$.

- **MAP Estimate:** The MAP estimate is $\hat{\theta} = 2$, as this value has the highest posterior probability (0.6).
- **Performance (Probability of Error):** Given $X = x$, the probability of error associated with our MAP estimate is $P(\hat{\theta} \neq \Theta|X = x) = 1 - p_{\Theta|X}(\hat{\theta}|x) = 1 - 0.6 = 0.4$. The MAP rule is optimal because it minimizes this conditional probability of error for every possible x.



4.2 Case 2: Discrete Θ, Continuous X (Signal Detection)

This is a standard signal detection problem, e.g., $\Theta \in \{1, 2, 3\}$ is a transmitted symbol and $X = \Theta + W$ is the received signal, corrupted by continuous noise $W \sim N(0, \sigma^2)$.

- **Bayes' Rule:** $p_{\Theta|X}(\theta|x) = \frac{p_\Theta(\theta)f_{X|\Theta}(x|\theta)}{f_X(x)}$
- **Evidence:** $f_X(x) = \sum_{\theta'} p_\Theta(\theta')f_{X|\Theta}(x|\theta')$

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

4.3 Case 3: Continuous Θ, Continuous X (Parameter Estimation)

This is the classic parameter estimation problem, e.g., estimating a signal amplitude Θ from a noisy measurement $X = \Theta + W$.

- **Bayes' Rule:** $f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta)f_{X|\Theta}(x|\theta)}{f_X(x)}$
- **Evidence:** $f_X(x) = \int f_\Theta(\theta')f_{X|\Theta}(x|\theta')d\theta'$

Since $P(\hat{\Theta} = \Theta) = 0$ for continuous variables, we use a different performance metric: the **Mean Squared Error (MSE)**. The LMS estimate $E[\Theta|X = x]$ is defined as the estimate that minimizes this MSE.

4.4 Case 4: Continuous Θ, Discrete X (Coin Bias Example)

This is a core problem: inferring the unknown bias Θ of a coin (a continuous value in $[0, 1]$) after observing $K = k$ heads in n discrete tosses.

- **Bayes' Rule:** $f_{\Theta|K}(\theta|k) = \frac{f_\Theta(\theta)p_{K|\Theta}(k|\theta)}{p_K(k)}$
- **Likelihood:** $p_{K|\Theta}(k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$ (Binomial PMF)

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

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

This posterior distribution is known as the **Beta distribution**.

Point Estimates for the Coin Bias Problem:

- **MAP Estimate:** We maximize $f_{\Theta|K}(\theta|k)$ by finding the peak of $\theta^k(1-\theta)^{n-k}$. Taking the derivative with respect to θ and setting it to zero yields $k(1-\theta) = (n-k)\theta$, which solves to $\hat{\theta}_{MAP} = k/n$.
- **LMS Estimate:** We must compute the mean of the Beta posterior. Using the known integral $\int_0^1 \theta^\alpha(1-\theta)^\beta d\theta = \frac{\alpha!\beta!}{(\alpha+\beta+1)!}$, we find:

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

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

5 Summary

- Bayesian inference starts with a **prior** $p_\Theta(\cdot)$ and an **observation model** $p_{X|\Theta}(\cdot|\cdot)$.
- It uses **Bayes' rule** to compute the **posterior** $p_{\Theta|X}(\cdot|x)$ after observing data $X = x$.
- An **estimator** $\hat{\Theta} = g(X)$ is a rule; an **estimate** $\hat{\theta} = g(x)$ is a number.
- **MAP** estimates maximize the posterior. This is optimal for hypothesis testing as it minimizes the probability of error.
- **LMS** estimates compute the mean of the posterior, $E[\Theta|X = x]$. This is optimal for estimation as it minimizes the Mean Squared Error, $E[(\hat{\Theta} - \Theta)^2]$.

Lecture 17: Linear Models With Normal Noise

Instructor: Prof. Abolfazl Hashemi

1 Introduction to Linear Models with Normal Noise

This lecture explores a particularly important and widely used class of models in estimation theory: linear models where both the underlying parameters and the observation noise are assumed to follow normal (Gaussian) distributions.

The general form of the model we consider involves observations X_i that are linear combinations of unknown parameters Θ_j, corrupted by additive noise W_i:

$$X_i = \sum_{j=1}^{m} a_{ij}\Theta_j + W_i$$

Here, a_{ij} are known coefficients. A key assumption in this lecture is that the noise terms W_i and the parameters Θ_j are mutually independent random variables, and all follow normal distributions.

This model structure is highly prevalent in various fields due to several advantageous properties:

- **Convenience and Tractability:** Normal distributions have convenient mathematical properties that simplify analysis significantly.
- **Bayes' Rule Application:** When priors and likelihoods are normal, the resulting posterior distribution is also normal. This property is known as conjugacy and greatly simplifies Bayesian inference.
- **Coincidence of Estimators:** For these models, the Maximum A Posteriori (MAP) estimate and the Least Mean Squares (LMS) estimate (which is equivalent to the conditional expectation $E[\Theta|X]$) coincide.
- **Simple Estimator Formulas:** The resulting MAP/LMS estimators are often linear functions of the observations, leading to straightforward calculation.
- **Analytical Performance:** Measures like the Mean Squared Error (MSE) can often be calculated analytically.

We will illustrate these concepts using examples, culminating in a trajectory estimation problem.

2 Recognizing Normal PDFs

Before diving into estimation, it's crucial to be able to recognize the probability density function (PDF) of a normal random variable, possibly scaled or unnormalized. A random variable X following a normal distribution with mean μ and variance σ^2, denoted $X \sim N(\mu, \sigma^2)$, has the PDF:

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

The key feature is that the exponent is a quadratic function of x, specifically $-(x-\mu)^2/(2\sigma^2)$.

Any function of the form $f(x) = c \cdot e^{-Q(x)}$, where c is a constant and $Q(x)$ is a quadratic function of x with a positive coefficient for the x^2 term, corresponds to a normal PDF (possibly unnormalized).

For instance, consider $g(x) = c \cdot e^{-8(x-3)^2}$. Comparing this to the standard form, we can identify:

$$\frac{(x-\mu)^2}{2\sigma^2} = 8(x-3)^2$$

This implies $\mu = 3$ and $2\sigma^2 = 1/8$, so $\sigma^2 = 1/16$. Thus, $g(x)$ represents the PDF of a $N(3, 1/16)$ random variable, scaled by some constant c.

More generally, if we encounter a function like:

$$f_X(x) = c \cdot e^{-(\alpha x^2 + \beta x + \gamma)}$$

where $\alpha > 0$, we can rewrite the exponent by completing the square to match the normal PDF structure:

$$\alpha x^2 + \beta x + \gamma = \alpha \left(x^2 + \frac{\beta}{\alpha} x \right) + \gamma = \alpha \left(x + \frac{\beta}{2\alpha} \right)^2 - \frac{\beta^2}{4\alpha} + \gamma$$

Comparing $\alpha \left(x + \frac{\beta}{2\alpha} \right)^2$ with $\frac{(x-\mu)^2}{2\sigma^2}$, we identify:

$$\mu = -\frac{\beta}{2\alpha}$$

$$\frac{1}{2\sigma^2} = \alpha \implies \sigma^2 = \frac{1}{2\alpha}$$

Thus, any PDF proportional to $e^{-(\alpha x^2 + \beta x + \gamma)}$ with $\alpha > 0$ corresponds to a normal distribution $N(-\beta/2\alpha, 1/(2\alpha))$.

3 Estimating a Normal Random Variable with Additive Normal Noise

Let's start with the simplest case: estimating a single unknown parameter Θ based on a single observation X, where the observation is the sum of the parameter and independent noise W.

$$X = \Theta + W$$

Assume both Θ and W are normal random variables and are independent. Specifically, let's first consider the standard case: $\Theta \sim N(0, 1)$ and $W \sim N(0, 1)$.

Our goal is to find the posterior PDF $f_{\Theta|X}(\theta|x)$ using Bayes' rule:

$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta) f_{X|\Theta}(x|\theta)}{f_X(x)}$$

where $f_X(x) = \int_{-\infty}^{\infty} f_\Theta(\theta) f_{X|\Theta}(x|\theta) d\theta$ is the marginal PDF of X, acting as a normalization constant.

First, we need the likelihood function $f_{X|\Theta}(x|\theta)$. Given a specific value $\Theta = \theta$, the observation $X = \theta + W$. Since $W \sim N(0, 1)$, X conditioned on $\Theta = \theta$ is a shifted normal random variable: $X|\{\Theta = \theta\} \sim N(\theta, 1)$. Thus, the likelihood is:

$$f_{X|\Theta}(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$$

The prior PDF for Θ is:

$$f_\Theta(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}}$$

Now, applying Bayes' rule (ignoring the denominator $f_X(x)$ for now, as it's just a normalizing constant with respect to θ):

$$\begin{aligned}
f_{\Theta|X}(\theta|x) &\propto f_\Theta(\theta) f_{X|\Theta}(x|\theta) \\
&\propto \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}}\right)\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}\right) \\
&\propto \exp\left\{-\frac{\theta^2}{2} - \frac{(x-\theta)^2}{2}\right\} \\
&= \exp\left\{-\frac{1}{2}(\theta^2 + x^2 - 2x\theta + \theta^2)\right\} \\
&= \exp\left\{-\frac{1}{2}(2\theta^2 - 2x\theta + x^2)\right\} \\
&= \exp\left\{-\left(\theta^2 - x\theta + \frac{x^2}{2}\right)\right\}
\end{aligned}$$

The expression in the exponent, $\theta^2 - x\theta + x^2/2$, is quadratic in θ. The coefficient of θ^2 is $1 (> 0)$. This confirms that the posterior distribution $f_{\Theta|X}(\theta|x)$ is normal. To find its mean and variance, we complete the square for the terms involving θ:

$$\theta^2 - x\theta = \left(\theta - \frac{x}{2}\right)^2 - \frac{x^2}{4}$$

Substituting this back into the exponent:

$$-\left(\left(\theta - \frac{x}{2}\right)^2 - \frac{x^2}{4} + \frac{x^2}{2}\right) = -\left(\left(\theta - \frac{x}{2}\right)^2 + \frac{x^2}{4}\right)$$

So, the posterior PDF is proportional to:

$$f_{\Theta|X}(\theta|x) \propto \exp\left\{-\left(\theta - \frac{x}{2}\right)^2\right\} \exp\left\{-\frac{x^2}{4}\right\}$$

Since the term $\exp\{-x^2/4\}$ does not depend on θ, it gets absorbed into the normalization constant. The part depending on θ is:

$$f_{\Theta|X}(\theta|x) \propto \exp\left\{-\frac{(\theta - x/2)^2}{1}\right\}$$

Comparing this to the general normal PDF form $\exp\{-\frac{(\theta-\mu_{\text{post}})^2}{2\sigma^2_{\text{post}}}\}$, we identify the posterior mean μ_{post} and posterior variance σ^2_{post}:

$$\mu_{\text{post}} = \frac{x}{2}$$

$$2\sigma^2_{\text{post}} = 1 \implies \sigma^2_{\text{post}} = \frac{1}{2}$$

Therefore, the posterior distribution is $\Theta|\{X = x\} \sim N(x/2, 1/2)$.

The MAP estimate $\hat{\theta}_{MAP}$ is the value of θ that maximizes the posterior PDF. For a normal distribution, this is simply the mean.

$$\hat{\theta}_{MAP} = \mu_{\text{post}} = \frac{x}{2}$$

The LMS estimate $\hat{\theta}_{LMS}$ is the conditional expectation $E[\Theta|X = x]$. For any distribution, this is the mean of the conditional distribution.

$$\hat{\theta}_{LMS} = E[\Theta|X = x] = \mu_{\text{post}} = \frac{x}{2}$$

In this case, MAP and LMS estimates coincide. The estimator, viewed as a function of the random variable X, is:

$$\hat{\Theta}_{MAP} = \hat{\Theta}_{LMS} = \mathbb{E}[\Theta|X] = \frac{X}{2}$$

This estimator is a linear function of the observation X.

These key findings hold even for general normal priors and noise: if $\Theta \sim N(\mu_0, \sigma_0^2)$ and $W \sim N(0, \sigma_W^2)$, and $X = \Theta + W$, then:

- The posterior $f_{\Theta|X}(\theta|x)$ is normal.
- The MAP and LMS estimators coincide.
- The estimator $\hat{\Theta} = \mathbb{E}[\Theta|X]$ is a linear function of X, specifically of the form $\hat{\Theta} = aX + b$. (The exact formula involves a weighted average of the prior mean μ_0 and the observation x).

4 Estimation with Multiple Observations

Now, consider the case where we have multiple independent observations $X_1, ..., X_n$, all related to the same unknown parameter Θ.

$$X_1 = \Theta + W_1$$
$$\vdots$$
$$X_n = \Theta + W_n$$

Assume Θ, W_1, ..., W_n are mutually independent. Let the prior for Θ be $\Theta \sim N(x_0, \sigma_0^2)$ and the noise terms be $W_i \sim N(0, \sigma_i^2)$. Note that we use x_0 to denote the prior mean to distinguish it from the observations $x_1, ..., x_n$.

We again use Bayes' rule for the vector of observations $X = (X_1, ..., X_n)$:

$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta) f_{X|\Theta}(x|\theta)}{f_X(x)}$$

The prior is $f_\Theta(\theta) \propto \exp\{-\frac{(\theta - x_0)^2}{2\sigma_0^2}\}$.

To find the joint likelihood $f_{X|\Theta}(x|\theta)$, we use the fact that given $\Theta = \theta$, the observations $X_i = \theta + W_i$ are conditionally independent. This is because the W_i are independent. Given $\Theta = \theta$, each X_i is normal: $X_i|\{\Theta = \theta\} \sim N(\theta, \sigma_i^2)$. The individual likelihood for observation X_i is:

$$f_{X_i|\Theta}(x_i|\theta) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x_i - \theta)^2}{2\sigma_i^2}}$$

Due to conditional independence, the joint likelihood is the product of the individual likelihoods:

$$f_{X|\Theta}(x|\theta) = \prod_{i=1}^{n} f_{X_i|\Theta}(x_i|\theta) \propto \prod_{i=1}^{n} \exp\left\{-\frac{(x_i-\theta)^2}{2\sigma_i^2}\right\} = \exp\left\{-\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2\sigma_i^2}\right\}$$

Now, the posterior is proportional to the product of the prior and the joint likelihood:

$$\begin{aligned} f_{\Theta|X}(\theta|x) &\propto f_\Theta(\theta) f_{X|\Theta}(x|\theta) \\ &\propto \exp\left\{-\frac{(\theta-x_0)^2}{2\sigma_0^2}\right\} \exp\left\{-\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2\sigma_i^2}\right\} \\ &= \exp\left\{-\frac{1}{2}\left(\frac{(\theta-x_0)^2}{\sigma_0^2} + \sum_{i=1}^{n}\frac{(x_i-\theta)^2}{\sigma_i^2}\right)\right\} \\ &= \exp\left\{-\frac{1}{2}\sum_{i=0}^{n}\frac{(x_i-\theta)^2}{\sigma_i^2}\right\} \end{aligned}$$

Let $quad(\theta) = \sum_{i=0}^{n}\frac{(x_i-\theta)^2}{\sigma_i^2}$. The posterior is $f_{\Theta|X}(\theta|x) \propto e^{-\frac{1}{2}quad(\theta)}$. Since $quad(\theta)$ is a sum of quadratic functions in θ, it is itself a quadratic function in θ. The coefficient of θ^2 in $quad(\theta)$ is $\sum_{i=0}^{n}\frac{1}{\sigma_i^2}$, which is positive. Therefore, the posterior distribution is normal.

To find the MAP/LMS estimate, we need to find the value of θ that minimizes $quad(\theta)$ (or equivalently, maximizes the posterior PDF). We can do this by taking the derivative with respect to θ and setting it to zero:

$$\frac{d}{d\theta}quad(\theta) = \frac{d}{d\theta}\sum_{i=0}^{n}\frac{(x_i-\theta)^2}{\sigma_i^2} = \sum_{i=0}^{n}\frac{-2(x_i-\theta)}{\sigma_i^2} = -2\sum_{i=0}^{n}\left(\frac{x_i}{\sigma_i^2} - \frac{\theta}{\sigma_i^2}\right)$$

Setting the derivative to zero:

$$\sum_{i=0}^{n}\frac{x_i}{\sigma_i^2} - \sum_{i=0}^{n}\frac{\theta}{\sigma_i^2} = 0$$

$$\sum_{i=0}^{n}\frac{x_i}{\sigma_i^2} = \theta\sum_{i=0}^{n}\frac{1}{\sigma_i^2}$$

Solving for θ gives the estimate:

$$\hat{\theta}_{MAP} = \hat{\theta}_{LMS} = E[\Theta|X=x] = \frac{\sum_{i=0}^{n}\frac{x_i}{\sigma_i^2}}{\sum_{i=0}^{n}\frac{1}{\sigma_i^2}}$$

This confirms that the MAP and LMS estimates coincide and yield this formula.

4.1 Interpretation of the Estimate

The estimate $\hat{\theta} = \frac{\sum_{i=0}^{n} x_i/\sigma_i^2}{\sum_{i=0}^{n} 1/\sigma_i^2}$ has a clear interpretation as a weighted average. We can rewrite it as:

$$\hat{\theta} = \sum_{i=0}^{n} w_i x_i, \quad \text{where } w_i = \frac{1/\sigma_i^2}{\sum_{j=0}^{n} 1/\sigma_j^2}$$

The weights w_i sum to 1 ($\sum_{i=0}^{n} w_i = 1$). The estimate is a weighted average of the prior mean x_0 and all the observations $x_1, ..., x_n$. The weight w_i given to x_i is inversely proportional to its associated variance σ_i^2 (where σ_0^2 is the prior variance and σ_i^2 for $i \geq 1$ is the noise variance for X_i). This makes intuitive sense: data points (including the prior mean) with smaller variance (i.e., higher precision or certainty) receive higher weight in the final estimate.

Key conclusions for the multiple observation case remain consistent with the single observation case under normality:

- The posterior distribution $f_{\Theta|X}(\theta|x)$ is normal.
- The LMS and MAP estimates coincide, given by the weighted average formula.
- The estimate is a linear function of the prior mean and the observations: $\hat{\theta} = a_0 x_0 + a_1 x_1 + \cdots + a_n x_n$.

5 The Mean Squared Error

A crucial aspect of estimation is evaluating the performance of an estimator. For the LMS estimator, the relevant performance measure is the Mean Squared Error (MSE). We consider both the conditional MSE given the observations, and the overall (unconditional) MSE.

The conditional MSE given $X = x$ is the variance of the posterior distribution:

$$\text{MSE}_{\text{cond}} = \mathbb{E}[(\Theta - \hat{\Theta})^2 | X = x] = \mathbb{E}[(\Theta - E[\Theta|X = x])^2 | X = x] = \text{var}(\Theta|X = x)$$

Since we found that the posterior $f_{\Theta|X}(\theta|x)$ is proportional to $e^{-\frac{1}{2} quad(\theta)}$, where $quad(\theta) = \sum_{i=0}^{n} \frac{(\theta - x_i)^2}{\sigma_i^2}$, we can identify the posterior variance. Recall that for a normal distribution $N(\mu, \sigma^2)$, the PDF is proportional to $e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}$. We need to rewrite $quad(\theta)$ in the form $\frac{(\theta-\mu_{\text{post}})^2}{\sigma_{\text{post}}^2} + \text{const}$. Expanding $quad(\theta)$:

$$quad(\theta) = \sum_{i=0}^{n} \frac{\theta^2 - 2x_i\theta + x_i^2}{\sigma_i^2} = \left(\sum_{i=0}^{n} \frac{1}{\sigma_i^2}\right)\theta^2 - 2\left(\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}\right)\theta + \left(\sum_{i=0}^{n} \frac{x_i^2}{\sigma_i^2}\right)$$

This is of the form $A\theta^2 + B\theta + C$. Comparing the exponent $-\frac{1}{2} quad(\theta)$ with the standard normal exponent $-\frac{(\theta-\mu_{\text{post}})^2}{2\sigma_{\text{post}}^2}$, we look at the coefficient of θ^2:

$$-\frac{1}{2}\left(\sum_{i=0}^{n} \frac{1}{\sigma_i^2}\right) = -\frac{1}{2\sigma_{\text{post}}^2}$$

This directly gives the posterior variance:

$$\sigma_{\text{post}}^2 = \text{var}(\Theta|X = x) = \frac{1}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

Therefore, the conditional MSE is:

$$\mathbb{E}[(\Theta - \hat{\Theta})^2 | X = x] = \frac{1}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

An important observation is that this conditional variance (and thus the conditional MSE) does *not* depend on the specific values observed $x = (x_1, ..., x_n)$. It depends only on the prior variance σ_0^2 and the noise variances σ_i^2.

The overall (unconditional) MSE is the expectation of the conditional MSE over all possible observations X:

$$\text{MSE}_{\text{overall}} = E[(\Theta - \hat{\Theta})^2] = E\left[\mathbb{E}[(\Theta - \hat{\Theta})^2|X]\right]$$

Since the conditional MSE is constant and does not depend on X, its expectation is just the constant itself:

$$E[(\Theta - \hat{\Theta})^2] = \frac{1}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

So, for linear models with normal noise, the conditional and unconditional MSE for the LMS/MAP estimator are identical.

5.1 Examples

- **Equal Variances:** If all variances are equal, $\sigma_0^2 = \sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2$, then the sum in the denominator has $n + 1$ identical terms:

$$\sum_{i=0}^{n} \frac{1}{\sigma_i^2} = \sum_{i=0}^{n} \frac{1}{\sigma^2} = \frac{n+1}{\sigma^2}$$

  The estimate becomes the simple average: $\hat{\theta} = \frac{\sum_{i=0}^{n} x_i}{n+1}$. The MSE (conditional and unconditional) is:

$$\text{MSE} = \frac{1}{(n+1)/\sigma^2} = \frac{\sigma^2}{n+1}$$

  As the number of observations n increases, the MSE decreases, reflecting increasing accuracy.

- **Single Observation Case Revisited:** Consider $X = \Theta + W$, with $\Theta \sim N(0, 1)$ and $W \sim N(0, 1)$ independent. Here, $n = 1$, $x_0 = 0$ (prior mean), $\sigma_0^2 = 1$, x_1 is the observation, $\sigma_1^2 = 1$. The estimate is $\hat{\Theta} = \frac{x_0/\sigma_0^2 + x_1/\sigma_1^2}{1/\sigma_0^2 + 1/\sigma_1^2} = \frac{0/1 + X/1}{1/1 + 1/1} = \frac{X}{2}$, which matches our previous result. The MSE is:

$$\text{MSE} = \mathbb{E}[(\Theta - \hat{\Theta})^2|X = x] = \frac{1}{1/\sigma_0^2 + 1/\sigma_1^2} = \frac{1}{1/1 + 1/1} = \frac{1}{2}$$

  The conditional MSE is constant, equal to $1/2$.

6 Multiple Parameters: Trajectory Estimation Example

The framework extends naturally to estimating multiple unknown parameters $\Theta = (\Theta_1, ..., \Theta_m)$. Consider estimating the parameters of a trajectory. Suppose the position $x(t)$ at time t follows a quadratic model:

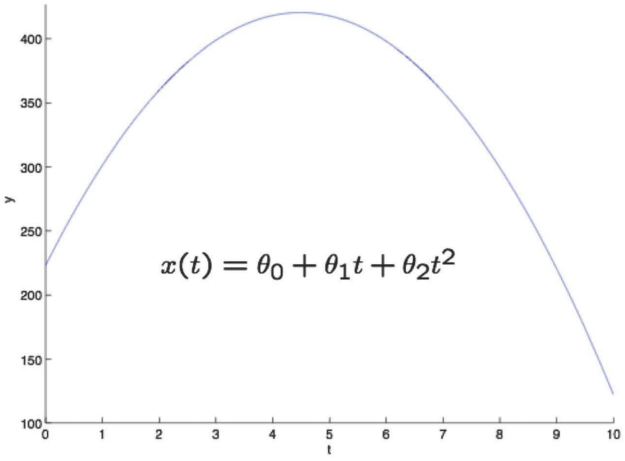
$$x(t) = \theta_0 + \theta_1 t + \theta_2 t^2$$

Here, $\theta_0, \theta_1, \theta_2$ are the unknown parameters determining the trajectory (e.g., initial position, initial velocity, acceleration/2).

We model these parameters as random variables $\Theta_0, \Theta_1, \Theta_2$. Let's assume they are independent with prior distributions $f_{\Theta_j}(\theta_j)$. We take measurements of the position at different times $t_1, ..., t_n$. Each measurement X_i is corrupted by noise W_i:

$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

We assume a noise model, typically that the W_i are independent and identically distributed (i.i.d.) random variables, independent of the Θ_j.



6.1 Model with Normality Assumptions

Let's make specific normality assumptions:

- Priors: $\Theta_j \sim N(0, \sigma_j^2)$ for $j = 0, 1, 2$. Assume independence. (Using zero mean priors is common if there's no strong prior belief, but non-zero means could also be used).
- Noise: $W_i \sim N(0, \sigma^2)$ for $i = 1, ..., n$. Assume independence, and independence from Θ_j.

We want to find the posterior PDF for the vector $\Theta = (\Theta_0, \Theta_1, \Theta_2)$ given the vector of observations $X = (X_1, ..., X_n)$. Using Bayes' rule:

$$f_{\Theta|X}(\theta|x) \propto f_\Theta(\theta) f_{X|\Theta}(x|\theta)$$

The joint prior PDF $f_\Theta(\theta)$ is the product of individual priors due to independence:

$$f_\Theta(\theta) = f_{\Theta_0}(\theta_0) f_{\Theta_1}(\theta_1) f_{\Theta_2}(\theta_2) \propto \exp\left\{-\frac{\theta_0^2}{2\sigma_0^2}\right\} \exp\left\{-\frac{\theta_1^2}{2\sigma_1^2}\right\} \exp\left\{-\frac{\theta_2^2}{2\sigma_2^2}\right\}$$

$$f_\Theta(\theta) \propto \exp\left\{-\frac{1}{2}\left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2}\right)\right\}$$

The joint likelihood $f_{X|\Theta}(x|\theta)$ is found using the conditional independence of X_i given $\Theta = \theta$. Given $\theta = (\theta_0, \theta_1, \theta_2)$, each observation is $X_i = (\theta_0 + \theta_1 t_i + \theta_2 t_i^2) + W_i$. Since $W_i \sim N(0, \sigma^2)$, we have $X_i | \{\Theta = \theta\} \sim N(\theta_0 + \theta_1 t_i + \theta_2 t_i^2, \sigma^2)$. The individual likelihood is:

$$f_{X_i|\Theta}(x_i|\theta) \propto \exp\left\{-\frac{(x_i - (\theta_0 + \theta_1 t_i + \theta_2 t_i^2))^2}{2\sigma^2}\right\}$$

The joint likelihood is the product:

$$f_{X|\Theta}(x|\theta) = \prod_{i=1}^{n} f_{X_i|\Theta}(x_i|\theta) \propto \prod_{i=1}^{n} \exp\left\{-\frac{(x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2}{2\sigma^2}\right\}$$

$$f_{X|\Theta}(x|\theta) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2\right\}$$

Combining the prior and the likelihood, the posterior PDF is:

$$f_{\Theta|X}(\theta|x) \propto \exp\left\{-\frac{1}{2}\left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2\right\}$$

Let this be written as $f_{\Theta|X}(\theta|x) \propto c(x)\exp\{-\frac{1}{2}Q(\theta_0, \theta_1, \theta_2)\}$, where $c(x)$ collects terms not depending on θ, and

$$Q(\theta_0, \theta_1, \theta_2) = \left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2}\right) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2$$

The expression $Q(\theta_0, \theta_1, \theta_2)$ is a quadratic function of the parameters $\theta_0, \theta_1, \theta_2$. This implies that the joint posterior distribution $f_{\Theta|X}(\theta|x)$ is a multivariate normal distribution.

The MAP estimate $\hat{\theta}_{MAP} = (\hat{\theta}_{0,MAP}, \hat{\theta}_{1,MAP}, \hat{\theta}_{2,MAP})$ is found by maximizing the posterior PDF, which is equivalent to minimizing the quadratic function $Q(\theta_0, \theta_1, \theta_2)$ with respect to $\theta_0, \theta_1, \theta_2$. This minimization can be performed by setting the partial derivatives $\partial Q/\partial\theta_j$ to zero for $j = 0, 1, 2$. This results in a system of linear equations for $\theta_0, \theta_1, \theta_2$.

7 General Linear Normal Models

The trajectory example is a specific instance of a broader class of linear normal models. In general, if the parameters $\Theta = (\Theta_1, ..., \Theta_m)$ and observations $X = (X_1, ..., X_n)$ are such that they can all be expressed as linear functions of some underlying independent normal random variables (which includes the priors for Θ_j and the noise terms W_i), then several key properties hold:

- The joint posterior distribution $f_{\Theta|X}(\theta|x)$ is a multivariate normal distribution. Its PDF is proportional to $\exp\{-Quadratic(\theta_1, ..., \theta_m)\}$.
- The MAP estimate $\hat{\Theta}_{MAP}$ is found by minimizing this quadratic function. This leads to a system of linear equations.
- The MAP estimate for each parameter $\hat{\Theta}_{MAP,j}$ is a linear function of the observations $X = (X_1, ..., X_n)$.
- **MAP = LMS:** In the multivariate normal case, the mode (MAP estimate) coincides with the mean (LMS estimate). Thus, $\hat{\Theta}_{MAP,j} = E[\Theta_j|X]$.
- **Marginal Posteriors:** The marginal posterior PDF for each individual parameter, $f_{\Theta_j|X}(\theta_j|x)$, is also normal.
- **Joint vs. Marginal MAP:** The MAP estimate for Θ_j obtained from the joint posterior ($\hat{\Theta}_{MAP,j}$) is the same as the MAP estimate obtained from the marginal posterior $f_{\Theta_j|X}(\theta_j|x)$. This is a property of multivariate normal distributions where the mode of the joint distribution projects onto the modes of the marginal distributions.

- **Constant Conditional MSE:** The conditional MSE for estimating Θ_i, given $X = x$, which is the variance of the marginal posterior distribution $\text{var}(\Theta_i|X = x)$, does not depend on the observed values x. Consequently, the overall MSE $E[(\hat{\Theta}_{i,MAP} - \Theta_i)^2]$ is equal to this constant conditional MSE.

8 Illustration: Free-Falling Object Trajectory

Let's consider a specific numerical example of trajectory estimation. Suppose an object is falling under gravity, so its vertical position follows $x(t) = \Theta_0 + \Theta_1 t + \Theta_2 t^2$. We know the acceleration due to gravity, so $\Theta_2 = -g/2 \approx -9.81/2 = -4.905$. Let's assume Θ_2 is known and constant for simplicity in this version (although the original slide notation varies slightly across pages). We want to estimate the initial position Θ_0 and initial velocity Θ_1.

Assume the following priors and noise model:

- $\Theta_0 \sim N(\mu_0, \sigma_0^2)$, e.g., $\mu_0 = 200, \sigma_0^2 = 50^2$
- $\Theta_1 \sim N(\mu_1, \sigma_1^2)$, e.g., $\mu_1 = 50, \sigma_1^2 = 50^2$
- $\Theta_2 = -9.81/2 = -4.905$ (Known constant)
- Noise $W_i \sim N(0, \sigma^2)$, e.g., $\sigma^2 = 50^2$

Measurements X_i are taken at times t_i:

$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

The posterior PDF for (Θ_0, Θ_1) given $X = x$ is proportional to:

$$f_{\Theta_0,\Theta_1|X}(\theta_0, \theta_1|x) \propto f_{\Theta_0}(\theta_0) f_{\Theta_1}(\theta_1) \prod_{i=1}^{n} f_{X_i|\Theta_0,\Theta_1}(x_i|\theta_0, \theta_1)$$

$$\propto \exp\left\{-\frac{(\theta_0 - \mu_0)^2}{2\sigma_0^2}\right\} \exp\left\{-\frac{(\theta_1 - \mu_1)^2}{2\sigma_1^2}\right\} \exp\left\{-\sum_{i=1}^{n} \frac{(x_i - \theta_0 - \theta_1 t_i - \Theta_2 t_i^2)^2}{2\sigma^2}\right\}$$

The MAP estimate $(\hat{\theta}_{0,MAP}, \hat{\theta}_{1,MAP})$ is found by minimizing the negative logarithm of this expression (ignoring constants), which means minimizing:

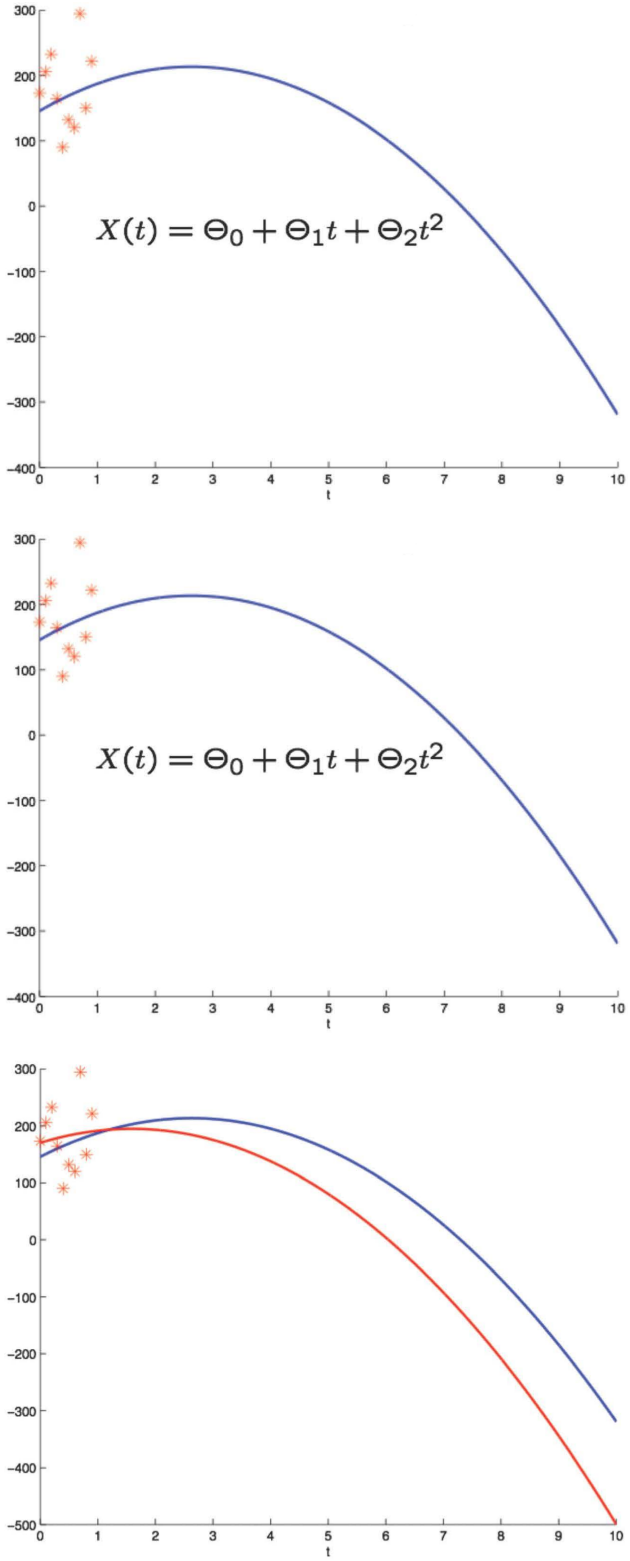
$$\frac{(\theta_0 - \mu_0)^2}{2\sigma_0^2} + \frac{(\theta_1 - \mu_1)^2}{2\sigma_1^2} + \sum_{i=1}^{n} \frac{(x_i - \theta_0 - \theta_1 t_i - \Theta_2 t_i^2)^2}{2\sigma^2}$$

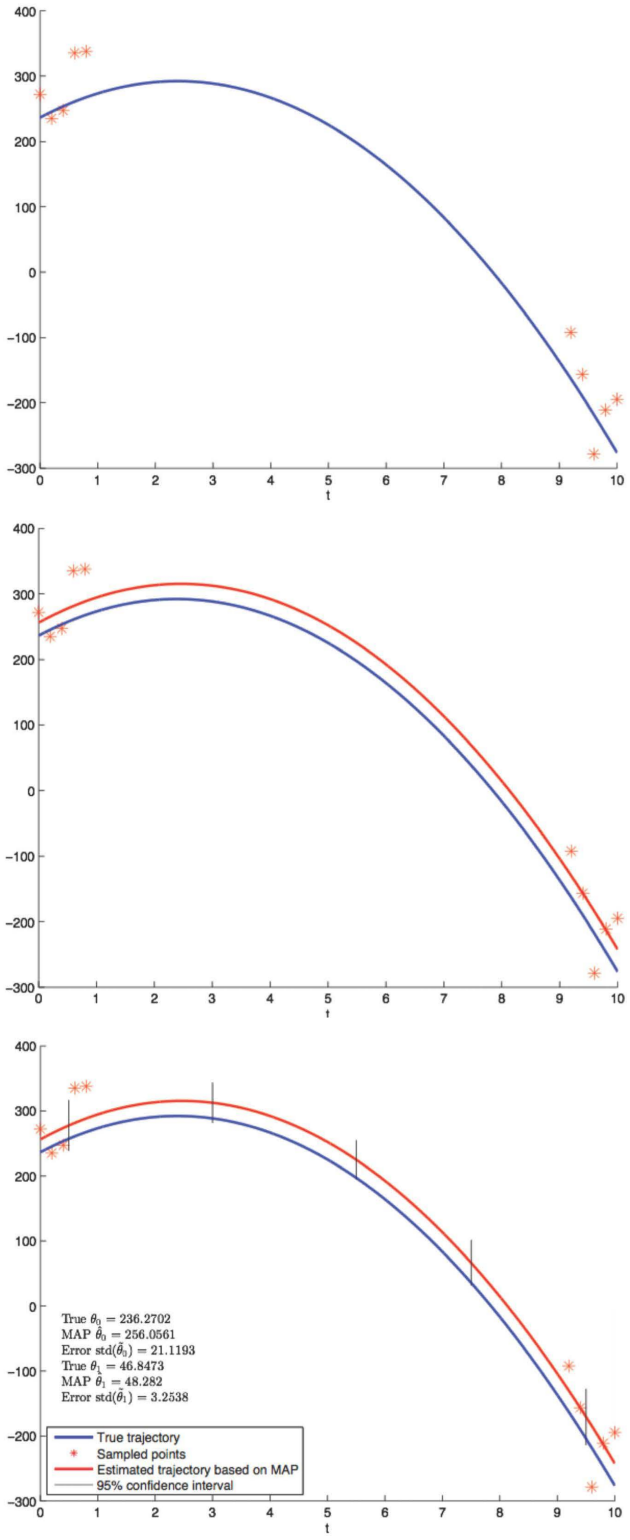
Using the example values ($\mu_0 = 200, \mu_1 = 50, \sigma_0^2 = \sigma_1^2 = \sigma^2 = 50^2, \Theta_2 = -4.905 \approx -9.81/2$. Note: slide 15 uses $\Theta_2 = -9.81$ directly in the objective, suggesting it might be simplifying $g = 9.81$ and $\Theta_2 = -g$. Let's stick to the objective shown on slide 15, which seems to absorb the $1/(2\sigma^2)$ scaling): Minimize w.r.t. θ_0, θ_1:

$$(\theta_0 - 200)^2 + (\theta_1 - 50)^2 + \sum_{i=1}^{n} (x_i - \theta_0 - \theta_1 t_i + (9.81/2) t_i^2)^2$$

(Note: The slides use $+9.81 t_i^2$ inside the square, which likely assumes x_i measurements are adjusted for the known gravity term, or there's a sign convention difference. Assuming the formula on slide 15 is the intended objective based on their priors and measurements).

Solving this minimization problem (a standard least squares problem with regularization from the prior) yields the MAP estimates $\hat{\theta}_0$ and $\hat{\theta}_1$. The figures illustrate this process: noisy measurements are generated from a true trajectory, and the MAP estimation procedure finds an estimated trajectory that balances fitting the data and adhering to the prior beliefs.





The final slide in the illustration shows numerical results for a specific realization:

- True $\theta_0 = 236.2702$
- MAP estimate $\hat{\theta}_0 = 256.0561$

- Standard deviation of the estimator $\hat{\Theta}_0$ (square root of posterior variance $\text{var}(\Theta_0|X = x)$): $std(\hat{\theta}_0) = 21.1193$
- True $\theta_1 = 46.8473$
- MAP estimate $\hat{\theta}_1 = 48.282$
- Standard deviation of the estimator $\hat{\Theta}_1$: $std(\hat{\theta}_1) = 3.2538$

These standard deviations quantify the uncertainty in the estimates based on the posterior distribution. The "95% confidence interval" shown in the plot is likely derived from these posterior standard deviations, typically as $\hat{\theta}_j \pm 1.96 \times std(\hat{\theta}_j)$. The plot visually demonstrates how the estimated trajectory, derived from the MAP estimates, fits the noisy data while being influenced by the prior expectations.

Lecture 18: Least Mean Squares (LMS) Estimation Estimation

Instructor: Prof. Abolfazl Hashemi

1 Introduction to Least Mean Squares (LMS) Estimation

This lecture introduces the Least Mean Squares (LMS) estimation criterion, a fundamental approach for determining point estimates of unknown parameters. Unlike the Maximum A Posteriori (MAP) method, which seeks the most probable value of the parameter, LMS estimation aims to find an estimate that minimizes the average squared error.

Specifically, given an observation $X = x$, we want to find an estimate $\hat{\theta}$ for the unknown parameter Θ that minimizes the conditional Mean Squared Error (MSE):

$$\text{MSE}_{\text{cond}}(\hat{\theta}|x) = E[(\Theta - \hat{\theta})^2|X = x]$$

As we will show, the solution that minimizes this conditional MSE is the conditional expectation of the parameter given the observation:

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

This provides a powerful and widely applicable method for generating estimators. We will delve into its mathematical properties, compare it with MAP estimation, and work through an illustrative example.

2 LMS Estimation Without Observations

Let's first consider the simplest scenario: estimating an unknown parameter Θ when we have no observations, only a prior distribution $f_\Theta(\theta)$ (or $p_\Theta(\theta)$ for discrete Θ). We want to choose a single numerical value $\hat{\theta}$ as our best guess for Θ.

How should we choose this $\hat{\theta}$?

- One approach is the MAP rule: choose $\hat{\theta}$ that maximizes the prior $f_\Theta(\theta)$. This gives the most likely value before seeing any data.
- Another approach relates to minimizing the error. The LMS criterion in this context is to choose $\hat{\theta}$ to minimize the overall Mean Squared Error (MSE), which is calculated based solely on the prior distribution:

  $$\text{MSE}(\hat{\theta}) = E[(\Theta - \hat{\theta})^2] = \int_{-\infty}^{\infty} (\theta - \hat{\theta})^2 f_\Theta(\theta) d\theta$$

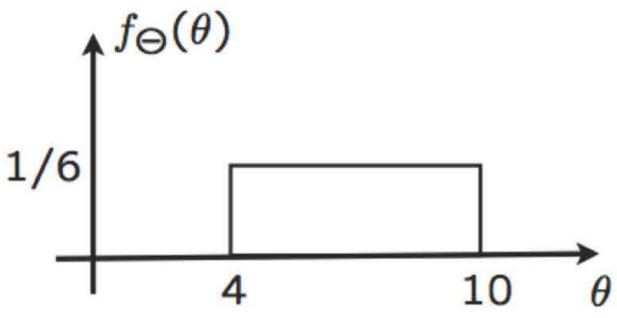
  (or a sum if Θ is discrete).

2.1 Example: Uniform Prior

Suppose Θ is uniformly distributed between 4 and 10: $\Theta \sim U[4, 10]$.

$$f_\Theta(\theta) = \begin{cases} 1/6 & \text{if } 4 \leq \theta \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

For this uniform prior:



- The MAP estimate could be any value between 4 and 10, as they all have the maximum probability density.
- The LMS estimate seeks to minimize $E[(\Theta - \hat{\theta})^2]$.

2.2 Derivation of the LMS Estimate (No Observations)

To find the value $\hat{\theta}$ that minimizes $g(\hat{\theta}) = E[(\Theta - \hat{\theta})^2]$, we can differentiate $g(\hat{\theta})$ with respect to $\hat{\theta}$ and set the derivative to zero. We can differentiate inside the expectation (assuming mild conditions):

$$\begin{aligned}
\frac{d}{d\hat{\theta}} E[(\Theta - \hat{\theta})^2] &= E\left[\frac{d}{d\hat{\theta}}(\Theta - \hat{\theta})^2\right] \\
&= E[2(\Theta - \hat{\theta}) \cdot (-1)] \\
&= -2E[\Theta - \hat{\theta}] \\
&= -2(E[\Theta] - E[\hat{\theta}]) \\
&= -2(E[\Theta] - \hat{\theta}) \quad \text{(since } \hat{\theta} \text{ is a constant here)}
\end{aligned}$$

Setting the derivative to zero:

$$-2(E[\Theta] - \hat{\theta}) = 0 \implies \hat{\theta} = E[\Theta]$$

The second derivative is $\frac{d^2}{d\hat{\theta}^2} g(\hat{\theta}) = \frac{d}{d\hat{\theta}}[-2(E[\Theta] - \hat{\theta})] = 2 > 0$, confirming this is a minimum.

Therefore, the LMS estimate in the absence of observations is the mean (expected value) of the prior distribution.

For the uniform example $\Theta \sim U[4, 10]$, the mean is $E[\Theta] = (4 + 10)/2 = 7$. So, $\hat{\theta}_{\text{LMS}} = 7$.

2.3 Optimal Mean Squared Error (No Observations)

The minimum achievable MSE occurs when we use the LMS estimate $\hat{\theta} = E[\Theta]$. The value of this minimum MSE is:

$$\min_{\hat{\theta}} E[(\Theta - \hat{\theta})^2] = E[(\Theta - E[\Theta])^2]$$

This expression is precisely the definition of the variance of Θ.

$$\text{Optimal MSE} = \text{var}(\Theta)$$

For the uniform example $\Theta \sim U[4, 10]$, the variance is $\text{var}(\Theta) = \frac{(b-a)^2}{12} = \frac{(10-4)^2}{12} = \frac{6^2}{12} = \frac{36}{12} = 3$. The minimum MSE achievable with a constant estimate is 3.

3 LMS Estimation Based on an Observation X

Now, let's incorporate an observation X. We have:

- An unknown parameter Θ with prior $f_\Theta(\theta)$.
- An observation X related to Θ via a likelihood model $f_{X|\Theta}(x|\theta)$.
- A specific observed value $X = x$.

We seek a point estimate $\hat{\theta}$ for Θ using the information provided by x.

The LMS principle is now applied within the conditional probability space, given $X = x$. We choose $\hat{\theta}$ to minimize the conditional Mean Squared Error:

$$\min_{\hat{\theta}} E[(\Theta - \hat{\theta})^2 | X = x]$$

This expectation is calculated using the posterior distribution $f_{\Theta|X}(\theta|x)$. The argument used previously for the unconditional case applies directly here: the value $\hat{\theta}$ that minimizes $E[(\Theta - \hat{\theta})^2 | X = x]$ is the mean of the distribution used in the expectation, which is the posterior distribution.

Therefore, the LMS estimate of Θ given $X = x$ is the conditional expectation:

$$\hat{\theta}_{\text{LMS}} = E[\Theta | X = x] = \int_{-\infty}^{\infty} \theta f_{\Theta|X}(\theta|x) d\theta$$

The LMS estimator is the function $\hat{\Theta}_{\text{LMS}} = g(X)$ that maps any possible observation X to the corresponding conditional expectation:

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

3.1 Optimality Properties

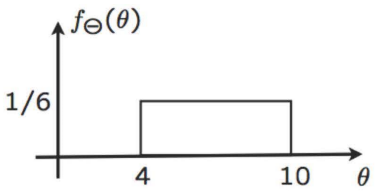
The LMS estimator $E[\Theta|X]$ possesses strong optimality properties:

1. $E[\Theta]$ minimizes $E[(\Theta - c)^2]$ over all constants c.
2. $E[\Theta|X = x]$ minimizes $E[(\Theta - c)^2 | X = x]$ over all constants c, for a fixed observation x.
3. $\hat{\Theta}_{\text{LMS}} = E[\Theta|X]$ minimizes the overall MSE $E[(\Theta - g(X))^2]$ over all possible estimators $g(X)$ (i.e., all functions of the data).

Property 3 is particularly significant. It states that among all conceivable ways to use the observation X to estimate Θ, the conditional expectation $E[\Theta|X]$ yields the lowest average squared error. No other function of X can perform better in this overall MSE sense.

4 Performance of the LMS Estimator

We can evaluate the performance of the LMS estimator both conditionally (after observing x) and unconditionally (before observing X).



- **Conditional MSE:** Given that we observed $X = x$, the expected squared error of our estimate $\hat{\theta} = E[\Theta|X = x]$ is:

$$\text{MSE}_{\text{cond}}(x) = E[(\Theta - E[\Theta|X = x])^2|X = x]$$

This is simply the variance of the posterior distribution:

$$\text{MSE}_{\text{cond}}(x) = \text{var}(\Theta|X = x)$$

This value generally depends on x. It quantifies the remaining uncertainty about Θ after observing x.

- **Overall MSE:** The overall expected performance of the estimator $\hat{\Theta} = E[\Theta|X]$ before any observation is made is the expectation of the conditional MSE over all possible X:

$$\text{MSE}_{\text{overall}} = E[(\Theta - E[\Theta|X])^2] = E[\text{var}(\Theta|X)]$$

This is also known as the minimum overall MSE achievable by any estimator based on X.

5 Comparison with MAP Estimation

The LMS estimator $E[\Theta|X = x]$ minimizes the expected squared error, while the MAP estimator maximizes the posterior PDF $f_{\Theta|X}(\theta|x)$. These are distinct criteria and may yield different estimates.

However, they coincide under certain conditions:

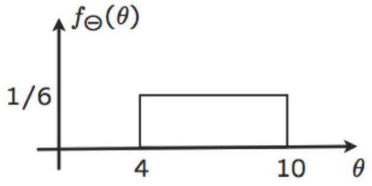
- If the posterior distribution $f_{\Theta|X}(\theta|x)$ is unimodal (single peak) and symmetric around its mean.
- A prominent example is when the posterior is a normal distribution, as its mean, median, and mode are all identical.
- As established in the previous lecture, linear-normal models (linear relationship between X and Θ, normal prior, normal noise) result in normal posteriors. In such cases, $\hat{\theta}_{\text{LMS}} = \hat{\theta}_{\text{MAP}}$.

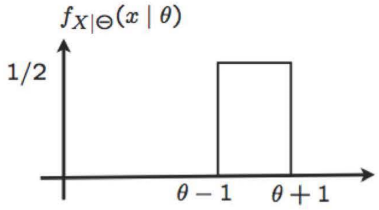
LMS is primarily used for estimation problems where the cost of error grows quadratically. It's less directly applicable to hypothesis testing, where MAP (or likelihood ratios) are more common.

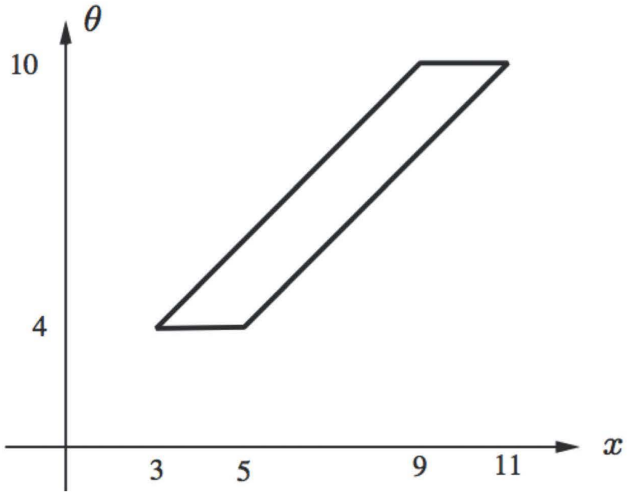
6 Example: Uniform Prior and Uniform Noise Revisited

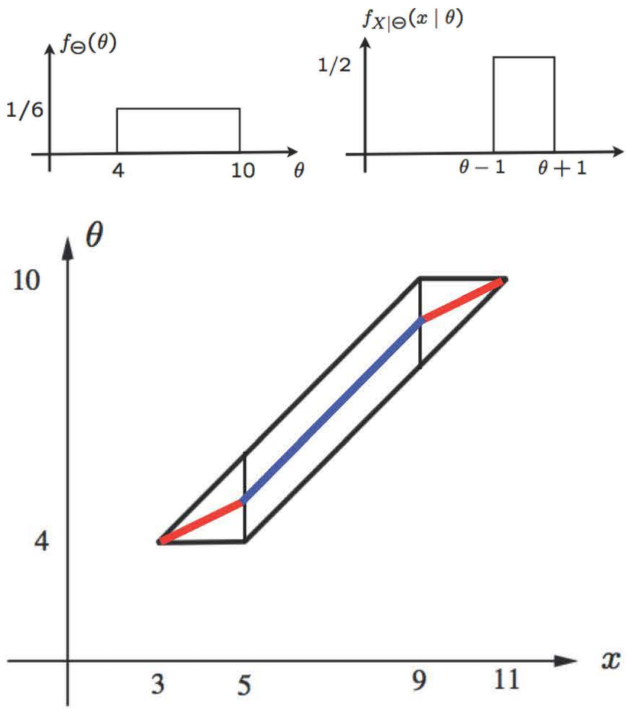
Let's re-examine the example with $\Theta \sim U[4, 10]$ and $X|\{\Theta = \theta\} \sim U[\theta - 1, \theta + 1]$.

We previously found the posterior distribution $f_{\Theta|X}(\theta|x)$ to be uniform over different intervals depending on x:


$f_\Theta(\theta)$
1/6
4
10
θ


$f_{X|\Theta}(x \mid \theta)$
1/2
$\theta - 1$
$\theta + 1$


θ
10
4
3
5
9
11
x



- If $3 < x < 5$: $f_{\Theta|X}(\theta|x) \sim U[4, x+1]$.
- If $5 \le x \le 9$: $f_{\Theta|X}(\theta|x) \sim U[x-1, x+1]$.
- If $9 < x < 11$: $f_{\Theta|X}(\theta|x) \sim U[x-1, 10]$.

The LMS estimate $\hat{\theta}_{\text{LMS}}(x) = E[\Theta|X = x]$ is the midpoint of the interval over which the posterior is uniform:

$$\hat{\theta}_{\text{LMS}}(x) = E[\Theta|X = x] = \begin{cases} (4 + (x+1))/2 = (x+5)/2 & \text{if } 3 < x < 5 \\ ((x-1) + (x+1))/2 = x & \text{if } 5 \le x \le 9 \\ ((x-1) + 10)/2 = (x+9)/2 & \text{if } 9 < x < 11 \end{cases}$$

This estimator is a continuous, piecewise linear function of x.

6.1 Conditional MSE for the Example

The conditional MSE is $\text{var}(\Theta|X = x)$. For a $U[a, b]$ distribution, the variance is $(b-a)^2/12$.

- If $3 < x < 5$: Interval is $[4, x+1]$. Length $b - a = x - 3$. $\text{var}(\Theta|X = x) = \frac{(x-3)^2}{12}$.
- If $5 \le x \le 9$: Interval is $[x-1, x+1]$. Length $b - a = 2$. $\text{var}(\Theta|X = x) = \frac{2^2}{12} = \frac{1}{3}$.
- If $9 < x < 11$: Interval is $[x-1, 10]$. Length $b - a = 11 - x$. $\text{var}(\Theta|X = x) = \frac{(11-x)^2}{12}$.

The conditional MSE is smallest (equal to $1/3$) when x falls in the central region $[5, 9]$, and increases quadratically as x approaches the boundaries 3 or 11. This indicates that observations in the middle range provide more certainty about Θ compared to observations near the edges.

7 LMS Estimation with Multiple Variables

The LMS framework extends easily to multiple observations or multiple unknown parameters.

- **Multiple Observations:** If we have observations $X = (X_1, ..., X_n)$, the LMS estimate for a scalar Θ is:
$$\hat{\theta}_{\text{LMS}} = E[\Theta | X_1 = x_1, ..., X_n = x_n]$$
This requires finding the posterior distribution $f_{\Theta|X_1,...,X_n}(\theta|x_1, ..., x_n)$ and calculating its mean.
- **Multiple Unknowns (Vector Parameter):** If $\Theta = (\Theta_1, ..., \Theta_m)$ is a vector, the LMS estimate $\hat{\Theta} = (\hat{\Theta}_1, ..., \hat{\Theta}_m)$ minimizes the expected squared Euclidean distance $E[||\Theta - \hat{\Theta}||^2 | X = x]$. This minimization is achieved by estimating each component separately using its conditional expectation:
$$\hat{\Theta}_{j,\text{LMS}} = E[\Theta_j | X = x] \quad \text{for } j = 1, ..., m$$
This requires finding the marginal posterior distributions for each Θ_j (given $X = x$) and calculating their means. In many cases, it's easier to find the mean of the joint posterior distribution $f_{\Theta|X}(\theta|x)$.

8 Properties of the Estimation Error

Let $\hat{\Theta} = E[\Theta|X]$ be the LMS estimator and $\tilde{\Theta} = \Theta - \hat{\Theta}$ be the estimation error.

- **Conditional Mean of Error is Zero:** As shown before, $E[\tilde{\Theta}|X = x] = 0$ for any x. This implies $E[\tilde{\Theta}|X] = 0$. By iterated expectations, the overall mean error is also zero: $E[\tilde{\Theta}] = E[E[\tilde{\Theta}|X]] = E[0] = 0$. The LMS estimator is unbiased conditionally and unconditionally.
- **Orthogonality Principle:** The error $\tilde{\Theta}$ is orthogonal (uncorrelated) to any function $h(X)$ of the data, including the estimator $\hat{\Theta}$ itself.
$$E[\tilde{\Theta} h(X)] = 0$$
Specifically, $E[(\Theta - E[\Theta|X])E[\Theta|X]] = 0$. Since both $\tilde{\Theta}$ and $\hat{\Theta}$ have zero mean (assuming $E[\Theta] = 0$ for simplicity, otherwise center them), this implies:
$$\text{cov}(\tilde{\Theta}, \hat{\Theta}) = 0$$
The error is uncorrelated with the estimate.
- **Variance Decomposition:** Using the orthogonality $E[\tilde{\Theta}\hat{\Theta}] = 0$ and $\tilde{\Theta} = \Theta - \hat{\Theta} \implies \Theta = \hat{\Theta} + \tilde{\Theta}$:
$$\begin{aligned}
\text{var}(\Theta) &= E[(\Theta - E[\Theta])^2] \\
&= E[(\hat{\Theta} + \tilde{\Theta} - E[\hat{\Theta} + \tilde{\Theta}])^2] \\
&= E[(\hat{\Theta} - E[\hat{\Theta}] + \tilde{\Theta} - E[\tilde{\Theta}])^2] \\
&= E[(\hat{\Theta} - E[\hat{\Theta}] + \tilde{\Theta})^2] \quad (\text{since } E[\tilde{\Theta}] = 0) \\
&= E[(\hat{\Theta} - E[\hat{\Theta}])^2 + \tilde{\Theta}^2 + 2\tilde{\Theta}(\hat{\Theta} - E[\hat{\Theta}])] \\
&= E[(\hat{\Theta} - E[\hat{\Theta}])^2] + E[\tilde{\Theta}^2] + 2E[\tilde{\Theta}\hat{\Theta}] - 2E[\tilde{\Theta}]E[\hat{\Theta}] \\
&= \text{var}(\hat{\Theta}) + E[\tilde{\Theta}^2] + 2(0) - 2(0)E[\hat{\Theta}] \\
&= \text{var}(\hat{\Theta}) + \text{var}(\tilde{\Theta}) \quad (\text{since } E[\tilde{\Theta}] = 0, E[\tilde{\Theta}^2] = \text{var}(\tilde{\Theta}))
\end{aligned}$$

So, $\text{var}(\Theta) = \text{var}(\hat{\Theta}_{\text{LMS}}) + \text{var}(\tilde{\Theta})$. This is also equivalent to the Law of Total Variance: $\text{var}(\Theta) = \text{var}(E[\Theta|X]) + E[\text{var}(\Theta|X)]$. The variance of the original variable is decomposed into the variance of the estimator plus the variance of the error (which is also the overall MSE).

Lecture 19: Linear Least Mean Squares (LLMS) Estimation

Instructor: Prof. Abolfazl Hashemi

1 Introduction

In the previous lecture, we established that the Least Mean Squares (LMS) estimator, given by the conditional expectation $\hat{\Theta}_{LMS} = E[\Theta|X]$, minimizes the Mean Squared Error (MSE) $E[(\Theta - g(X))^2]$ over all possible estimators $g(X)$. However, computing this conditional expectation can be challenging for several reasons.

1.1 Challenges in LMS Estimation

While theoretically optimal in the mean-square sense, LMS estimation faces practical difficulties:

1. **Model Dependency:** The result relies heavily on the accuracy of the assumed prior $f_\Theta(\theta)$ and likelihood $f_{X|\Theta}(x|\theta)$. If the models are incorrect, the resulting estimator may be far from optimal.
2. **Computational Burden:**
   - Finding the posterior $f_{\Theta|X}(\theta|x)$ involves calculating the evidence $f_X(x) = \int f_\Theta(\theta')f_{X|\Theta}(x|\theta')d\theta'$, which can be a difficult integral.
   - Calculating the conditional mean $E[\Theta|X = x] = \int \theta f_{\Theta|X}(\theta|x)d\theta$ requires another, potentially complex, integration.
   - These integrals often lack closed-form solutions, necessitating numerical methods (like Markov Chain Monte Carlo - MCMC) especially in high dimensions.
3. **Complexity of the Estimator:** The resulting estimator $\hat{\Theta} = E[\Theta|X]$ can be a highly nonlinear and complex function of the observations X, making its analysis (e.g., finding its distribution or moments) challenging. Linear-normal models are a notable exception where the estimator remains linear.

1.2 Addressing the Challenges

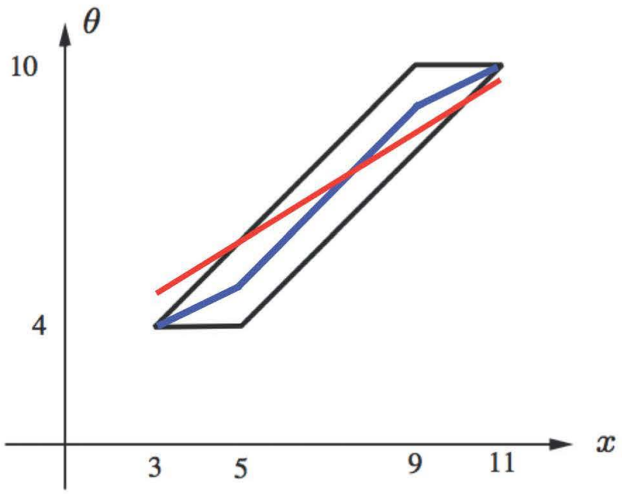
To address these difficulties, we introduce the Linear Least Mean Squares (LLMS) estimation framework. Instead of searching for the best estimator among *all* functions $g(X)$, we restrict our search to a simpler class: linear estimators of the form $\hat{\Theta} = aX + b$.

The goal of LLMS estimation is to find the scalar coefficients a and b that minimize the MSE specifically within this class of linear estimators:

$$\min_{a,b} E[(\Theta - (aX + b))^2]$$

The key advantages of this approach are:

- The solution for a and b is simple and depends only on first and second moments (means, variances, covariances) of Θ and X. Full distributional knowledge is not required.
- The resulting estimator is, by construction, easy to implement and analyze.



We will derive the solution, examine its properties, and compare it to the general LMS estimator through an example.

2 The LLMS Formulation

We are given an unknown random variable Θ and an observation random variable X. Our objective is to estimate Θ using an estimator that is a linear function of X.

Comparison of Estimator Classes:

- **General Estimators:** $\hat{\Theta} = g(X)$ for any function g. The optimal estimator in terms of minimizing overall MSE $E[(\Theta - g(X))^2]$ is $\hat{\Theta}_{LMS} = E[\Theta|X]$.
- **Linear Estimators:** $\hat{\Theta} = aX + b$. The optimal estimator in this restricted class, minimizing $E[(\Theta - (aX + b))^2]$, is denoted $\hat{\Theta}_{LLMS}$ or $\hat{\Theta}_L$.

The LLMS problem is to find the specific values of a and b that achieve this minimum.

The figure illustrates the concept using the example from the previous lecture. The blue curve represents the potentially nonlinear LMS estimator $E[\Theta|X]$. The red line represents the best linear approximation to this curve, which is the LLMS estimator $\hat{\Theta}_L = aX + b$. The LLMS estimator provides the best linear fit to Θ in the mean-square sense.

Important Note: If the true conditional expectation $E[\Theta|X]$ happens to be a linear function of X, then the best overall estimator is already linear. In this case, the LLMS estimator will coincide with the LMS estimator: $\hat{\Theta}_{LLMS} = \hat{\Theta}_{LMS}$. This occurs, for instance, in the linear-normal models discussed previously.

3 Derivation of the LLMS Estimator

We want to minimize the cost function $J(a, b) = E[(\Theta - aX - b)^2]$ with respect to a and b. We use calculus by setting the partial derivatives to zero.

Step 1: Minimize with respect to b (assuming a is fixed)

$$\frac{\partial J}{\partial b} = \frac{\partial}{\partial b} E[(\Theta - aX - b)^2] = E\left[\frac{\partial}{\partial b}(\Theta - aX - b)^2\right]$$
$$= E[2(\Theta - aX - b) \cdot (-1)] = -2E[\Theta - aX - b]$$

Setting the derivative to zero:

$$E[\Theta - aX - b] = 0 \implies E[\Theta] - aE[X] - b = 0$$
$$b^* = E[\Theta] - aE[X]$$

This shows that for any given slope a, the optimal intercept b^* ensures that the estimator $aX + b^*$ has the same mean as Θ, i.e., $E[aX + b^*] = aE[X] + (E[\Theta] - aE[X]) = E[\Theta]$. This means the LLMS estimator is unbiased.

Step 2: Substitute b^* and minimize with respect to a Substitute $b^* = E[\Theta] - aE[X]$ back into the cost function:

$$J(a, b^*) = E[(\Theta - aX - (E[\Theta] - aE[X]))^2]$$
$$= E[((\Theta - E[\Theta]) - a(X - E[X]))^2]$$

Now minimize this with respect to a:

$$\frac{\partial J(a, b^*)}{\partial a} = E\left[\frac{\partial}{\partial a}((\Theta - E[\Theta]) - a(X - E[X]))^2\right]$$
$$= E[2((\Theta - E[\Theta]) - a(X - E[X])) \cdot (-(X - E[X]))]$$
$$= -2E[(X - E[X])(\Theta - E[\Theta]) - a(X - E[X])^2]$$

Setting the derivative to zero:

$$E[(X - E[X])(\Theta - E[\Theta])] - a^* E[(X - E[X])^2] = 0$$

Recognizing the definitions of covariance and variance:

$$\mathrm{Cov}(\Theta, X) - a^* \mathrm{var}(X) = 0$$

Solving for the optimal slope a^*:

$$a^* = \frac{\mathrm{Cov}(\Theta, X)}{\mathrm{var}(X)}$$

(Assuming $\mathrm{var}(X) > 0$).

Step 3: Combine a^* and b^* The optimal linear estimator $\hat{\Theta}_L = a^* X + b^*$ is:

$$\hat{\Theta}_L = \frac{\mathrm{Cov}(\Theta, X)}{\mathrm{var}(X)} X + \left(E[\Theta] - \frac{\mathrm{Cov}(\Theta, X)}{\mathrm{var}(X)} E[X]\right)$$

Rearranging gives the standard form:

$$\hat{\Theta}_L = E[\Theta] + \frac{\mathrm{Cov}(\Theta, X)}{\mathrm{var}(X)}(X - E[X])$$

Using the correlation coefficient $\rho_{\Theta X} = \frac{\mathrm{Cov}(\Theta, X)}{\sigma_\Theta \sigma_X}$ and variances $\sigma_\Theta^2 = \mathrm{var}(\Theta)$, $\sigma_X^2 = \mathrm{var}(X)$:

$$\hat{\Theta}_L = E[\Theta] + \rho_{\Theta X} \frac{\sigma_\Theta}{\sigma_X}(X - E[X])$$

4 Properties of the LLMS Solution

The LLMS estimator $\hat{\Theta}_L = E[\Theta] + \frac{\text{Cov}(\Theta,X)}{\text{var}(X)}(X - E[X])$ has several important properties:

- **Depends only on Moments:** Its calculation requires only the means $E[\Theta], E[X]$, variances $\text{var}(\Theta), \text{var}(X)$, and the covariance $\text{Cov}(\Theta, X)$. The full probability distributions are not needed. This makes it practical even when distributions are unknown or complex, as these moments can often be estimated from data.
- **Interpretation:** The estimate starts at the prior mean $E[\Theta]$ and adjusts based on the observation X. The adjustment is proportional to the deviation of X from its mean, $X - E[X]$. The proportionality constant $a^* = \text{Cov}(\Theta, X)/\text{var}(X)$ reflects how strongly Θ and X are linearly related and scales the deviation by the variability of X.
- **Effect of Correlation:**
  - If $\text{Cov}(\Theta, X) > 0$ (positive correlation, $\rho > 0$), an observation X above its mean ($X > E[X]$) leads to an estimate $\hat{\Theta}_L$ above the prior mean ($\hat{\Theta}_L > E[\Theta]$).
  - If $\text{Cov}(\Theta, X) < 0$ (negative correlation, $\rho < 0$), an observation X above its mean leads to an estimate $\hat{\Theta}_L$ below the prior mean.
  - If $\text{Cov}(\Theta, X) = 0$ (uncorrelated, $\rho = 0$), then $\hat{\Theta}_L = E[\Theta]$. The observation X provides no information useful for a *linear* estimate, and the best linear estimate is just the prior mean.
- **Unbiasedness:** As shown in the derivation, $E[\hat{\Theta}_L] = E[\Theta]$. The LLMS estimator is unbiased.

4.1 LLMS Error Variance (Minimum MSE for Linear Estimators)

The performance of the LLMS estimator is measured by its MSE. Let $\tilde{\Theta}_L = \Theta - \hat{\Theta}_L$ be the error.

$$\text{MSE}_{LLMS} = E[\tilde{\Theta}_L^2] = E[(\Theta - \hat{\Theta}_L)^2]$$

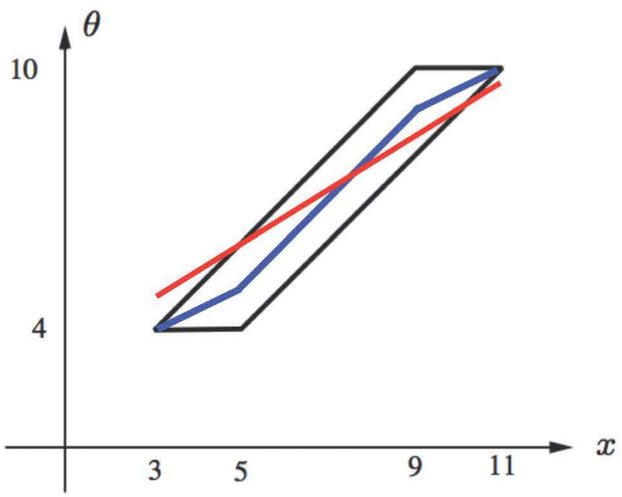
Substituting $\hat{\Theta}_L = E[\Theta] + a^*(X - E[X])$ where $a^* = \text{Cov}(\Theta, X)/\text{var}(X)$:

$$\begin{aligned}
E[(\Theta - \hat{\Theta}_L)^2] &= E[((\Theta - E[\Theta]) - a^*(X - E[X]))^2] \\
&= E[(\Theta - E[\Theta])^2] - 2a^*E[(\Theta - E[\Theta])(X - E[X])] + (a^*)^2E[(X - E[X])^2] \\
&= \text{var}(\Theta) - 2a^*\text{Cov}(\Theta, X) + (a^*)^2\text{var}(X) \\
&= \text{var}(\Theta) - 2\frac{\text{Cov}(\Theta, X)}{\text{var}(X)}\text{Cov}(\Theta, X) + \left(\frac{\text{Cov}(\Theta, X)}{\text{var}(X)}\right)^2 \text{var}(X) \\
&= \text{var}(\Theta) - 2\frac{(\text{Cov}(\Theta, X))^2}{\text{var}(X)} + \frac{(\text{Cov}(\Theta, X))^2}{\text{var}(X)} \\
&= \text{var}(\Theta) - \frac{(\text{Cov}(\Theta, X))^2}{\text{var}(X)}
\end{aligned}$$

Using the correlation coefficient $\rho^2 = \frac{(\text{Cov}(\Theta,X))^2}{\text{var}(\Theta)\text{var}(X)}$:

$$E[(\Theta - \hat{\Theta}_L)^2] = \text{var}(\Theta) - \rho^2\text{var}(\Theta) = (1 - \rho^2)\text{var}(\Theta)$$

This is the minimum MSE achievable by any linear estimator.



- The reduction in variance from the prior variance $\text{var}(\Theta)$ depends on the square of the correlation coefficient, ρ^2.
- If $\rho = 0$, the MSE is $\text{var}(\Theta)$, meaning the linear estimator provides no improvement over simply using the prior mean.
- If $|\rho| = 1$, the MSE is 0. This occurs when Θ and X have a perfect linear relationship, allowing Θ to be determined exactly from X via a linear function.

5 Example: Uniform Prior and Noise Revisited

Let's apply the LLMS formula to the example where $\Theta \sim U[4, 10]$ and $X|\{\Theta = \theta\} \sim U[\theta - 1, \theta + 1]$. We previously calculated the necessary moments:

- $E[\Theta] = 7$
- $E[X] = 7$
- $\text{var}(\Theta) = 3$ ($\sigma_\Theta = \sqrt{3}$)
- $\text{var}(X) = 10/3$ ($\sigma_X = \sqrt{10/3}$)
- $\text{Cov}(\Theta, X) = 3$

The LLMS estimator is:

$$\hat{\Theta}_L = E[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - E[X])$$

$$\hat{\Theta}_L = 7 + \frac{3}{10/3}(X - 7) = 7 + \frac{9}{10}(X - 7)$$

$$\hat{\Theta}_L = 7 + 0.9X - 6.3 = 0.9X + 0.7$$

This linear estimator $\hat{\Theta}_L = 0.9X + 0.7$ is the best linear approximation in the mean square sense.

The figure shows both the nonlinear LMS estimator $E[\Theta|X]$ (blue piecewise curve) and the LLMS estimator $\hat{\Theta}_L$ (red straight line). The LLMS provides a simpler alternative when the LMS is complex or computationally expensive.

The minimum MSE achieved by this linear estimator is:

$$E[(\Theta - \hat{\Theta}_L)^2] = (1 - \rho^2)\text{var}(\Theta)$$

First, find ρ:

$$\rho = \frac{\text{Cov}(\Theta, X)}{\sigma_\Theta \sigma_X} = \frac{3}{\sqrt{3}\sqrt{10/3}} = \frac{3}{\sqrt{10}}$$

$$\rho^2 = \frac{9}{10} = 0.9$$

$$\text{MSE}_{LLMS} = (1 - 0.9) \times 3 = 0.1 \times 3 = 0.3$$

The best linear estimator reduces the variance from the prior variance of 3 down to 0.3.

6 LLMS for Inferring Coin Bias Revisited

Consider the coin flip example: $\Theta \sim U[0, 1]$ (prior bias), $X|\Theta \sim \text{Binomial}(n, \Theta)$ (number of heads in n flips). We previously found the moments:

- $E[\Theta] = 1/2$
- $E[X] = n/2$
- $\text{var}(\Theta) = 1/12$
- $\text{var}(X) = n(n + 2)/12$
- $\text{Cov}(\Theta, X) = n/12$

Applying the LLMS formula:

$$\hat{\Theta}_{LLMS} = E[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - E[X])$$

$$\hat{\Theta}_{LLMS} = \frac{1}{2} + \frac{n/12}{n(n + 2)/12}\left(X - \frac{n}{2}\right)$$

$$\hat{\Theta}_{LLMS} = \frac{1}{2} + \frac{1}{n + 2}\left(X - \frac{n}{2}\right)$$

As simplified before, this yields:

$$\hat{\Theta}_{LLMS} = \frac{X + 1}{n + 2}$$

In this case, the LLMS estimator matches the LMS estimator $E[\Theta|X] = \frac{X+1}{n+2}$, because the LMS estimator happens to be linear in X.

7 LLMS with Multiple Observations

The LLMS framework readily extends to estimating Θ using multiple observations $X = (X_1, ..., X_n)$. We now seek the best estimator that is a linear function of *all* observations:

$$\hat{\Theta} = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + b$$

The goal is to find the coefficients $a_1, ..., a_n$ and the intercept b that minimize the MSE:

$$\min_{a_1,...,a_n,b} E[(\Theta - (a_1X_1 + \cdots + a_nX_n + b))^2]$$

Similar to the single observation case, we can solve this by setting the partial derivatives with respect to b and each a_i to zero.

- Setting $\partial/\partial b = 0$ yields $b = E[\Theta] - \sum_{i=1}^n a_i E[X_i]$. This ensures the estimator is unbiased.
- Setting $\partial/\partial a_j = 0$ for each $j = 1, ..., n$ yields a system of n linear equations involving the coefficients $a_1, ..., a_n$ and the covariances $\text{Cov}(\Theta, X_j)$ and $\text{Cov}(X_i, X_j)$. This system is often written in matrix form involving the covariance matrix of X.

The solution requires only the means $E[\Theta], E[X_i]$ and all pairwise covariances $\text{Cov}(\Theta, X_i), \text{Cov}(X_i, X_j)$.

If the true conditional expectation $E[\Theta|X_1, ..., X_n]$ is a linear function of $X_1, ..., X_n$, then the resulting LLMS estimator will be identical to the LMS estimator.

If multiple parameters Θ_j need to be estimated, the LLMS procedure is typically applied separately to find the best linear estimator $\hat{\Theta}_j = \sum_i a_{ji}X_i + b_j$ for each Θ_j.

8 Example: LLMS with Multiple Noisy Observations

Consider the model from Lecture 15:

$$X_i = \Theta + W_i, \quad i = 1, ..., n$$

where $\Theta, W_1, ..., W_n$ are uncorrelated random variables. Let $E[\Theta] = x_0$, $\text{var}(\Theta) = \sigma_0^2$. Let $E[W_i] = 0$, $\text{var}(W_i) = \sigma_i^2$.

We seek the LLMS estimator $\hat{\Theta}_{LLMS} = a_1X_1 + \cdots + a_nX_n + b$.

If we additionally assume normality and independence (as in Lecture 17), we found:

$$\hat{\Theta}_{LMS} = E[\Theta|X] = \frac{\frac{x_0}{\sigma_0^2} + \sum_{i=1}^n \frac{X_i}{\sigma_i^2}}{\frac{1}{\sigma_0^2} + \sum_{i=1}^n \frac{1}{\sigma_i^2}}$$

Since this $\hat{\Theta}_{LMS}$ is already linear in $X_1, ..., X_n$, it must be the LLMS solution under these normality and independence assumptions: $\hat{\Theta}_{LLMS} = \hat{\Theta}_{LMS}$.

Now, consider the case with general distributions, retaining only the assumptions about means, variances, and uncorrelatedness. The LLMS solution depends only on these moments. We need to compute $E[X_i]$ and $\text{Cov}(\Theta, X_i)$, $\text{Cov}(X_i, X_j)$.

- $E[X_i] = E[\Theta + W_i] = E[\Theta] + E[W_i] = x_0 + 0 = x_0$.
- $\text{Cov}(\Theta, X_i) = \text{Cov}(\Theta, \Theta + W_i) = \text{Cov}(\Theta, \Theta) + \text{Cov}(\Theta, W_i)$. Since Θ, W_i are uncorrelated, $\text{Cov}(\Theta, W_i) = 0$. So, $\text{Cov}(\Theta, X_i) = \text{var}(\Theta) = \sigma_0^2$.
- For $i = j$: $\text{var}(X_i) = \text{var}(\Theta + W_i) = \text{var}(\Theta) + \text{var}(W_i)$ (since Θ, W_i uncorrelated) $= \sigma_0^2 + \sigma_i^2$.
- For $i \neq j$: $\text{Cov}(X_i, X_j) = \text{Cov}(\Theta + W_i, \Theta + W_j) = \text{Cov}(\Theta, \Theta) + \text{Cov}(\Theta, W_j) + \text{Cov}(W_i, \Theta) + \text{Cov}(W_i, W_j)$. Since all variables are uncorrelated, the cross-terms are zero. So, $\text{Cov}(X_i, X_j) = \text{var}(\Theta) = \sigma_0^2$.

These are exactly the same first and second moments as in the independent normal case. Because the LLMS solution depends only on these moments, the resulting LLMS estimator $\hat{\Theta}_{LLMS}$ must be the same formula as derived in the normal case, even for general distributions (as long as they share these moments and uncorrelatedness property).

$$\hat{\Theta}_{LLMS} = \frac{\frac{x_0}{\sigma_0^2} + \sum_{i=1}^{n} \frac{X_i}{\sigma_i^2}}{\frac{1}{\sigma_0^2} + \sum_{i=1}^{n} \frac{1}{\sigma_i^2}}$$

(Note: Deriving this directly from the LLMS linear system requires more algebra but confirms the result).

9 Importance of Data Representation in LLMS

A crucial difference between LMS and LLMS lies in their sensitivity to data transformations.

- **LMS:** $E[\Theta|X]$ uses all information in X. If $Y = f(X)$ is some transformation (e.g., $Y = X^3$), then conditioning on Y contains the same (or less, if f is not invertible) information as conditioning on X. If f is invertible, $E[\Theta|X]$ and $E[\Theta|Y]$ represent the same underlying conditional distribution mean, just expressed as different functions of their respective conditioning variables.
- **LLMS:** The estimator is restricted to be linear in the *given* data representation.
  - The best linear estimator based on X is $\hat{\Theta}_1 = aX + b$.
  - The best linear estimator based on $Y = X^3$ is $\hat{\Theta}_2 = cY + d = cX^3 + d$.
  - In general, $\hat{\Theta}_1 \neq \hat{\Theta}_2$. Transforming the data changes the space of functions considered.
- **Feature Engineering:** This sensitivity allows for improving LLMS performance by applying nonlinear transformations to the original data X to create new features, and then finding the best linear estimator based on these engineered features.
  - E.g., Use features (X, X^2, X^3). Find $\hat{\Theta} = a_1X + a_2X^2 + a_3X^3 + b$ minimizing MSE. This is still an LLMS problem, but in a higher-dimensional feature space.
  - E.g., Use features $(X, e^X, \log X)$. Find $\hat{\Theta} = a_1X + a_2e^X + a_3 \log X + b$ minimizing MSE.
- By including relevant nonlinear transformations of the data as linear features, LLMS can approximate complex nonlinear relationships more effectively.

Lecture 20: Inequalities, Convergence, and the Weak Law of Large Numbers

Instructor: Prof. Abolfazl Hashemi

1 Probability Inequalities

Often, we do not know the full probability distribution (PDF or PMF) of a random variable X, but we might know its mean, variance, or other properties. Probability inequalities provide a way to make useful statements about the likelihood of certain events (often, events where the random variable takes on extreme values) based only on this limited information.

1.1 The Markov Inequality

The Markov inequality provides an upper bound on the probability that a *non-negative* random variable takes on a value significantly larger than its mean.

Statement: If X is a random variable such that $X \geq 0$ (takes only non-negative values) and its expected value $E[X]$ is finite, then for any constant $a > 0$:

$$P(X \geq a) \leq \frac{E[X]}{a}$$

Intuition: If a non-negative random variable has a small average value, it's unlikely that the variable often takes on very large values, because large values would significantly increase the average. The inequality quantifies this intuition.

Proof (Continuous Case): Recall the definition of the expected value for a non-negative continuous random variable:

$$E[X] = \int_0^\infty x f_X(x) dx$$

We can split the integral:

$$E[X] = \int_0^a x f_X(x) dx + \int_a^\infty x f_X(x) dx$$

Since $X \geq 0$, both x and $f_X(x)$ are non-negative in the integrals. Therefore, the first integral is non-negative:

$$\int_0^a x f_X(x) dx \geq 0$$

This implies:

$$E[X] \geq \int_a^\infty x f_X(x) dx$$

Now, within the remaining integral, x is always greater than or equal to a ($x \geq a$). Therefore, we can replace x with a inside the integral, which potentially makes the integral smaller:

$$E[X] \geq \int_a^\infty a f_X(x) dx = a \int_a^\infty f_X(x) dx$$

The integral $\int_a^\infty f_X(x)dx$ is, by definition, the probability $P(X \geq a)$.

$$E[X] \geq aP(X \geq a)$$

Since $a > 0$, we can divide by a to get the Markov inequality:

$$P(X \geq a) \leq \frac{E[X]}{a}$$

(A similar proof holds for discrete random variables using sums instead of integrals).

Examples:

1. Let $X \sim$ Exponential($\lambda = 1$). Here $X \geq 0$ and $E[X] = 1/\lambda = 1$. Applying Markov for $a > 0$:

$$P(X \geq a) \leq \frac{1}{a}$$

   For comparison, the exact probability is $P(X \geq a) = e^{-a}$. For $a = 2$, Markov gives $P(X \geq 2) \leq 1/2 = 0.5$, while the exact value is $e^{-2} \approx 0.135$. The bound is correct but can be quite loose.

2. Let $X \sim$ Uniform$[-4, 4]$. Can we apply Markov to find $P(X \geq 3)$? No, because X is not always non-negative. The condition $X \geq 0$ is essential for the proof and the validity of the inequality. (We could apply it to X^2 or $|X|$ if needed, as these are non-negative).

1.2 The Chebyshev Inequality

The Chebyshev inequality provides a bound on the probability that a random variable deviates from its mean by more than a certain amount. It uses both the mean and the variance and does not require the random variable to be non-negative.

Statement: Let X be a random variable with finite mean $\mu = E[X]$ and finite variance $\sigma^2 = \text{var}(X)$. Then for any constant $c > 0$:

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

Intuition: If the variance (a measure of spread) is small, the probability of observing a value far from the mean must also be small.

Proof: Let $Y = (X - \mu)^2$. Since Y is a squared value, $Y \geq 0$. We can apply the Markov inequality to Y. The mean of Y is $E[Y] = E[(X - \mu)^2] = \sigma^2$. Now consider the event $|X - \mu| \geq c$. Since $c > 0$, this is equivalent to the event $(X - \mu)^2 \geq c^2$, or $Y \geq c^2$. Applying the Markov inequality to Y with $a = c^2$ (note $a > 0$ since $c > 0$):

$$P(Y \geq c^2) \leq \frac{E[Y]}{c^2}$$

Substituting back $Y = (X - \mu)^2$ and $E[Y] = \sigma^2$:

$$P((X - \mu)^2 \geq c^2) \leq \frac{\sigma^2}{c^2}$$

Since $P((X-\mu)^2 \geq c^2) = P(|X-\mu| \geq c)$, we get the Chebyshev inequality:

$$P(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

Alternative Form: Often, the deviation c is expressed in terms of standard deviations, $c = k\sigma$, where $k > 0$. Substituting $c = k\sigma$ into the inequality:

$$P(|X-\mu| \geq k\sigma) \leq \frac{\sigma^2}{(k\sigma)^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

This means the probability that X falls k or more standard deviations away from its mean is at most $1/k^2$. For example, the probability of being 2 or more standard deviations away is at most $1/2^2 = 1/4$. The probability of being 3 or more standard deviations away is at most $1/3^2 = 1/9$.

Example: Let $X \sim \text{Exponential}(\lambda = 1)$. We know $\mu = 1$ and $\sigma^2 = 1$. Let's bound $P(X \geq 3)$. The event $X \geq 3$ implies $X - 1 \geq 2$. Since X is non-negative, $X - 1$ could be negative, but $X \geq 3$ definitely implies $|X - 1| \geq 2$. So, $P(X \geq 3) \leq P(|X-1| \geq 2)$. Using Chebyshev with $\mu = 1, \sigma^2 = 1, c = 2$:

$$P(|X-1| \geq 2) \leq \frac{\sigma^2}{c^2} = \frac{1}{2^2} = \frac{1}{4} = 0.25$$

So, Chebyshev gives $P(X \geq 3) \leq 0.25$. Compare: Markov gave $P(X \geq 3) \leq 1/3 \approx 0.333$. Exact value is $e^{-3} \approx 0.05$. Chebyshev provides a tighter bound than Markov in this case, but both are quite loose compared to the exact value. The power of these inequalities lies in their generality - they apply regardless of the specific distribution shape, using only moments.

2 The Weak Law of Large Numbers (WLLN)

The Law of Large Numbers is a cornerstone theorem connecting probability theory to observed frequencies and averages. It mathematically justifies the intuitive idea that as we collect more data, the sample average tends to approach the true underlying mean. The Weak Law (WLLN) formalizes this using the concept of convergence in probability.

Setup:

- Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables.
- Assume they have a finite mean $E[X_i] = \mu$ and a finite variance $\text{var}(X_i) = \sigma^2$.
- Define the sample mean (or sample average) after n observations: $M_n = \frac{X_1 + \cdots + X_n}{n}$.

Properties of the Sample Mean:

- Mean: $E[M_n] = E\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n E[X_i] = \frac{1}{n}(n\mu) = \mu$. The sample mean is an unbiased estimator of the population mean.
- Variance: Using the independence of X_i:

$$\text{var}(M_n) = \text{var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\text{var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \text{var}(X_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

  The variance of the sample mean decreases as n increases, indicating that M_n becomes more concentrated around its mean μ.

Proof of WLLN using Chebyshev: We can apply the Chebyshev inequality to the random variable M_n. Its mean is μ and its variance is σ^2/n. For any fixed constant $\epsilon > 0$:

$$P(|M_n - \mu| \geq \epsilon) \leq \frac{\text{var}(M_n)}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

Now, consider the limit as the sample size n goes to infinity:

$$\lim_{n\to\infty} P(|M_n - \mu| \geq \epsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2}$$

Since σ^2 and ϵ^2 are finite positive constants, the right-hand side goes to 0:

$$\lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0$$

Since probabilities are non-negative, $P(|M_n - \mu| \geq \epsilon) \geq 0$. By the Squeeze Theorem, we must have:

$$\lim_{n\to\infty} P(|M_n - \mu| \geq \epsilon) = 0$$

This is the statement of the Weak Law of Large Numbers.

WLLN Statement: If $X_1, X_2, \ldots$ are i.i.d. random variables with finite mean μ and finite variance σ^2, then for any $\epsilon > 0$:

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) = 0$$

(Note: The WLLN actually holds even if the variance is infinite, as long as the mean is finite, but the proof using Chebyshev requires finite variance).

2.1 Interpretations of the WLLN

The WLLN formalizes two key intuitive ideas:

1. **Averaging Reduces Noise:** If we take repeated noisy measurements $X_i = \mu + W_i$ of a true value μ, where W_i are i.i.d. noise terms with $E[W_i] = 0$, then $E[X_i] = \mu$. The sample mean $M_n = \mu + \frac{1}{n}\sum W_i$. The WLLN implies $P(|M_n - \mu| \geq \epsilon) \to 0$, meaning the average of the measurements M_n becomes a highly reliable estimate of μ for large n.
2. **Frequencies Approach Probabilities:** If we repeat an experiment independently n times and let $X_i = 1$ if an event A occurs on trial i and $X_i = 0$ otherwise ($P(A) = p$), then $E[X_i] = p$. The sample mean $M_n = \frac{\sum X_i}{n}$ is the relative frequency of event A in n trials. The WLLN implies $P(|M_n - p| \geq \epsilon) \to 0$, meaning the observed relative frequency gets arbitrarily close to the true probability p with high probability as n increases. This provides the theoretical basis for estimating probabilities from experimental frequencies.

2.2 Application: The Pollster's Problem

A classic application is estimating the proportion p of a population favoring a certain option (e.g., voting "yes").

- We randomly sample n individuals.
- $X_i = 1$ if person i favors "yes", $X_i = 0$ otherwise. Assume $X_i \sim \text{Bernoulli}(p)$ are i.i.d.

- The sample proportion is $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. This is our estimate of p.
- We want to determine the sample size n needed to achieve a certain accuracy with high probability. For example, we want the error $|M_n - p|$ to be less than 0.01 (1 percentage point) with, say, 95% probability. This means we want $P(|M_n - p| \geq 0.01) \leq 0.05$.

Using the Chebyshev bound derived from WLLN:

$$P(|M_n - p| \geq \epsilon) \leq \frac{\text{var}(X_i)}{n\epsilon^2} = \frac{p(1-p)}{n\epsilon^2}$$

Setting $\epsilon = 0.01$. The term $p(1-p)$ is unknown but is maximized when $p = 0.5$, with $p(1-p) \leq 1/4$. Using this worst-case value:

$$P(|M_n - p| \geq 0.01) \leq \frac{1/4}{n(0.01)^2} = \frac{1}{4n(0.0001)} = \frac{2500}{n}$$

To ensure this probability is ≤ 0.05:

$$\frac{2500}{n} \leq 0.05 \implies n \geq \frac{2500}{0.05} = \frac{2500}{1/20} = 50000$$

According to the Chebyshev bound, a sample size of $n = 50,000$ is needed to guarantee the error is less than 0.01 with at least 95% probability.

If we tried $n = 10,000$, the bound gives:

$$P(|M_{10000} - p| \geq 0.01) \leq \frac{2500}{10000} = 0.25$$

This only guarantees the probability is at most 25%. (In practice, due to the Central Limit Theorem, the required sample size is much smaller, typically around $n = 1000$ to $n = 2500$ for this level of accuracy, because the Chebyshev bound is often very conservative).

3 Convergence in Probability

The WLLN statement, $\lim_{n\to\infty} P(|M_n - \mu| \geq \epsilon) = 0$, motivates a general definition for how a sequence of random variables can converge to a constant.

Definition: A sequence of random variables $Y_1, Y_2, \ldots$ is said to **converge in probability** to a constant a if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|Y_n - a| \geq \epsilon) = 0$$

We denote this by $Y_n \xrightarrow{P} a$.

The WLLN can thus be concisely stated as: If X_i are i.i.d. with mean μ (and finite variance for the Chebyshev proof), then $M_n \xrightarrow{P} \mu$.

3.1 Understanding Convergence in Probability

It's helpful to contrast this with the familiar convergence of a sequence of deterministic numbers, $a_n \to a$.

- **Ordinary Convergence ($a_n \to a$):** For any desired closeness $\epsilon > 0$, the terms a_n must *eventually* (for all $n \geq n_0$) fall within the interval $[a-\epsilon, a+\epsilon]$ and *stay* there.
- **Convergence in Probability ($Y_n \xrightarrow{P} a$):** For any desired closeness $\epsilon > 0$, the probability that Y_n falls *outside* the interval $[a-\epsilon, a+\epsilon]$ must go to zero as $n \to \infty$. This doesn't mean Y_n can never be far from a for large n, but the chance of this happening becomes vanishingly small. Essentially, the probability distribution of Y_n becomes increasingly concentrated around a.

3.2 Properties of Convergence in Probability

Convergence in probability behaves well with arithmetic operations and continuous functions.

- If $X_n \xrightarrow{P} a$ and $Y_n \xrightarrow{P} b$, then:
  - $X_n + Y_n \xrightarrow{P} a+b$
  - $X_n Y_n \xrightarrow{P} ab$
- **Continuous Mapping Theorem:** If g is a function that is continuous at a, and $X_n \xrightarrow{P} a$, then $g(X_n) \xrightarrow{P} g(a)$. For example, if $M_n \xrightarrow{P} \mu$, then $M_n^2 \xrightarrow{P} \mu^2$.
- **Limitation regarding Expectations:** A crucial point is that $Y_n \xrightarrow{P} a$ does **not** imply $E[Y_n] \to a$. The limit of the expectation may not equal the expectation of the limit.

3.3 Examples of Convergence in Probability

Example 1 (Expectation vs. Probability Limit): Consider the sequence Y_n defined by the PMF:

$$p_{Y_n}(y) = \begin{cases} 1-1/n & \text{if } y = 0 \\ 1/n & \text{if } y = n^2 \\ 0 & \text{otherwise} \end{cases}$$

Does $Y_n \xrightarrow{P} 0$? For any $\epsilon > 0$, if n is large enough so $n^2 \geq \epsilon$:

$$P(|Y_n - 0| \geq \epsilon) = P(Y_n = n^2) = \frac{1}{n}$$

Since $\lim_{n\to\infty}(1/n) = 0$, we have $Y_n \xrightarrow{P} 0$. However, the expectation is:

$$E[Y_n] = 0 \cdot \left(1 - \frac{1}{n}\right) + n^2 \cdot \left(\frac{1}{n}\right) = n$$

Here, $\lim_{n\to\infty} E[Y_n] = \infty$. The expectation diverges even though the random variable converges in probability to 0. The rare but increasingly large value n^2 prevents the expectation from converging.

Example 2 (Minimum of Uniforms): Let $X_1, X_2, \ldots$ be i.i.d. $U[0,1]$. Let $Y_n = \min\{X_1, \ldots, X_n\}$. Does $Y_n \xrightarrow{P} 0$? For $0 < \epsilon < 1$:

$$P(|Y_n - 0| \geq \epsilon) = P(Y_n \geq \epsilon)$$

The minimum is $\geq \epsilon$ if and only if all X_i are $\geq \epsilon$.

$$P(Y_n \geq \epsilon) = P(X_1 \geq \epsilon, \ldots, X_n \geq \epsilon) = \prod_{i=1}^{n} P(X_i \geq \epsilon)$$

Since $P(X_i \geq \epsilon) = 1 - \epsilon$ for $X_i \sim U[0,1]$:

$$P(Y_n \geq \epsilon) = (1 - \epsilon)^n$$

As $n \to \infty$, since $0 < 1 - \epsilon < 1$, we have $(1 - \epsilon)^n \to 0$. Thus, $Y_n = \min\{X_1, \ldots, X_n\} \xrightarrow{P} 0$.

4 Related Topics and Further Convergence Concepts

This lecture introduced basic inequalities and the WLLN, which uses convergence in probability. There are several related and more advanced topics:

- **Better Bounds/Approximations:** While Markov and Chebyshev are general, they are often loose. Other tools provide tighter bounds or approximations for tail probabilities, such as the Chernoff bound (using moment generating functions) and the Central Limit Theorem (providing a normal approximation for the distribution of sums/averages).
- **Other Types of Convergence:** Besides convergence in probability, other important modes exist:
  - **Almost Sure Convergence (Convergence with Probability 1):** $P(\lim_{n\to\infty} Y_n = a) = 1$. This is a stronger condition than convergence in probability. The **Strong Law of Large Numbers (SLLN)** states that under similar conditions to WLLN, the sample mean M_n converges almost surely to μ.
  - **Convergence in Distribution:** The sequence of cumulative distribution functions $F_{Y_n}(y)$ converges to a limiting CDF $F_Y(y)$ at all points where F_Y is continuous. This is the mode of convergence related to the Central Limit Theorem.

Lecture 21: The Central Limit Theorem (CLT)

Instructor: Prof. Abolfazl Hashemi

1 Overview: From WLLN to CLT

In our study of inequalities and convergence, we established the Weak Law of Large Numbers (WLLN). The WLLN states that for a sequence of independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$, their sample mean M_n converges in probability to the true mean μ:

$$M_n = \frac{X_1 + \cdots + X_n}{n} \xrightarrow{P} \mu$$

This is a statement about convergence, telling us that the distribution of M_n becomes tightly concentrated around μ. However, it does not tell us the *shape* of the distribution.

The Central Limit Theorem (CLT) answers this question. It describes the shape of the distribution of the sum $S_n = X_1 + \cdots + X_n$ (or its standardized version) as n becomes large. It is one of the most profound and useful results in all of probability and statistics.

This lecture will cover:

- A precise statement of the CLT.
- The universality and practical usefulness of the theorem.
- Examples of how to use the CLT for approximations.
- A specific refinement for approximating discrete random variables (the 1/2 continuity correction).
- An application to statistical polling.

2 Different Scalings of Sums of Random Variables

Let $X_1, \ldots, X_n$ be i.i.d. random variables with a finite mean μ and a finite variance σ^2. Let $S_n = X_1 + \cdots + X_n$. We can examine the behavior of this sum under different scalings:

- **The Sum S_n:**
  - $E[S_n] = E[X_1 + \cdots + X_n] = n\mu$
  - $\text{var}(S_n) = \text{var}(X_1 + \cdots + X_n) = n\sigma^2$ (by independence)

  As $n \to \infty$, the variance $n\sigma^2 \to \infty$. The distribution of S_n spreads out and does not converge to a stable form.
- **The Sample Mean M_n:**
  - $M_n = S_n/n$
  - $E[M_n] = E[S_n/n] = n\mu/n = \mu$
  - $\text{var}(M_n) = \text{var}(S_n/n) = (1/n^2)\text{var}(S_n) = (1/n^2)(n\sigma^2) = \sigma^2/n$

As $n \to \infty$, the variance $\sigma^2/n \to 0$. The distribution of M_n shrinks and becomes concentrated at the single point μ. This is the WLLN, but it does not give us a limiting *shape*.

- **An Intermediate Scaling:**
  - $S_n/\sqrt{n}$
  - $E[S_n/\sqrt{n}] = n\mu/\sqrt{n} = \mu\sqrt{n}$
  - $\text{var}(S_n/\sqrt{n}) = (1/n)\text{var}(S_n) = (1/n)(n\sigma^2) = \sigma^2$

  This scaling results in a constant variance, but the mean $\mu\sqrt{n}$ diverges (assuming $\mu \neq 0$). This also fails to converge.

The correct approach to get a stable, non-degenerate limiting distribution is to standardize the random variable.

3 The Central Limit Theorem (CLT)

To standardize S_n, we first center it by subtracting its mean, and then scale it by its standard deviation.

- $S_n = X_1 + \cdots + X_n$
- $E[S_n] = n\mu$
- $\text{var}(S_n) = n\sigma^2 \implies \text{std. dev.}(S_n) = \sqrt{n}\sigma$

We define the standardized sum Z_n as:

$$Z_n = \frac{S_n - E[S_n]}{\text{std. dev.}(S_n)} = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

Let's find the mean and variance of this new random variable Z_n:

$$E[Z_n] = E\left[\frac{S_n - n\mu}{\sqrt{n}\sigma}\right] = \frac{E[S_n] - n\mu}{\sqrt{n}\sigma} = \frac{n\mu - n\mu}{\sqrt{n}\sigma} = 0$$

$$\text{var}(Z_n) = \text{var}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma}\right) = \text{var}\left(\frac{S_n}{\sqrt{n}\sigma}\right) = \frac{1}{n\sigma^2}\text{var}(S_n) = \frac{1}{n\sigma^2}(n\sigma^2) = 1$$

For any n, Z_n is a random variable with a mean of 0 and a variance of 1. The CLT states that as n increases, the *distribution* of Z_n converges to the standard normal distribution.

The Central Limit Theorem (CLT): Let $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables with finite mean μ and finite, non-zero variance σ^2. Let $S_n = X_1 + \cdots + X_n$. Let $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$. Let Z be a standard normal random variable, $Z \sim N(0,1)$, with CDF $\Phi(z)$.

Then, for every z:

$$\lim_{n\to\infty} P(Z_n \leq z) = \Phi(z)$$

3.1 Usefulness of the CLT

The CLT is arguably the most important theorem in probability for practical applications.

- **Universality:** It is a universal result. It applies to *any* underlying distribution for the X_i, whether they are discrete (like Bernoulli, Poisson) or continuous (like Uniform, Exponential), as long as they have a finite mean and variance.
- **Simplicity:** It allows us to make probabilistic statements using only the mean μ and variance σ^2, without needing to know the full PMF or PDF of the X_i.
- **Computational Shortcut:** Calculating the exact PMF/PDF of S_n requires $n-1$ convolutions, which is computationally intractable for large n. The CLT provides a simple, excellent approximation.
- **Justification for Normal Models:** It explains why the normal distribution appears so frequently in nature. Many real-world random phenomena (e.g., measurement errors, noise in a signal, height of a person) are the aggregate result of many small, independent random factors. The CLT states that the sum of such factors will be approximately normally distributed.

3.2 Theoretical vs. Practical Implications

- **Theory:** The CLT is a precise mathematical statement about the convergence of the CDF of Z_n to the normal CDF $\Phi(z)$. Stronger versions of the theorem exist, showing convergence of PDFs/PMFs under certain conditions. The theorem can also be generalized to cases where X_i are not identically distributed, or even have weak dependence. The standard proof uses transforms (Moment Generating Functions), by showing that the MGF of Z_n converges to the MGF of a standard normal: $\lim_{n\to\infty} E[e^{sZ_n}] = e^{s^2/2}$.
- **Practice:** The practical use of the CLT is as an **approximation** for a finite n. We treat Z_n as if it *is* a standard normal random variable. This is equivalent to treating the sum S_n as if it is a normal random variable:

  $$S_n \approx N(n\mu, n\sigma^2)$$

  This approximation is generally considered good for "moderate" n, often $n > 30$ is cited as a rule of thumb. However, the quality of the approximation depends heavily on the underlying distribution of X_i. If X_i is itself symmetric and unimodal (bell-shaped), the convergence is very fast. If X_i is highly skewed (like an exponential), it requires a larger n for the sum to become symmetric and bell-shaped.

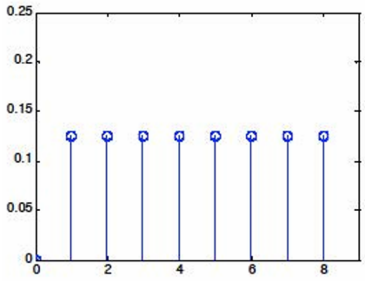
3.3 Visualizing the CLT

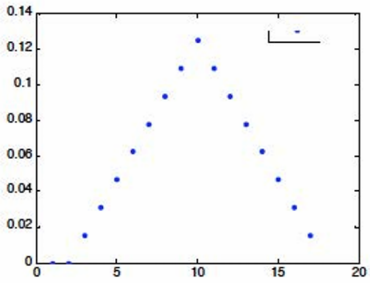
Symmetric Case: Sum of Uniform Random Variables

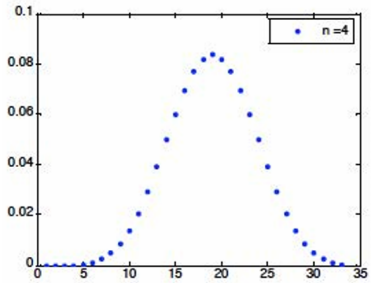
Consider $X_i \sim U[0, 8]$ (discrete, integer values). The PMF for S_n (shifted and scaled) is shown for increasing n. The original PMF ($n = 1$) is symmetric. Convergence to the normal shape is very fast.

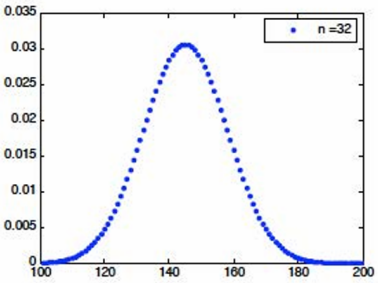
Non-Symmetric Case: Sum of Geometric Random Variables

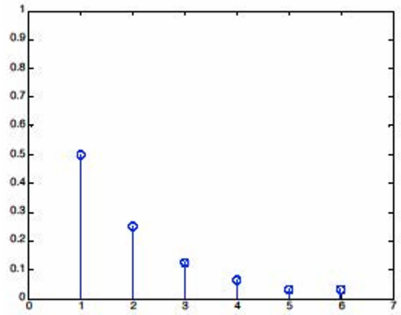
Consider $X_i \sim$ Geometric(p). The PMF for S_n is shown for increasing n. The original PMF ($n = 1$) is highly skewed. Convergence is much slower, but even for $n = 32$, the distribution's shape is clearly approaching the symmetric, bell-shaped normal curve.

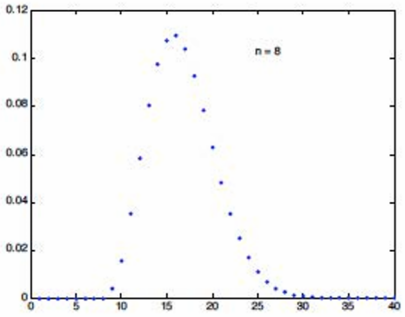


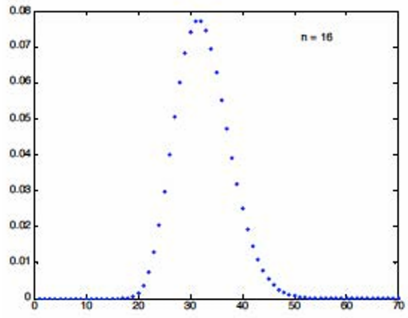


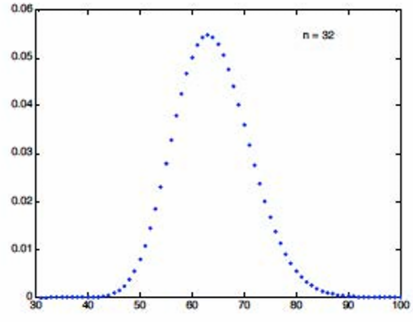

n =4


n =32




n = 8


n = 16


n = 32

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520

4 Examples of CLT Approximations

The standard problem form is: given $P(S_n \leq a) \approx b$, find one parameter given the other two.

Setup for Examples 1-4: Let X_i be the weights of packages, which are i.i.d. exponential random variables with parameter $\lambda = 1/2$. From the properties of the exponential distribution:

- Mean: $\mu = E[X_i] = 1/\lambda = 2$
- Variance: $\sigma^2 = \text{var}(X_i) = 1/\lambda^2 = 4$.
- Standard Deviation: $\sigma = \sqrt{4} = 2$.

Let $S_n = X_1 + \cdots + X_n$. By the CLT, $S_n \approx N(n\mu, n\sigma^2) = N(2n, 4n)$. The standardized sum is $Z_n = \frac{S_n - 2n}{\sqrt{4n}} = \frac{S_n - 2n}{2\sqrt{n}}$.

4.1 Example 1: Find a Probability

If we load $n = 100$ packages, what is $P(S_{100} \geq 210)$?

- We approximate $S_{100} \approx N(2 \cdot 100, 4 \cdot 100) = N(200, 400)$.
- The standard deviation is $\sqrt{400} = 20$.
- Standardize the value 210: $z = \frac{210 - \mu_{S_n}}{\sigma_{S_n}} = \frac{210 - 200}{20} = \frac{10}{20} = 0.5$.
- $P(S_{100} \geq 210) \approx P(Z \geq 0.5)$, where $Z \sim N(0, 1)$.
- $P(Z \geq 0.5) = 1 - P(Z < 0.5) = 1 - \Phi(0.5)$.

- From the normal table, $\Phi(0.5) \approx 0.6915$.
- $P(S_{100} \geq 210) \approx 1 - 0.6915 = 0.3085$.

4.2 Example 2: Find a Threshold

Let $n = 100$. Find the capacity a such that $P(S_{100} \geq a) \approx 0.05$.

- We are approximating $S_{100} \sim N(200, 400)$.
- We want to find a such that $P(S_{100} \geq a) \approx 0.05$.
- Standardize a: $P\left(\frac{S_{100}-200}{20} \geq \frac{a-200}{20}\right) \approx 0.05$.
- $P\left(Z \geq \frac{a-200}{20}\right) \approx 0.05$.
- We need to find the value z such that $P(Z \geq z) = 0.05$, or $P(Z < z) = 0.95$.
- From the normal table, $\Phi(1.64) \approx 0.9495$ and $\Phi(1.65) \approx 0.9505$. We use $z \approx 1.645$.
- $\frac{a-200}{20} \approx 1.645 \implies a \approx 200 + 20(1.645) = 200 + 32.9 = 232.9$.

4.3 Example 3: Find Sample Size

Find n such that $P(S_n \geq 210) \approx 0.05$.

- We are approximating $S_n \sim N(2n, 4n)$.
- We want $P\left(\frac{S_n-2n}{2\sqrt{n}} \geq \frac{210-2n}{2\sqrt{n}}\right) \approx 0.05$.
- $P\left(Z \geq \frac{210-2n}{2\sqrt{n}}\right) \approx 0.05$.
- From Example 2, the critical value for Z is 1.645.
- $\frac{210-2n}{2\sqrt{n}} \approx 1.645 \implies 210 - 2n \approx 3.29\sqrt{n}$.
- Let $y = \sqrt{n}$. We have a quadratic equation: $2n+3.29\sqrt{n}-210 \approx 0 \implies 2y^2+3.29y-210 \approx 0$.
- $y = \frac{-b\pm\sqrt{b^2-4ac}}{2a} = \frac{-3.29\pm\sqrt{3.29^2-4(2)(-210)}}{4} = \frac{-3.29\pm\sqrt{10.82+1680}}{4}$.
- Since $y = \sqrt{n} > 0$: $y \approx \frac{-3.29+\sqrt{1690.82}}{4} \approx \frac{-3.29+41.12}{4} = \frac{37.83}{4} \approx 9.46$.
- $n = y^2 \approx (9.46)^2 \approx 89.5$. Thus, $n \approx 89$ packages.

4.4 Example 4: Alternative Formulation

Find $P(N > 100)$, where N is the number of packages needed for the total weight to exceed 210.

- The event $\{N > 100\}$ means "it takes more than 100 packages to exceed 210 lbs."
- This is logically identical to the event $\{S_{100} \leq 210\}$, meaning "the sum of the first 100 packages did not exceed 210 lbs."
- $P(N > 100) = P(S_{100} \leq 210)$.
- From Example 1, $P(S_{100} \geq 210) \approx 0.3085$.

- $P(S_{100} \leq 210) = 1 - P(S_{100} > 210)$. Since the normal is continuous, $P(S_{100} > 210) = P(S_{100} \geq 210)$.
- $P(N > 100) \approx 1 - 0.3085 = 0.6915$.

5 Normal Approximation to the Binomial PMF

A primary application of the CLT is the De Moivre-Laplace approximation to the binomial distribution.

- Let $X_i \sim \text{Bernoulli}(p)$. Then $E[X_i] = p$ and $\text{var}(X_i) = p(1-p)$.
- Let $S_n = X_1 + \cdots + X_n$. S_n follows a Binomial(n, p) distribution.
- By the CLT, $S_n \approx N(np, np(1-p))$.

Example: $n = 36, p = 0.5$. Find $P(S_{36} \leq 21)$.

- Mean: $np = 36(0.5) = 18$.
- Variance: $np(1-p) = 36(0.5)(0.5) = 9$.
- Standard deviation: $\sqrt{9} = 3$.
- Exact calculation: $P(S_{36} \leq 21) = \sum_{k=0}^{21} \binom{36}{k}(0.5)^{36} = 0.8785$.
- Naive CLT approximation: $P(S_n \leq 21) \approx P\left(Z \leq \frac{21-18}{3}\right) = P(Z \leq 1) = \Phi(1) \approx 0.8413$.

This approximation is not very accurate. The discrepancy arises because we are approximating a discrete random variable (which places probability mass at integer points) with a continuous one.

5.1 The 1/2 Correction for Integer Random Variables

The binomial S_n only takes integer values. The event $S_n \leq 21$ is the sum of the probabilities $P(S_n = 0), \ldots, P(S_n = 21)$. When we approximate this sum with an integral under the normal curve, the bar at $k = 21$ is best represented by the interval $[20.5, 21.5]$. To approximate $P(S_n \leq 21)$, we should integrate the normal PDF up to the right edge of this bar, which is 21.5.

This is the **continuity correction** (or 1/2 correction):

$$P(S_n \leq 21) \approx P(S_n^{\text{normal}} \leq 21.5)$$

Applying this correction:

$$P(S_n \leq 21.5) \approx P\left(Z \leq \frac{21.5 - 18}{3}\right) = P\left(Z \leq \frac{3.5}{3}\right) \approx P(Z \leq 1.17)$$

From the table, $\Phi(1.17) \approx 0.8790$. This is an excellent approximation of the exact value 0.8785.

5.2 Approximating the PMF

We can also use the continuity correction to approximate the probability of a single value, $P(S_n = k)$. We approximate this as the area under the normal curve over the interval $[k - 0.5, k + 0.5]$.

$$P(S_n = k) \approx P(k - 0.5 \leq S_n^{\text{normal}} \leq k + 0.5)$$

$$P(S_n = k) \approx \Phi\left(\frac{k + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - 0.5 - np}{\sqrt{np(1-p)}}\right)$$

Example: Find $P(S_{36} = 19)$ for $p = 0.5$.

- Exact answer: $\binom{36}{19}(0.5)^{36} \approx 0.1251$.
- CLT approximation:

$$P(18.5 \leq S_n \leq 19.5) \approx P\left(\frac{18.5 - 18}{3} \leq Z \leq \frac{19.5 - 18}{3}\right)$$

$$= P(0.166... \leq Z \leq 0.5) \approx \Phi(0.5) - \Phi(0.17)$$

$$\approx 0.6915 - 0.5675 = 0.1240$$

This is again a very close approximation.

6 The Pollster's Problem Revisited

We wish to find n such that $P(|M_n - p| \geq 0.01) \leq 0.05$. $M_n = S_n/n$. The event is $P(|S_n/n - p| \geq 0.01) = P(|S_n - np| \geq 0.01n)$. Standardizing, we want:

$$P\left(\left|\frac{S_n - np}{\sqrt{np(1-p)}}\right| \geq \frac{0.01n}{\sqrt{np(1-p)}}\right) \leq 0.05$$

$$P\left(|Z| \geq \frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right) \leq 0.05$$

The term $p(1-p)$ is unknown. We use the worst-case (largest) value $p(1-p) = 0.25$ (at $p = 0.5$), which makes the variance largest and the probability bound most conservative.

$$P\left(|Z| \geq \frac{0.01\sqrt{n}}{\sqrt{0.25}}\right) \leq 0.05 \implies P(|Z| \geq 0.02\sqrt{n}) \leq 0.05$$

By symmetry, this is $2 \cdot P(Z \geq 0.02\sqrt{n}) \leq 0.05$, or $P(Z \geq 0.02\sqrt{n}) \leq 0.025$. This implies $P(Z < 0.02\sqrt{n}) \geq 0.975$. From the normal table, $\Phi(1.96) = 0.975$. So, we need:

$$0.02\sqrt{n} \geq 1.96$$

$$\sqrt{n} \geq \frac{1.96}{0.02} = 98$$

$$n \geq 98^2 = 9604$$

This sample size of 9,604 is far more efficient than the $n \geq 25,000$ required by the Chebyshev inequality, demonstrating the power of the CLT.

Lecture 22: Classical Statistics I

Instructor: Prof. Abolfazl Hashemi

1 Introduction to Classical Statistics

This lecture transitions from the Bayesian framework of inference to the **classical** (or **frequentist**) framework. The core philosophy changes:

- **Main Idea:** The unknown parameter, θ, is a fixed, deterministic constant. It is **not** a random variable.
- Our goal is to estimate this unknown constant θ.

2 Classical Statistics vs. Bayesian Inference

Let's contrast the two major approaches to statistics:

- **Bayesian Inference (Recap):**
  - The unknown parameter θ is treated as a **random variable**.
  - We must specify a **prior distribution** $p_\theta(\theta)$ or $f_\theta(\theta)$ that represents our belief about θ before seeing data.
  - The observation X is also a random variable.
  - We use Bayes' rule to compute the **posterior distribution** $p_{\theta|X}(\theta|x)$ or $f_{\theta|X}(\theta|x)$. This posterior distribution is the complete answer to the inference problem.
- **Classical Statistics:**
  - The unknown parameter θ is a **fixed, deterministic constant**. There is no prior distribution.
  - The observation X is a random variable, and its distribution *depends* on the value of the constant θ.
  - We write this as $p_X(x;\theta)$ or $f_X(x;\theta)$. This is the likelihood of observing x if the parameter's true value is θ.
  - **Crucial point:** $p_X(x;\theta)$ is **not** a conditional probability in the Bayesian sense, because θ is not a random variable.
  - We are not working with a single probabilistic model, but rather a **family of models**, one for each possible value of θ.
  - An **estimator**, $\hat{\Theta} = g(X)$, is a function of the observations. It is a random variable, and we use its value (the **estimate**) to guess the true value of θ.

3 Problem Types in Classical Statistics

Using the classical setup, where θ is a constant determining the model $p_X(x;\theta)$, we can define several types of problems:

- **Parameter Estimation:** The parameter θ can be any value within a continuous (e.g., $\theta \in [0,1]$) or discrete set. The goal is to design an estimator $\hat{\Theta} = g(X)$ such that the estimation error, $\hat{\Theta} - \theta$, is "small" in some probabilistic sense.
- **Binary Hypothesis Testing:** We must decide between two possible values for the parameter.
$$H_0 : \theta = 1/2 \quad \text{versus} \quad H_1 : \theta = 3/4$$
- **Composite Hypothesis Testing:** At least one of the hypotheses corresponds to a set of values.
$$H_0 : \theta = 1/2 \quad \text{versus} \quad H_1 : \theta \neq 1/2$$

This lecture will focus on parameter estimation.

4 Estimating a Mean and Estimator Properties

The most common estimation problem is finding an unknown mean.

- Let $X_1, \ldots, X_n$ be i.i.d. observations.
- Let $\theta = \mathbb{E}[X_i]$ be the unknown mean.
- Let $\sigma^2 = \text{Var}(X_i)$ be the variance (which may also be unknown).

A natural estimator for θ is the **sample mean**:

$$\hat{\Theta}_n = M_n = \frac{X_1 + \cdots + X_n}{n}$$

We can evaluate this estimator based on several desirable properties:

- **Bias:** The bias of an estimator is the expected difference between the estimator and the true parameter.
$$b(\hat{\Theta}_n) = E[\hat{\Theta}_n - \theta] = E[\hat{\Theta}_n] - \theta$$
For the sample mean, we find $E[\hat{\Theta}_n]$:
$$E[\hat{\Theta}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}(n\theta) = \theta$$
Since $E[\hat{\Theta}_n] = \theta$ (or $b(\hat{\Theta}_n) = 0$) for all possible values of θ, the sample mean is an **unbiased** estimator.
- **Consistency:** An estimator is consistent if it converges in probability to the true parameter as $n \to \infty$.
$$\hat{\Theta}_n \xrightarrow{P} \theta$$
By the Weak Law of Large Numbers (WLLN), we know that the sample mean M_n converges in probability to the true mean μ (which is θ in our notation). Therefore, $\hat{\Theta}_n$ is a **consistent** estimator.

- **Mean Squared Error (MSE):** The MSE is the expected squared estimation error:

$$\text{MSE} = E[(\hat{\Theta}_n - \theta)^2]$$

4.1 Mean Squared Error Decomposition

For any estimator $\hat{\Theta}$ of a constant parameter θ, we can decompose the MSE. Let $Z = \hat{\Theta} - \theta$ be the error. We know $E[Z^2] = \text{Var}(Z) + (E[Z])^2$.

$$E[(\hat{\Theta} - \theta)^2] = \text{Var}(\hat{\Theta} - \theta) + (E[\hat{\Theta} - \theta])^2$$

Since θ is a constant, $\text{Var}(\hat{\Theta} - \theta) = \text{Var}(\hat{\Theta})$. The second term is the bias squared, $(b(\hat{\Theta}))^2$. This gives the fundamental decomposition:

$$\textbf{MSE} = \text{Var}(\hat{\Theta}) + (\text{bias})^2$$

For an unbiased estimator (like the sample mean), the bias is zero, so its MSE is simply its variance:

$$\text{MSE}(\hat{\Theta}_n) = \text{Var}(\hat{\Theta}_n) + 0 = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

The $\sqrt{\text{Var}(\hat{\Theta})}$ is often called the **standard error** of the estimator.

5 Confidence Intervals (CIs)

A point estimate $\hat{\theta}$ gives a single number, but no sense of our certainty. A **confidence interval** provides a range of values that is likely to contain the true parameter θ.

Definition (Confidence Interval): A $1-\alpha$ confidence interval for a parameter θ is an interval $[\hat{\Theta}^-, \hat{\Theta}^+]$ calculated from the data $X_1, \ldots, X_n$ such that

$$P(\hat{\Theta}^- \leq \theta \leq \hat{\Theta}^+) \geq 1 - \alpha$$

for all possible values of θ. The value $1-\alpha$ (e.g., 0.95) is the **confidence level**.

The interpretation is critical: θ is fixed, and the interval endpoints $\hat{\Theta}^-$ and $\hat{\Theta}^+$ are random variables (because they depend on the random data X_i). The property $P(\ldots) \geq 1-\alpha$ means that if we repeated the entire experiment (sampling n observations) many times, and constructed an interval each time, at least $100(1-\alpha)\%$ of these random intervals would successfully "capture" the true, fixed parameter θ.

5.1 CI for the Mean (Known Variance σ^2)

We can use the CLT to construct an approximate CI for the mean θ using the sample mean estimator $\hat{\Theta}_n = M_n$. The CLT states that $Z_n = \frac{\hat{\Theta}_n - \theta}{\sigma/\sqrt{n}}$ is approximately $N(0,1)$.

To construct a 95% CI ($1-\alpha = 0.95 \implies \alpha = 0.05$), we find the value $z_{\alpha/2}$ from the normal table such that $P(Z > z_{\alpha/2}) = \alpha/2 = 0.025$. This value is $z_{0.025} = 1.96$, because $\Phi(1.96) = 0.975$. By symmetry, $P(-1.96 \leq Z \leq 1.96) = 0.95$.

$$P\left(-1.96 \leq \frac{\hat{\Theta}_n - \theta}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95$$

We now "invert" the inequality to solve for θ:

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \hat{\Theta}_n - \theta \leq 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

$$P\left(\hat{\Theta}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \theta \leq \hat{\Theta}_n + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

Thus, our 95% confidence interval is:

$$\left[\hat{\Theta}_n - 1.96\frac{\sigma}{\sqrt{n}}, \quad \hat{\Theta}_n + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

5.2 CIs for the Mean (Unknown Variance σ^2)

The previous formula requires σ, which is often unknown. We have three options:

1. **Use an Upper Bound:** If we know σ is bounded, we can use a worst-case value. For Bernoulli(θ) variables, $\sigma^2 = \theta(1-\theta) \leq 1/4$, so $\sigma \leq 1/2$. Using this gives a conservative (wider) interval.
2. **Use a Plug-in Estimate:** We can estimate σ^2 from the data.
   - For Bernoulli: $\hat{\sigma}^2 = \hat{\Theta}_n(1-\hat{\Theta}_n)$.
   - In general: Use the **sample variance**. The **ML estimator** for $v = \sigma^2$ is $\hat{v}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\Theta}_n)^2$.

   We then plug this $\hat{\sigma} = \sqrt{\hat{v}_n}$ into the CI formula in place of σ.
3. **Use the t-distribution:** The plug-in method (Option 2) relies on two approximations: the CLT, and $\hat{\sigma} \approx \sigma$. For small n, this second approximation adds significant uncertainty. If the X_i are *exactly* normal, the distribution of the statistic $T = \frac{\hat{\Theta}_n - \theta}{\hat{S}_n/\sqrt{n}}$ (where $\hat{S}_n^2 = \frac{1}{n-1}\sum(X_i - \hat{\Theta}_n)^2$ is the *unbiased* sample variance) is not normal. It follows a **t-distribution** with $n-1$ degrees of freedom. This distribution is wider than the normal, accounting for the uncertainty in our variance estimate. For small n, one should use the t-distribution tables instead of the $N(0,1)$ tables to find the critical value (e.g., 1.96 is replaced by a larger value).

6 Method of Moments Estimators

The "plug-in" approach from Option 2 is a general technique called the **Method of Moments**. The idea is to equate theoretical expectations (which are functions of θ) with their corresponding sample averages and solve for θ.

- $\mathbb{E}[X] \longleftrightarrow \frac{1}{n}\sum X_i$
- $\mathbb{E}[g(X)] \longleftrightarrow \frac{1}{n}\sum g(X_i)$
- $\text{Var}(X) = \mathbb{E}[X^2] - (E[X])^2 \longleftrightarrow \left(\frac{1}{n}\sum X_i^2\right) - \left(\frac{1}{n}\sum X_i\right)^2$

This provides a natural estimator for almost any statistical quantity:

- **Mean Estimator:** $\hat{\Theta}_X = M_n = \frac{1}{n}\sum X_i$
- **Variance Estimator:** $\hat{v}_X = \frac{1}{n}\sum(X_i - M_n)^2$

- **Covariance Estimator:** $\hat{\text{Cov}}(X,Y) = \frac{1}{n}\sum(X_i - M_{n,X})(Y_i - M_{n,Y})$

7 Maximum Likelihood (ML) Estimation

The Method of Moments is intuitive, but a more formal and powerful principle is **Maximum Likelihood (ML) estimation**.

Idea: We have observed data $x = (x_1, \ldots, x_n)$. We ask: "What value of the parameter θ would make observing this data *most likely*?"

Likelihood Function: This is the PMF or PDF $p_X(x;\theta)$ or $f_X(x;\theta)$, viewed as a function of θ for the fixed data x.

ML Estimate: The ML estimate is the value $\hat{\theta}_{ML}$ that maximizes this function.

$$\hat{\theta}_{ML} = \arg\max_{\theta} p_X(x;\theta) \quad \text{or} \quad \hat{\theta}_{ML} = \arg\max_{\theta} f_X(x;\theta)$$

In practice, it is almost always easier to maximize the **log-likelihood** function, $\log p_X(x;\theta)$, since the log function is monotonic and turns products (from i.i.d. data) into sums.

7.1 ML vs. MAP

- MAP: $\hat{\theta}_{MAP} = \arg\max_\theta p_{X|\theta}(x|\theta)p_\theta(\theta)$ (likelihood $\times$ prior)
- ML: $\hat{\theta}_{ML} = \arg\max_\theta p_X(x;\theta)$ (where $p_X(x;\theta) = p_{X|\theta}(x|\theta)$)

ML estimation is equivalent to Bayesian MAP estimation if we assume a **uniform (flat) prior** for θ.

7.2 Properties of ML Estimators

For n i.i.d. observations, the ML estimator $\hat{\Theta}_n$ has excellent properties:

- It is **consistent** ($\hat{\Theta}_n \xrightarrow{P} \theta$).
- It is **asymptotically normal** (its error distribution approaches a normal distribution, allowing for CI construction).
- It is **asymptotically efficient**, meaning for large n, it has the smallest possible variance among "good" estimators.

7.3 Example 1: Parameter of Binomial

We observe $K = k$ heads in n trials. $\theta = p$ is the unknown parameter. 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. θ and set to 0:

$$\frac{dL}{d\theta} = \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0 \implies \frac{k}{\theta} = \frac{n-k}{1-\theta}$$

$$k(1-\theta) = (n-k)\theta \implies k - k\theta = n\theta - k\theta \implies k = n\theta$$

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

The ML estimator is the sample mean $\hat{\Theta}_{ML} = K/n$.

7.4 Example 2: Normal Mean and Variance

We observe $X_1, \ldots, X_n$ from $N(\mu, v)$, where $\theta = (\mu, v)$ is the 2D parameter vector. The log-likelihood function is:

$$L(\mu, v) = \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log v - \frac{(x_i - \mu)^2}{2v} \right)$$

$$L(\mu, v) = C - \frac{n}{2}\log v - \frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2$$

We must set the partial derivatives w.r.t. μ and v to zero.

1. $\frac{\partial L}{\partial \mu} = 0 - 0 - \frac{1}{2v}\sum_{i=1}^{n} 2(x_i - \mu)(-1) = \frac{1}{v}\sum_{i=1}^{n}(x_i - \mu) = 0$ This implies $\sum x_i - \sum \mu = 0 \implies \sum x_i - n\mu = 0 \implies \hat{\mu}_{ML} = \frac{1}{n}\sum x_i$. The ML estimator for the mean is the sample mean.
2. $\frac{\partial L}{\partial v} = 0 - \frac{n}{2v} - (-\frac{1}{2v^2})\sum_{i=1}^{n}(x_i - \mu)^2 = 0$ Plug in $\hat{\mu}_{ML}$ for μ:

$$\frac{1}{2v^2}\sum_{i=1}^{n}(x_i - \hat{\mu}_{ML})^2 = \frac{n}{2v}$$

$$\hat{v}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{ML})^2$$

The ML estimator for the variance is the sample variance (the biased version).

Lecture 23: Classical Statistics II

Instructor: Prof. Abolfazl Hashemi

1 Overview

This lecture continues our study of classical statistics, focusing on two central topics:

1. **Linear Regression:** We move from estimating a single parameter (like the mean) to estimating the relationship between two variables. We will develop a method to fit a line to a set of noisy data points.
2. **Binary Hypothesis Testing:** We will formalize the process of deciding between two competing hypotheses, H_0 and H_1, based on observed data.

2 Linear Regression

A very common problem in science and engineering is to find a model that describes the relationship between two (or more) variables. We are given n pairs of data points:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

We hypothesize that there is an approximately linear relationship between the variable x (the "explanatory" variable) and the variable y (the "dependent" variable). We model this relationship as:

$$Y \approx \theta_0 + \theta_1 X$$

Here, θ_0 (the intercept) and θ_1 (the slope) are the unknown, deterministic parameters that we wish to estimate from our data.

2.1 The Least Squares Approach

The goal of linear regression is to find the specific line $y = \hat{\theta}_0 + \hat{\theta}_1 x$ that provides the "best fit" to the n data points.

To do this, we must define what "best fit" means. For each data point i, the **residual** is the vertical error between the observed value y_i and the value predicted by the line, $\hat{\theta}_0 + \hat{\theta}_1 x_i$.

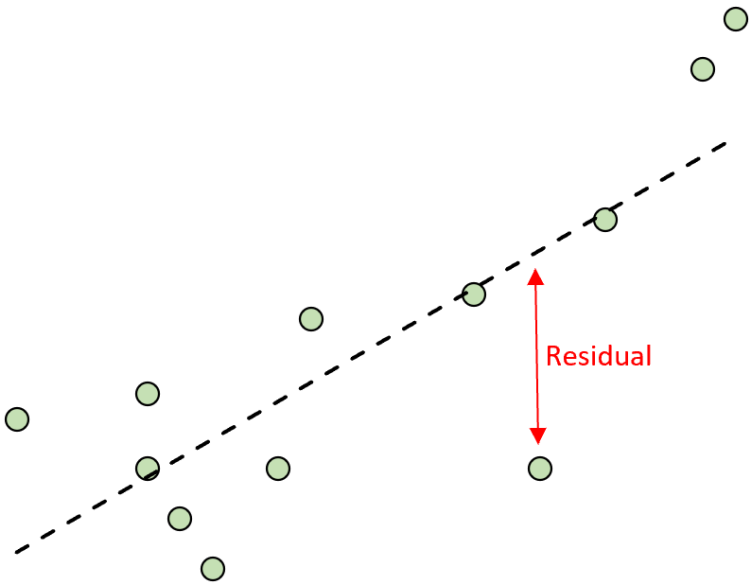
$$\text{Residual}_i = y_i - (\hat{\theta}_0 + \hat{\theta}_1 x_i)$$

The **method of least squares** defines the best line as the one that minimizes the sum of the squares of these residuals. We seek the parameters $(\hat{\theta}_0, \hat{\theta}_1)$ that solve this optimization problem:

$$(\hat{\theta}_0, \hat{\theta}_1) = \arg\min_{\theta_0, \theta_1} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

We define the cost function $J(\theta_0, \theta_1)$ as this sum of squared errors.

The figure shows a scatter plot of data points. The line represents a candidate model $y = \theta_0 + \theta_1 x$. The vertical lines represent the residuals, and the least squares method finds the line that minimizes the sum of the squares of the lengths of these vertical lines.



2.2 Solution to the Least Squares Problem

The cost function $J(\theta_0, \theta_1) = \sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)^2$ is a quadratic function of the two parameters θ_0 and θ_1. To find the minimum, we can set the partial derivatives with respect to each parameter equal to zero.

$$\frac{\partial J}{\partial \theta_0} = 0 \quad \text{and} \quad \frac{\partial J}{\partial \theta_1} = 0$$

1. $\frac{\partial J}{\partial \theta_0} = \sum_{i=1}^{n} 2(y_i - \theta_0 - \theta_1 x_i)(-1) = 0 \implies \sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i) = 0$

$$\sum_{i=1}^{n} y_i - n\theta_0 - \theta_1 \sum_{i=1}^{n} x_i = 0$$

Dividing by n, and letting $\bar{y} = \frac{1}{n}\sum y_i$ and $\bar{x} = \frac{1}{n}\sum x_i$ be the sample means:

$$\bar{y} - \theta_0 - \theta_1 \bar{x} = 0 \implies \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

This first result shows that the optimal line must pass through the "center of mass" $(\bar{x}, \bar{y})$ of the data.

2. $\frac{\partial J}{\partial \theta_1} = \sum_{i=1}^{n} 2(y_i - \theta_0 - \theta_1 x_i)(-x_i) = 0 \implies \sum_{i=1}^{n}(y_i x_i - \theta_0 x_i - \theta_1 x_i^2) = 0$ Substitute $\theta_0 = \bar{y} - \theta_1 \bar{x}$ into this second equation:

$$\sum_{i=1}^{n}(y_i x_i - (\bar{y} - \theta_1 \bar{x})x_i - \theta_1 x_i^2) = 0$$

$$\sum_{i=1}^{n} y_i x_i - \bar{y}\sum_{i=1}^{n} x_i + \theta_1 \bar{x}\sum_{i=1}^{n} x_i - \theta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum y_i x_i - \bar{y}(n\bar{x}) + \theta_1 \bar{x}(n\bar{x}) - \theta_1 \sum x_i^2 = 0$$

$$\theta_1 \left(n\bar{x}^2 - \sum x_i^2\right) = n\bar{x}\bar{y} - \sum y_i x_i$$

$$\hat{\theta}_1 = \frac{\sum y_i x_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

The final expressions are the sample covariance (numerator) divided by the sample variance (denominator).

Linear Regression Estimates:

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

2.3 Probabilistic Justifications for Least Squares

The choice to minimize the square of the errors (as opposed to absolute values or something else) seems arbitrary, but it has deep probabilistic justifications.

1. Maximum Likelihood (ML) Estimation Assume the data is generated by a probabilistic model:

$$Y_i = \theta_0 + \theta_1 x_i + W_i$$

where θ_0 and θ_1 are unknown constants, and W_i represents noise. If we assume the noise terms 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 (viewed as a function of θ_0, θ_1) is:

$$f_{Y|\Theta}(y|\theta_0, \theta_1) = \prod_{i=1}^{n} f_{Y_i|\Theta}(y_i|\theta_0, \theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2}\right\}$$

To find the ML estimate, we maximize this, or equivalently, maximize its logarithm:

$$\log f_{Y|\Theta}(y|\theta_0, \theta_1) = C - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

Maximizing this expression is **equivalent** to minimizing the sum of squared residuals:

$$\min_{\theta_0, \theta_1} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

Thus, the least squares method is identical to the Maximum Likelihood estimator under the assumption of i.i.d. Gaussian noise.

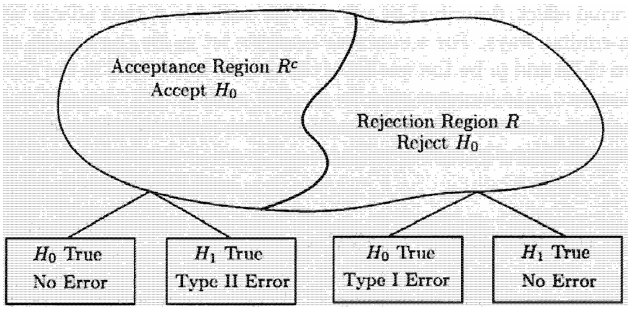
2. Bayesian Linear Regression We can also view this in a Bayesian framework by placing priors on θ_0 and θ_1. Assume $\theta_0 \sim N(0, \sigma_0^2)$, $\theta_1 \sim N(0, \sigma_1^2)$, and $W_i \sim N(0, \sigma^2)$, all independent. The model is $Y_i = \theta_0 + \theta_1 x_i + W_i$. The MAP estimate is found by maximizing the posterior, $f_{\Theta|Y}(\theta|y) \propto f_{Y|\Theta}(y|\theta) f_\Theta(\theta)$:

$$\hat{\theta}_{MAP} = \arg\max_{\theta_0, \theta_1} \left(\prod_{i=1}^{n} f_{Y_i|\Theta}(y_i|\theta) \right) f_{\Theta_0}(\theta_0) f_{\Theta_1}(\theta_1)$$

Taking the log and minimizing its negative, this is equivalent to solving:

$$\min_{\theta_0, \theta_1} \left(\sum_{i=1}^{n} \frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2} + \frac{\theta_0^2}{2\sigma_0^2} + \frac{\theta_1^2}{2\sigma_1^2} \right)$$

This is a "regularized" least squares problem. The prior terms act as penalties that pull the estimates toward their prior means (0), which prevents "overfitting" if the data is very noisy. If we let the prior variances $\sigma_0^2, \sigma_1^2 \to \infty$ (a "flat" or "uninformative" prior), the penalty terms vanish, and the MAP estimate becomes identical to the ML / least squares estimate.



2.4 Multiple and Nonlinear Regression

The same framework can be extended.

- **Multiple Linear Regression:** The model is linear in multiple variables: $Y \approx \theta_0 + \theta_1 X_1 + \theta_2 X_2$. We have n data points $(x_{i,1}, x_{i,2}, y_i)$. We minimize $J = \sum(y_i - \theta_0 - \theta_1 x_{i,1} - \theta_2 x_{i,2})^2$. This is solved by setting 3 partial derivatives to 0, resulting in a 3×3 system of linear equations.
- **Polynomial Regression:** The model is $Y \approx \theta_0 + \theta_1 X + \theta_2 X^2$. This is a nonlinear model in X, but it is *linear in the parameters* $\theta_0, \theta_1, \theta_2$. We can treat it as a multiple linear regression problem by defining new features: $X_1 = X$ and $X_2 = X^2$. We then minimize $J = \sum(y_i - \theta_0 - \theta_1 x_i - \theta_2 x_i^2)^2$. The solution is found in the same way.

3 Binary Hypothesis Testing

We now switch to a different problem: deciding between two competing hypotheses, H_0 and H_1. This is a decision problem, not an estimation problem.

3.1 Setup

In the classical framework, θ is a fixed, unknown constant. Our two hypotheses correspond to two different possible values (or sets of values) for θ.

- **Null Hypothesis (H_0):** $\theta = \theta_0$. This is the "default" or "no-change" hypothesis. Under H_0, the data X is drawn from the distribution $f_X(x; \theta_0)$.
- **Alternative Hypothesis (H_1):** $\theta = \theta_1$. This is the "effect" or "change" hypothesis. Under H_1, the data X is drawn from $f_X(x; \theta_1)$.

Our goal is to design a **decision rule** based on the observation $X = x$ to choose either H_0 or H_1.

3.2 Decision Rules and Types of Error

A decision rule partitions the entire observation space into two regions:

- **Rejection Region (R):** The set of all x for which we decide to **reject** H_0 (and accept H_1).
- **Acceptance Region (R^c):** The set of all x for which we decide to **accept** H_0.

When we make a decision, we can be correct or we can make one of two types of errors:

- **Type I Error (False Rejection):** We reject H_0 (i.e., $X \in R$) when H_0 is actually true.

$$\alpha(R) = P(\text{Reject } H_0 \mid H_0 \text{ is true}) = P(X \in R; H_0)$$

  This probability is called the **significance level** of the test.

- **Type II Error (False Acceptance):** We accept H_0 (i.e., $X \in R^c$) when H_1 is actually true.

$$\beta(R) = P(\text{Accept } H_0 \mid H_1 \text{ is true}) = P(X \in R^c; H_1)$$

  The value $1 - \beta = P(X \in R; H_1)$ is called the **power** of the test (the probability of correctly detecting H_1).

3.3 The Likelihood Ratio Test (LRT)

How do we choose the "best" rejection region R? In the Bayesian framework (Lecture 16), we chose the hypothesis that maximized the posterior probability. This MAP rule was: Decide H_1 if $P(H_1|x) > P(H_0|x)$. This is equivalent to $p_X(x|H_1)P(H_1) > p_X(x|H_0)P(H_0)$, which rearranges to:

$$\frac{p_X(x|H_1)}{p_X(x|H_0)} > \frac{P(H_0)}{P(H_1)} = \text{threshold}$$

The classical framework adopts this same structure, but without the Bayesian interpretation.

- Define the **Likelihood Ratio** $L(x)$:

$$L(x) = \frac{f_X(x;\theta_1)}{f_X(x;\theta_0)} \quad (\text{or } \frac{p_X(x;\theta_1)}{p_X(x;\theta_0)})$$

  This ratio measures how much more likely the observation x is under H_1 than under H_0.

- A **Likelihood Ratio Test (LRT)** is a decision rule of the form: Reject H_0 if $L(x) > \xi$.

- The rejection region is $R = \{x \mid L(x) > \xi\}$, where ξ is the **critical value** (or threshold) of the test.

3.4 Choosing the Threshold ξ

There is a fundamental tradeoff between α and β.

- If we make ξ very large, R becomes smaller. This decreases α (fewer false rejections) but increases β (more false acceptances).

- If we make ξ very small, R becomes larger. This increases α (more false rejections) but decreases β (more false acceptances, higher power).

The **Neyman-Pearson framework** is the standard method for choosing ξ:

1. Fix a maximum tolerable Type I error, α. This is the **significance level** (e.g., $\alpha = 0.05$).

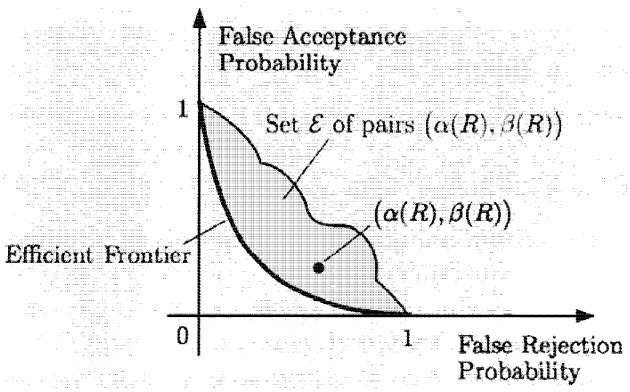
2. Find the critical value ξ that results in this exact probability:

$$P(L(X) > \xi; H_0) = \alpha$$

   This defines the test.

3. Then, calculate the resulting Type II error:

$$\beta = P(L(X) \leq \xi; H_1)$$



3.5 Neyman-Pearson Lemma

This framework is not arbitrary. The LRT is proven to be the most powerful test.

Neyman-Pearson Lemma: Among all possible decision rules (rejection regions R) that have a Type I error probability $\alpha(R) \leq \alpha$, the Likelihood Ratio Test (LRT) defined by $P(L(X) > \xi; H_0) = \alpha$ achieves the **smallest possible** Type II error probability β.

In other words, for a given "budget" of Type I error α, the LRT maximizes the power $(1 - \beta)$ of the test. It is the optimal decision rule in the classical framework.

3.6 Example: Testing Normal Means

Let $X \sim N(\theta, \sigma^2)$ with σ^2 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 $\log L(x) > \log \xi$:

$$\frac{(x^2 - 2x\theta_0 + \theta_0^2) - (x^2 - 2x\theta_1 + \theta_1^2)}{2\sigma^2} > \log \xi$$

$$\frac{2x(\theta_1 - \theta_0) + \theta_0^2 - \theta_1^2}{2\sigma^2} > \log \xi$$

Since $\theta_1 - \theta_0 > 0$ and $2\sigma^2 > 0$, we can rearrange this inequality to isolate x. The right side is just a new constant threshold γ.

$$2x(\theta_1 - \theta_0) > 2\sigma^2 \log \xi - \theta_0^2 + \theta_1^2 \implies x > \gamma$$

The LRT is a simple threshold test: **Reject** H_0 **if** $x > \gamma$.

2. Find the threshold γ for a given α. We set the Type I error probability to α:

$$P(\text{Reject } H_0; H_0) = P(X > \gamma; H_0) = \alpha$$

Under H_0, $X \sim N(\theta_0, \sigma^2)$. We standardize this probability:

$$P\left(\frac{X - \theta_0}{\sigma} > \frac{\gamma - \theta_0}{\sigma}; H_0\right) = P\left(Z > \frac{\gamma - \theta_0}{\sigma}\right) = \alpha$$

Let z_α be the critical value from the $N(0,1)$ table such that $P(Z > z_\alpha) = \alpha$ (or $\Phi(z_\alpha) = 1 - \alpha$).

$$\frac{\gamma - \theta_0}{\sigma} = z_\alpha \implies \gamma = \theta_0 + z_\alpha \sigma$$

3. Find the resulting Type II Error β.

$$\beta = P(\text{Accept } H_0; H_1) = P(X \leq \gamma; H_1)$$

Under H_1, $X \sim N(\theta_1, \sigma^2)$. We standardize this probability:

$$\beta = P\left(\frac{X - \theta_1}{\sigma} \leq \frac{\gamma - \theta_1}{\sigma}; H_1\right) = \Phi\left(\frac{\gamma - \theta_1}{\sigma}\right)$$

Substitute the expression for γ:

$$\beta = \Phi\left(\frac{(\theta_0 + z_\alpha \sigma) - \theta_1}{\sigma}\right) = \Phi\left(z_\alpha - \frac{\theta_1 - \theta_0}{\sigma}\right)$$

This final expression relates α (via z_α) and β to the "separation" of the means $(\theta_1 - \theta_0)$ relative to the noise σ.

Lecture 24: The Bernoulli Process

Instructor: Prof. Abolfazl Hashemi

1 Introduction and Definition

This lecture introduces the **Bernoulli process**, a fundamental concept in the study of stochastic processes that provides a model for a sequence of independent events occurring over time. We will define the process, explore its basic properties, particularly memorylessness, and apply these concepts to analyze the timing of successes and the merging and splitting of such processes. Finally, we will examine the Poisson approximation, which connects the Bernoulli process to the continuous-time Poisson process.

1.1 Definition of the Bernoulli Process

A Bernoulli process is formally defined as an infinite sequence of independent and identically distributed (i.i.d.) Bernoulli trials, denoted by $X_1, X_2, \ldots$. At each time step i (or trial i), the outcome X_i is a binary random variable such that:

$$P(X_i = 1) = p \quad \text{(Success or Arrival at the } i\text{-th trial)}$$

$$P(X_i = 0) = 1 - p \quad \text{(Failure at the } i\text{-th trial)}$$

The probability p is assumed to be constant across all trials.

The two key assumptions underpinning the Bernoulli process are:

1. **Independence**: The outcome of any trial X_i is probabilistically independent of the outcomes of all other trials X_j, where $i \neq j$.
2. **Time-Homogeneity**: The probability of success p is constant for all trials.

The Bernoulli process serves as a simple yet powerful model for various real-world phenomena involving discrete events over time, such as:

- The sequence of wins or losses in a repetitive game or lottery.
- The sequence of whether a bank receives a customer arrival during each one-second interval.
- The sequence of successful or failed transmissions at discrete time slots to a communication server.

The foundation of the Bernoulli process originates from the work of Jacob Bernoulli (1654-1705), who formalized the analysis of sequences of independent trials.

2 Stochastic Processes and Views

A **stochastic process** is a collection of random variables indexed by time, typically denoted as $\{X_t, t \in T\}$. For the Bernoulli process, the index T is the set of positive integers, $T = \{1, 2, 3, \ldots\}$, making it a discrete-time stochastic process.

2.1 First View: Sequence of Random Variables

In the first view, the Bernoulli process is treated as a collection of random variables $X_1, X_2, \ldots$. Key probabilistic characteristics of individual trials include:

- Expected Value: $E[X_i] = 1 \cdot p + 0 \cdot (1-p) = p$.
- Variance: $\text{var}(X_i) = E[X_i^2] - (E[X_i])^2 = p - p^2 = p(1-p)$.
- Probability Mass Function (PMF): $p_{X_i}(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$.
- Joint PMF: Due to independence, the joint distribution of any finite subset of n trials is the product of their marginal PMFs:

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

2.2 Second View: Sample Space

The second view considers the entire sample space Ω associated with the process. For an infinite sequence of Bernoulli trials, Ω is the set of all possible infinite sequences of zeros and ones:

$$\Omega = \{(x_1, x_2, x_3, \ldots) \mid x_i \in \{0, 1\}\}$$

Any event is a subset of this space. For example, the probability of observing an endless sequence of successes is:

$$P(X_i = 1 \text{ for all } i) = \lim_{n\to\infty} P(X_1 = 1, \ldots, X_n = 1) = \lim_{n\to\infty} p^n = 0 \quad \text{(assuming } p < 1)$$

3 Random Variables Associated with the Bernoulli Process

From the underlying Bernoulli process, several important related random variables can be derived.

3.1 Number of Successes/Arrivals (S) in n Time Slots

Let S be the total number of successes (ones) in the first n trials:

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

Since S is the sum of n independent Bernoulli(p) random variables, S follows the **Binomial distribution** with parameters n and p.

- Probability Mass Function:

$$P(S = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for } k = 0, 1, \ldots, n$$

- Expected Value:

$$E[S] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = np$$

- Variance: Due to independence of X_i:

$$\text{var}(S) = \text{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \text{var}(X_i) = np(1-p)$$

3.2 Time Until the First Success/Arrival (T_1)

Let T_1 be the time of the first success, defined as the trial number where the first 1 occurs.

$$T_1 = \min\{i \geq 1 \mid X_i = 1\}$$

T_1 follows the **Geometric distribution** with parameter p.

- Probability Mass Function: The first success occurs at time k if the first $k-1$ trials were failures, and the k-th trial was a success.

$$P(T_1 = k) = P(X_1 = 0, \ldots, X_{k-1} = 0, X_k = 1) = (1-p)^{k-1} p, \quad \text{for } k = 1, 2, \ldots$$

- Expected Value:

$$E[T_1] = \frac{1}{p}$$

- Variance:

$$\text{var}(T_1) = \frac{1-p}{p^2}$$

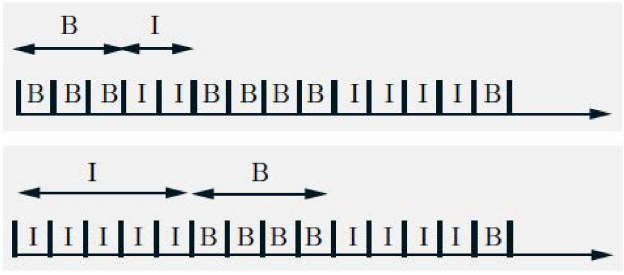
4 Memorylessness and Fresh-Start Properties

The Bernoulli process exhibits a strong memoryless property, which is crucial for its analysis.

4.1 Memorylessness and Fresh-Start

The fundamental property of a Bernoulli process is that, at any point in time n, the future sequence of outcomes $X_{n+1}, X_{n+2}, \ldots$ is independent of the past sequence $X_1, \ldots, X_n$ and is statistically identical to the original process starting at time 1. This is the **fresh-start property**.

- **Fresh-start after time n (deterministic)**: If we stop observing the process at a fixed time n, the process starting at X_{n+1} is independent of $X_1, \ldots, X_n$ and is a Bernoulli process with parameter p.
- **Fresh-start after time T_1 (random)**: Since the inter-arrival times are independent of the past, the process "forgets" the past upon the occurrence of a success. The process immediately following the first success at time T_1 is a new Bernoulli process starting at $T_1 + 1$. This reinforces the i.i.d. nature of the inter-arrival times.



4.2 Fresh-Start After a General Random Time N

The fresh-start property extends even to random times N, provided N is determined **causally** or "non-anticipatorily". A random time N is causal if the decision to stop or the value of N depends only on the history of the process up to time N.

Examples of causal random times (N):

- 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 $1, 1, 1$.

For any such causal time N, the sequence of outcomes starting after N, defined by the sequence of random variables $X_{N+1}, X_{N+2}, \ldots$, maintains the fresh-start properties:

- The future sequence $X_{N+1}, X_{N+2}, \ldots$ is a Bernoulli process with parameter p.
- The future sequence is independent of the past events $N, X_1, \ldots, X_N$.

4.3 The Distribution of Busy Periods

Consider a server that is busy (B) if $X_i = 1$ and idle (I) if $X_i = 0$. The sequence of outcomes forms a Bernoulli process.

The **first busy period** starts with the first busy slot and ends just before the first subsequent idle slot. Similarly, an idle period is a block of consecutive failures.

- The **length of a busy period** is the number of consecutive successes (1s) followed by a failure (0). This length follows a Geometric distribution on $\{1, 2, \ldots\}$ with parameter $1 - p$.
- The **length of an idle period** is the number of consecutive failures (0s) followed by a success (1). This length follows a Geometric distribution on $\{1, 2, \ldots\}$ with parameter p.

In the example sequence `B B B I I B B B B I I I I B`, the lengths of the periods are: 3 (B), 2 (I), 4 (B), 4 (I), 1 (B), and so forth.

5 Inter-Arrival Times and the Negative Binomial Distribution

5.1 Time of the k-th Success (Y_k)

Let Y_k denote the time (trial number) of the k-th success or arrival.

$$Y_k = \min\{t \geq k \mid \sum_{i=1}^{t} X_i = k\}$$

We can express Y_k as the sum of k inter-arrival times:

$$Y_k = T_1 + T_2 + \cdots + T_k$$

where T_1 is the time to the first success, and for $k \geq 2$, T_k is the **k-th inter-arrival time**, defined as the number of trials between the $(k-1)$-th success and the k-th success, $T_k = Y_k - Y_{k-1}$.

Due to the fresh-start property after every success, the sequence of inter-arrival times $T_1, T_2, \ldots$ are **i.i.d., Geometric(p)** random variables. Thus, T_2 is independent of T_1 and has the same Geometric(p) distribution.

5.2 Properties of Y_k

Y_k, the sum of k i.i.d. Geometric random variables, follows the **Negative Binomial distribution** (sometimes called the Pascal distribution).

- Expected Value: By linearity of expectation:

$$E[Y_k] = \sum_{i=1}^{k} E[T_i] = kE[T_1] = k \cdot \frac{1}{p} = \frac{k}{p}$$

- Variance: Due to the independence of the T_i:

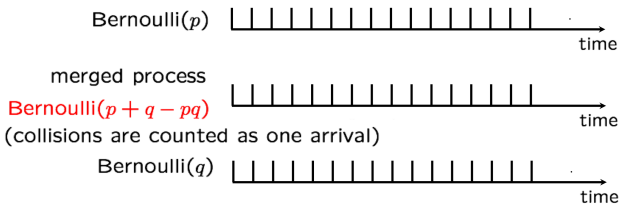
$$\text{var}(Y_k) = \sum_{i=1}^{k} \text{var}(T_i) = k\,\text{var}(T_1) = k \cdot \frac{1-p}{p^2} = \frac{k(1-p)}{p^2}$$

- Probability Mass Function ($p_{Y_k}(t)$): The k-th success occurs exactly at time t if there were precisely $k-1$ successes in the first $t-1$ trials, and the t-th trial was a success. The number of ways to arrange $k-1$ successes in $t-1$ trials is $\binom{t-1}{k-1}$.

$$p_{Y_k}(t) = P(Y_k = t) = \binom{t-1}{k-1} p^{k-1}(1-p)^{(t-1)-(k-1)} \cdot p = \binom{t-1}{k-1} p^k (1-p)^{t-k}, \quad t = k, k+1, \ldots$$

6 Operations on Bernoulli Processes

The Bernoulli process exhibits structural stability under certain operations, specifically merging and splitting.



6.1 Merging of Independent Bernoulli Processes

Consider two independent Bernoulli processes, Process 1 with success probability p and Process 2 with success probability q. We form a **merged process** where an arrival occurs if there is an arrival in either Process 1 OR Process 2 (a collision is counted as one arrival).

Let $X_i \sim \text{Bernoulli}(p)$ and $Z_i \sim \text{Bernoulli}(q)$ be the outcomes of the two processes at time i. The outcome of the merged process, M_i, is 1 if $X_i = 1$ or $Z_i = 1$.

$$P(M_i = 1) = P(X_i = 1 \cup Z_i = 1)$$

Using the addition rule for probability and the independence of X_i and Z_i:

$$P(M_i = 1) = P(X_i = 1) + P(Z_i = 1) - P(X_i = 1 \cap Z_i = 1) = p + q - pq$$

The merged process is also a **Bernoulli process** with parameter $p_{\text{merged}} = p + q - pq$.

If an arrival occurs in the merged process ($M_i = 1$), we can calculate the probability that it came from Process 1 using Bayes' rule:

$$P(\text{arrival in Process 1} \mid \text{arrival in merged process}) = P(X_i = 1 \mid M_i = 1)$$

Since $X_i = 1$ implies $M_i = 1$:

$$P(X_i = 1 \mid M_i = 1) = \frac{P(X_i = 1 \cap M_i = 1)}{P(M_i = 1)} = \frac{P(X_i = 1)}{P(M_i = 1)} = \frac{p}{p + q - pq}$$

6.2 Splitting of a Bernoulli Process

Consider a primary Bernoulli process $X_i \sim \text{Bernoulli}(p)$. We split the successes into two output streams using a secondary independent Bernoulli trial (an auxiliary coin flip) with success probability q for each success.

- **Stream 1** receives the arrival if $X_i = 1$ AND the auxiliary flip is a success (probability q).
- **Stream 2** receives the arrival if $X_i = 1$ AND the auxiliary flip is a failure (probability $1 - q$).

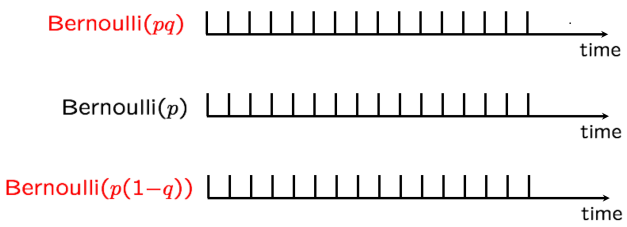
The coin flips used for splitting are assumed to be independent of the primary Bernoulli process X_i.

Let Y_i be the outcome of Stream 1 at time i.

$$P(Y_i = 1) = P(X_i = 1) \cdot P(\text{assigned to Stream 1}) = pq$$

Let Z_i be the outcome of Stream 2 at time i.

$$P(Z_i = 1) = P(X_i = 1) \cdot P(\text{assigned to Stream 2}) = p(1 - q)$$



The key result is that the two resulting streams, Stream 1 (Bernoulli(pq)) and Stream 2 (Bernoulli($p(1-q)$), are not only Bernoulli processes but are also **independent** of each other. This is because, at any time i, the condition $Y_i = 1$ depends on the outcome $X_i = 1$ and the auxiliary flip being q, while $Z_i = 1$ depends on $X_i = 1$ and the auxiliary flip being $1 - q$. For a given $X_i = 1$, Y_i and Z_i are determined by mutually exclusive events in the auxiliary space. For $X_i = 0$, both Y_i and Z_i are 0. When considering the sequence of trials, the outcomes (Y_i, Z_i) are independent of (Y_j, Z_j) for $i \neq j$.

7 Poisson Approximation to Binomial

The **Poisson distribution** arises as the limiting distribution of the Binomial distribution under specific conditions. This is important because the Poisson distribution is the basis for the continuous-time Poisson process, which is the continuous-time analog of the discrete-time Bernoulli process.

The approximation is accurate when the number of trials n is large, the probability of success p is small, but their product $\lambda = np$ is a moderate, fixed value. This corresponds to the **rare events regime**.

Let $S \sim \text{Binomial}(n, p)$ be the number of arrivals in n slots, with PMF:

$$p_S(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \quad k = 0, \ldots, n$$

Fact: Fix $k \geq 0$ and let $n \to \infty$ such that $p = \lambda/n$.

$$p_S(k) \to \frac{\lambda^k}{k!} e^{-\lambda}, \quad \text{for } k = 0, 1, \ldots$$

This is the PMF of a Poisson(λ) random variable.

The derivation involves separating the terms of $p_S(k)$:

$$p_S(k) = \frac{n(n-1)\cdots(n-k+1)}{k!} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

As $n \to \infty$:

- The first term $\frac{n(n-1)\cdots(n-k+1)}{n^k} \to 1$.
- The term $\left(\frac{\lambda}{n}\right)^k$ contributes $\frac{\lambda^k}{n^k}$.
- The term $\left(1 - \frac{\lambda}{n}\right)^{n-k} \to e^{-\lambda}$.

Combining these, we get the Poisson PMF. The result formalizes the notion that when the expected number of events λ is fixed, the actual distribution of the number of events is well approximated by the Poisson distribution, regardless of the large n and small p.

Lecture 25: The Poisson Process Part I

Instructor: Prof. Abolfazl Hashemi

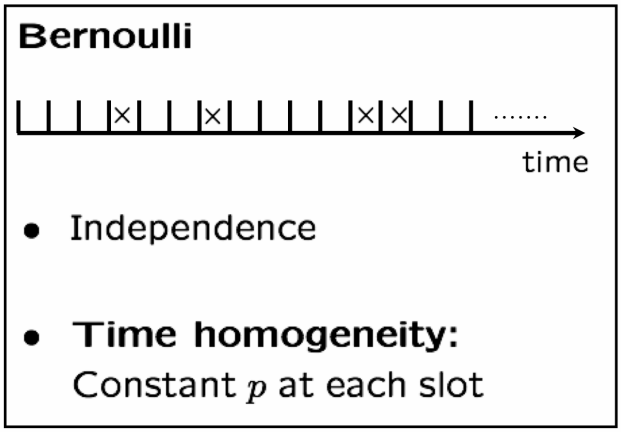
1 Definition of the Poisson Process

The **Poisson process** is the fundamental continuous-time model for describing events that occur randomly over time, such as arrivals, emissions, or occurrences of rare phenomena. It can be viewed as the continuous-time limit of the Bernoulli process. The process is characterized by an arrival rate, λ, and two core properties: independent increments and time homogeneity.

1.1 Key Properties and Small Interval Probabilities

The properties of a Poisson process with rate λ are formalized by considering a very small time interval of duration δ:

1. **Independent Increments**: The number of arrivals in disjoint time intervals are independent random variables.
2. **Time Homogeneity**: The arrival rate λ is constant over time.



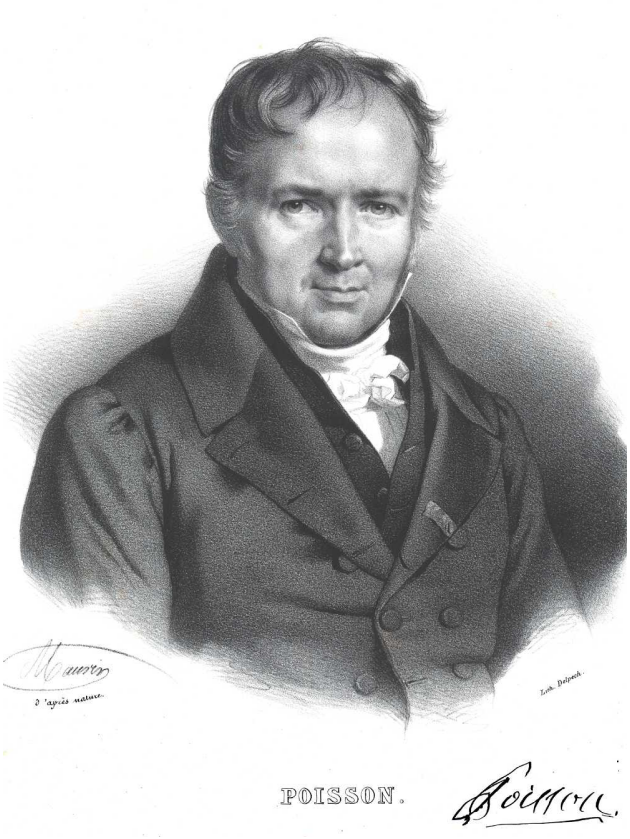
For an infinitesimally small δ, the probability of k arrivals, denoted $P(k, \delta)$, is approximated as follows:

- The probability of **no arrivals** ($k = 0$) is $P(0, \delta) \approx 1 - \lambda\delta]$.
- The probability of **exactly one arrival** ($k = 1$) is $P(1, \delta) \approx \lambda\delta$.
- The probability of **more than one arrival** ($k > 1$) is $P(k, \delta) \approx 0$.

A more rigorous expression incorporates terms of order δ^2 ($O(\delta^2)$):

$$P(k, \delta) = \begin{cases} 1 - \lambda\delta + O(\delta^2) & \text{if } k = 0 \\ \lambda\delta + O(\delta^2) & \text{if } k = 1 \\ 0 + O(\delta^2) & \text{if } k > 1 \end{cases}$$

The parameter λ is the constant **arrival rate** per unit time.



1.2 Applications

The Poisson process, named after Siméon Denis Poisson (1781-1840), is widely used across various domains to model rare and independent events.

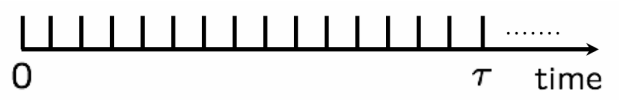
Applications include:

- Modeling deaths from external, random causes, such as the historical data on deaths from horse kicks in the Prussian army (1898).
- Describing physical phenomena, such as particle emissions and radioactive decay.
- Modeling communication and networking events, such as photon arrivals from a weak source or phone calls and service requests.
- Modeling rare occurrences in finance, such as financial market shocks.

2 Distribution of the Number of Arrivals

2.1 Poisson PMF for N_τ

Let N_τ be the number of arrivals in a fixed time interval of duration τ. To find the probability $P(k,\tau) = P(N_\tau = k)$, we use the Poisson approximation to the Binomial distribution. We divide the interval τ into $n = \tau/\delta$ small slots. The probability of success in each slot is $p = \lambda\delta + O(\delta^2)$. In the limit as $\delta \to 0$ (and thus $n \to \infty$) while maintaining the average number of arrivals $\lambda\tau$:



The number of arrivals N_τ follows the **Poisson distribution** with parameter $\lambda\tau$:

$$P(k,\tau) = P(N_\tau = k) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \ldots$$

2.2 Mean and Variance

Since N_τ is a Poisson random variable with parameter $\lambda\tau$, its mean and variance are equal.

- Expected Value:
$$E[N_\tau] = \lambda\tau$$
- Variance:
$$\text{var}(N_\tau) = \lambda\tau$$

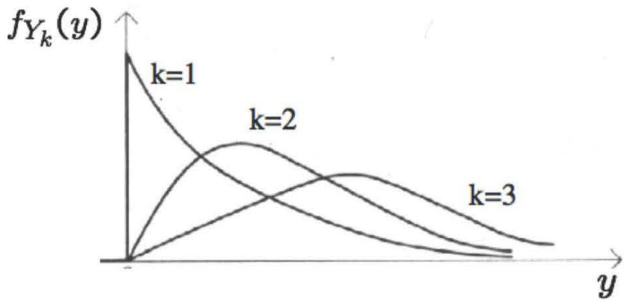
The expected value can be derived directly from the PMF:

$$E[N_\tau] = \sum_{k=0}^{\infty} k \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}$$

2.3 Example: Email Arrivals

Suppose emails arrive according to a Poisson process at a rate of $\lambda = 5$ messages per hour.

- **Mean and variance of mails received during a day**: For a day, the time duration is $\tau = 24$ hours.
$$\lambda\tau = 5 \times 24 = 120$$
$$E[N_{24}] = 120$$
$$\text{var}(N_{24}) = 120$$
- **Probability of one new message in the next hour**: Here, $\tau = 1$ hour and $k = 1$.
$$P(N_1 = 1) = \frac{(5 \cdot 1)^1 e^{-(5 \cdot 1)}}{1!} = 5e^{-5}$$
- **Probability of exactly two messages during each of the next three hours**: This involves three independent intervals of duration $\tau = 1$ hour. The probability of 2 messages in a single hour is $P(N_1 = 2) = \frac{5^2 e^{-5}}{2!} = \frac{25}{2} e^{-5}$. The overall probability is the product:
$$P(N_{H_1} = 2, N_{H_2} = 2, N_{H_3} = 2) = P(N_1 = 2)^3 = \left(\frac{25}{2} e^{-5}\right)^3$$



3 Timing of Arrivals

3.1 Time until the First Arrival (T_1)

The time T_1 until the first arrival is a continuous random variable. We find its cumulative distribution function (CDF) by relating the event $\{T_1 > t\}$ to the number of arrivals in $[0, t]$:

$$P(T_1 \leq t) = 1 - P(T_1 > t) = 1 - P(N_t = 0)$$

$$P(T_1 \leq t) = 1 - \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = 1 - e^{-\lambda t}, \quad t \geq 0$$

T_1 follows the **Exponential distribution** with rate λ. The probability density function (PDF) is found by differentiating the CDF:

$$f_{T_1}(t) = \lambda e^{-\lambda t}, \quad t \geq 0$$

3.2 Time of the k-th Arrival (Y_k)

The time Y_k of the k-th arrival is a continuous random variable. Its probability density function is derived by noting that the k-th arrival occurs in the infinitesimal interval $[y, y + \delta]$ if and only if exactly $k - 1$ arrivals occurred in $[0, y]$ and one arrival occurred in $[y, y + \delta]$. Y_k follows the **Erlang distribution**:

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \quad y \geq 0$$

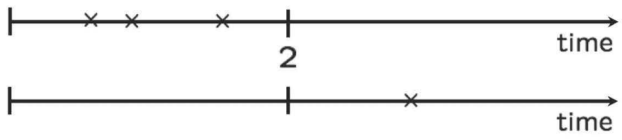
The Erlang distribution is also known as the distribution of the sum of k independent, identically distributed Exponential random variables.

4 Memorylessness and Interarrival Times

4.1 Memorylessness and Fresh-Start Property

The Poisson process exhibits a **memorylessness** and a **fresh-start property** which are analogous to those of the Bernoulli process. Given the relationship between the two processes, this is a plausible structural property.

- **Fresh-start at fixed time** t: If observation begins at time t, the process observed from that point forward is a Poisson process with rate λ, independent of the history until time t.
- **Interarrival Times**: Due to the memoryless nature of the Exponential distribution and the independent increments property, the time until the next arrival after any point t (or after any previous arrival Y_{k-1}) is still Exponential(λ).



- Thus, the interarrival times $T_k = Y_k - Y_{k-1}$ for $k \geq 1$ are **i.i.d. Exponential(λ)**.

Since Y_k is the sum of k i.i.d. Exponential(λ) random variables:

$$E[Y_k] = k/\lambda$$

$$\text{var}(Y_k) = k/\lambda^2$$

The property of having i.i.d. Exponential interarrival times can serve as an **equivalent definition** of the Poisson process.

5 Example: Poisson Fishing

Consider the example of fishing, where fish are caught as a Poisson process with rate $\lambda = 0.6$ fish per hour. The stopping rule is: fish for two hours; if at least one fish is caught ($N_2 \geq 1$), stop; otherwise ($N_2 = 0$), continue until the first fish is caught (T_1). The parameter for the initial two hours is $\lambda\tau = 0.6 \times 2 = 1.2$.

5.1 Probabilities of Total Time

- **P(fish for more than two hours)**: This occurs if and only if no fish are caught in the first two hours, $N_2 = 0$.
$$P(\text{Time} > 2) = P(N_2 = 0) = e^{-1.2}$$

- **P(fish for more than two and less than five hours)**: This event is equivalent to the first fish arriving between 2 and 5 hours, $2 < T_1 < 5$.

$$P(2 < T_1 < 5) = P(T_1 > 2) - P(T_1 > 5) = e^{-2\lambda} - e^{-5\lambda} = e^{-1.2} - e^{-3.0}$$

5.2 Expected Values

- **E[future fishing time | already fished for three hours]**: Due to memorylessness, the elapsed time of 3 hours is irrelevant, and the expected remaining time until the next fish is simply the mean interarrival time, $E[T_1] = 1/\lambda$.

$$E[\text{future time}] = 1/0.6 = 5/3 \text{ hours}$$

- **E[total fishing time T_{total}]**: Using the Law of Total Expectation:

$$E[T_{\text{total}}] = E[T_{\text{total}} \mid N_2 \geq 1]P(N_2 \geq 1) + E[T_{\text{total}} \mid N_2 = 0]P(N_2 = 0)$$

$$E[T_{\text{total}}] = 2 \cdot (1 - e^{-1.2}) + E[T_1 \mid T_1 > 2] \cdot e^{-1.2}$$

By memorylessness, $E[T_1 \mid T_1 > 2] = 2 + E[T_1] = 2 + 1/\lambda = 2 + 5/3 = 11/3$.

$$E[T_{\text{total}}] = 2(1 - e^{-1.2}) + (11/3)e^{-1.2} = 2 + \frac{5}{3}e^{-1.2}$$

- **E[total number of fish N_{total}]**: The total number of fish is $N_{\text{total}} = N_2 + I_{\{N_2=0\}}$, where $I_{\{N_2=0\}}$ is the indicator that the fishing continued (i.e., exactly one fish was caught during the continuation period).

$$E[N_{\text{total}}] = E[N_2] + P(N_2 = 0) = \lambda\tau + e^{-\lambda\tau}$$

$$E[N_{\text{total}}] = 1.2 + e^{-1.2}$$

Lecture 26: The Poisson Process Part II

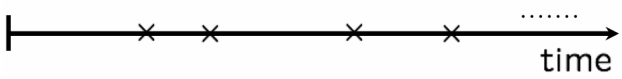
Instructor: Prof. Abolfazl Hashemi

1 The Sum of Independent Poisson Random Variables

This section explores a fundamental property regarding the closure of the Poisson distribution under summation. The analysis begins by considering the number of arrivals in consecutive, disjoint time intervals within a single Poisson process.

Consider a Poisson process with a rate of λ (for simplicity, we initially set $\lambda = 1$).

- Let M be the number of arrivals in an initial interval of length τ. M follows a Poisson distribution with parameter $E[M] = \lambda\tau$.
- Let N be the number of arrivals in a subsequent, consecutive interval of length ν. N follows a Poisson distribution with parameter $E[N] = \lambda\nu$.



Due to the property of **independent increments** inherent in the Poisson process, M and N are **independent** random variables.

The sum $M + N$ represents the total number of arrivals in the entire combined interval of length $\tau + \nu$. This total count must also follow a Poisson distribution, as the combined interval itself is subject to the original process over a duration $\tau + \nu$. Thus, $E[M + N] = \lambda(\tau + \nu) = \lambda\tau + \lambda\nu$.

General Theorem The **sum of independent Poisson random variables** with means (or parameters) μ and ν is a Poisson random variable with mean (or parameter) $\mu + \nu$. This result holds for any number of independent Poisson random variables.

$$M \sim \text{Poisson}(\mu), \quad N \sim \text{Poisson}(\nu), \quad M, N \text{ independent} \implies M + N \sim \text{Poisson}(\mu + \nu)$$

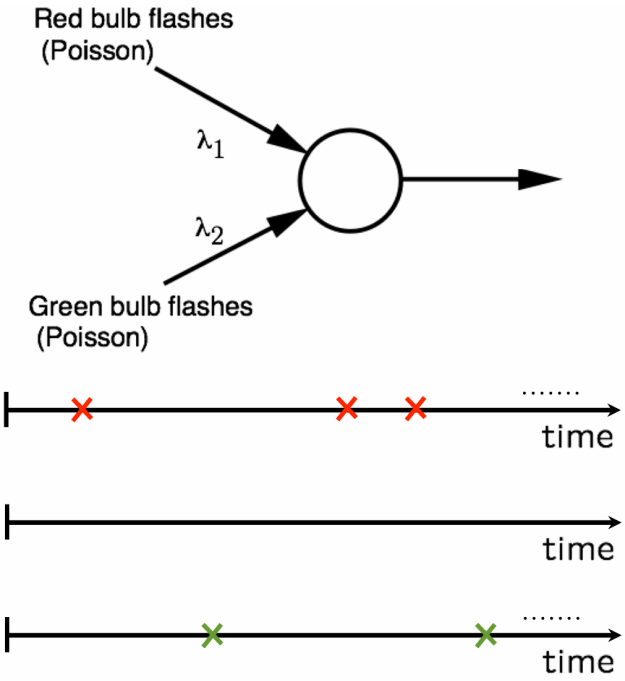
2 Merging of Independent Poisson Processes

The stability of the Poisson process is evident when multiple independent processes are combined or merged.

Consider two independent Poisson processes, Process 1 (Red bulb flashes) with rate λ_1 and Process 2 (Green bulb flashes) with rate λ_2. The **merged process** counts any arrival from either source.

Derivation via Small Intervals In a very small time interval δ:

- $P(\text{Red arrival}) \approx \lambda_1\delta$.
- $P(\text{Green arrival}) \approx \lambda_2\delta$.
- $P(\text{Simultaneous arrival}) \approx (\lambda_1\delta)(\lambda_2\delta) = O(\delta^2)$.



The probability of at least one arrival in the merged process is the sum of the individual probabilities, neglecting the vanishing probability of collision:

$$P(\text{Merged arrival}) \approx P(\text{Red}) + P(\text{Green}) \approx (\lambda_1 + \lambda_2)\delta$$

Since the probability of a merged arrival in δ is proportional to the time duration δ, and successive intervals are independent, the merged process is a **Poisson process** with rate $\lambda_{\text{merged}} = \lambda_1 + \lambda_2$.

2.1 Source of the Arrival

When an arrival is observed in the merged process, it is important to determine its source. The probability that an arbitrary arrival comes from Process 1 (Red) is independent of the arrival time t and is given by the ratio of the rates:

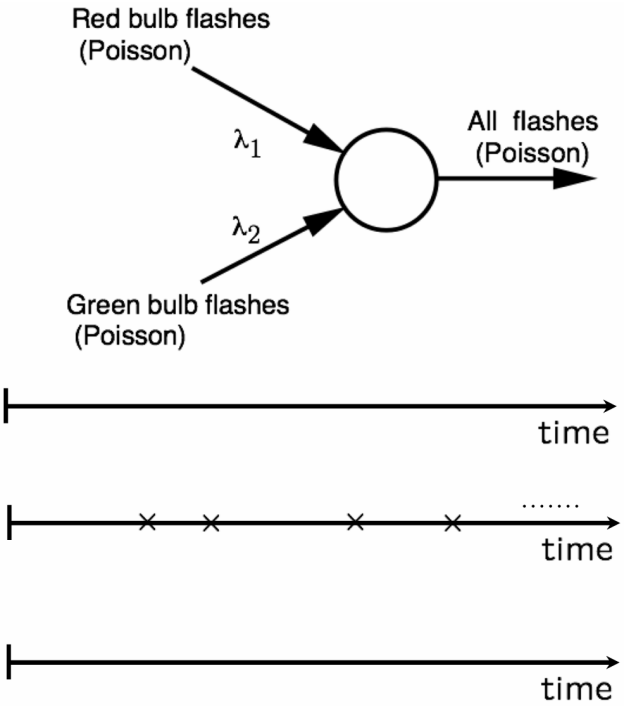
$$P(\text{Red} \mid \text{arrival}) = \frac{\text{Rate of Red arrivals}}{\text{Rate of All arrivals}} = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

This result holds for any specific arrival, irrespective of its position in the sequence. Consequently, the classification of whether an arrival is Red or Green is **independent for different arrivals**.

If one observes n arrivals in the merged process, the number of those arrivals that originated from the Red process follows a **Binomial distribution** with parameters n and $p = \frac{\lambda_1}{\lambda_1+\lambda_2}$.

Numerical Example Suppose λ_1 is the rate of Red flashes and λ_2 is the rate of Green flashes. The probability that the k-th arrival is Red is $p = \frac{\lambda_1}{\lambda_1+\lambda_2}$. The probability that exactly 4 out of the first 10 arrivals are Red is given by the Binomial PMF for $n = 10, k = 4$:

$$P(4 \text{ Red arrivals}) = \binom{10}{4} p^4 (1-p)^6$$



3 Applications of Merging: Reliability

3.1 Time until the First Burnout

Consider three lightbulbs with independent lifetimes X, Y, Z, where $X \sim$ Exponential(λ_X), $Y \sim$ Exponential(λ_Y), and $Z \sim$ Exponential(λ_Z). The exponential distribution characterizes the time until the first event (or arrival) in a Poisson process.

The time until the **first burnout** is $M = \min\{X, Y, Z\}$. Since the merged process is Poisson($\lambda_X + \lambda_Y + \lambda_Z$), the time to the first event in the merged process, $\min\{X, Y, Z\}$, is Exponential($\lambda_X + \lambda_Y + \lambda_Z$).

The expected time until the first burnout is the mean of this Exponential distribution:

$$E[\min\{X, Y, Z\}] = \frac{1}{\lambda_X + \lambda_Y + \lambda_Z}$$

3.2 Time until the Last Burnout

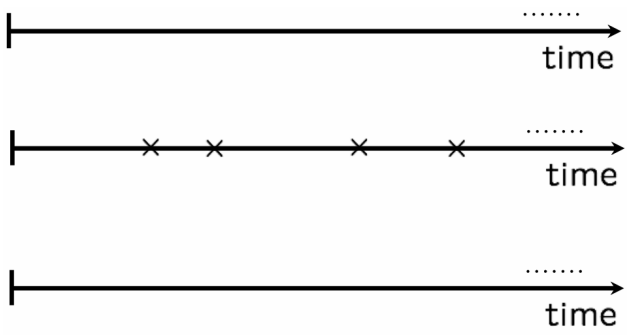
Finding the expected time until **all** bulbs burn out, $E[\max\{X, Y, Z\}]$, is a more complex calculation.

For the symmetric case where $\lambda_X = \lambda_Y = \lambda_Z = \lambda$, the expected time until the last burnout is found by relating the maximum to sums of independent Exponential random variables:

$$E[\max\{X, Y, Z\}] = \frac{1}{\lambda} + \frac{1}{2\lambda} + \frac{1}{3\lambda} = \frac{6 + 3 + 2}{6\lambda} = \frac{11}{6\lambda}$$

4 Splitting of a Poisson Process

A reverse operation to merging is splitting, where the arrivals of a single Poisson process are partitioned into multiple streams.



Consider a Poisson process with rate λ. Each arrival is classified into Stream 1 with an independent probability q, and into Stream 2 with probability $1-q$.

The resulting streams possess remarkable properties:

1. **Resulting Processes are Poisson**: Stream 1 is a Poisson process with rate $\lambda_1 = \lambda q$. Stream 2 is a Poisson process with rate $\lambda_2 = \lambda(1-q)$.
2. **Independence**: The two resulting streams are **independent** of each other.

This surprising independence simplifies the analysis of complex systems where external events are processed into different service queues.

5 Random Incidence and the Inspection Paradox

5.1 The Paradox Defined

Consider a Poisson process that has been running indefinitely, with interarrival times T_k that are i.i.d. Exponential(λ). The expected length of a typical interarrival interval is $E[T_k] = 1/\lambda$.

The **random incidence** problem (or the inspection paradox) arises when one attempts to measure the length of these intervals by arriving at an arbitrary random time t^* and measuring the length of the interval $V-U$ during which t^* falls.

Numerical Example If the interarrival times T_k have a mean of $1/\lambda = 15$ minutes ($\lambda = 4$/hour). When an observer shows up at a random time and measures the interval $V-U$, the average result is found to be $\mathbf{2}/\lambda = 30$ minutes, which is twice the expected length of a typical interval.

5.2 Analysis of Random Incidence

Let U be the time of the last arrival before the random arrival time t^*, and V be the time of the next arrival after t^*. $V-U$ is the observed interarrival interval.

- **Observed Length**: The interarrival interval $V-U$ seen by the random observer is **not** Exponential(λ).
- **Distribution of Observed Length**: For a Poisson process, the length of the interval containing the random time t^* follows an Erlang$(2,\lambda)$ distribution.
- **Expected Observed Length**: The expected length is $E[V-U] = 2/\lambda$.

5.3 The Source of the Bias

The paradox is a manifestation of **sampling bias**: when sampling at a random time, there is a proportional bias towards longer intervals. The observer is more likely to arrive during a longer interval than a shorter one.

General Renewal Process Example Consider a general renewal process where interarrival times T are i.i.d., equally likely to be 5 or 10 minutes.

- The expected value of a typical interarrival time is $E[T] = 0.5(5) + 0.5(10) = 7.5$ minutes.
- The probability of arriving during a 5-minute interval is proportional to its length: $P(\text{arrive during 5-minute i}$ $\frac{5}{5+10} = \frac{1}{3}$.
- The probability of arriving during a 10-minute interval is $P(\text{arrive during 10-minute interval}) = \frac{10}{5+10} = \frac{2}{3}$.
- The expected length of the interarrival interval during which you arrive is the length weighted by the probability of sampling it:

$$E[\text{Observed Length}] = 5 \cdot \frac{1}{3} + 10 \cdot \frac{2}{3} = \frac{25}{3} \approx 8.33 \text{ minutes}$$

The observed length (8.33 min) is greater than the true average length (7.5 min). This generalizes to all renewal processes (processes with i.i.d. interarrival times).

5.4 Implications for Sampling

The random incidence phenomenon illustrates that the **sampling method matters**.

- **Average family size**: Sampling a random family (uniformly) gives the true average. Sampling a random person's family is biased towards larger families.
- **Average bus occupancy**: Sampling a random bus (uniformly) gives the true average. Sampling a random passenger's bus is biased towards buses with higher occupancy.

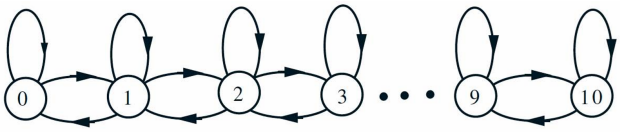
In each case, sampling based on the element (person, passenger) introduces a bias proportional to the size of the set (family size, bus occupancy).

Lecture 27: Markov Chains I

Instructor: Prof. Abolfazl Hashemi

1 Introduction to Markov Processes

This lecture introduces the concept of a **Markov process**, a specialized type of stochastic process characterized by the property of being memoryless. Specifically, we focus on **Discrete-Time Markov Chains (DTMCs)** with finite state spaces.



1.1 The Markov Property

A stochastic process $\{X_t, t \in T\}$ has the **Markov property** if the future state of the process, given the present state, is conditionally independent of the sequence of states that preceded it (the past). This is often summarized by the phrase: "Given the current state, the past doesn't matter."

Mathematically, for a discrete-time process X_n:

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j \mid X_n = i)$$

The state transition only depends on the value of the state at the current time n. The state at time $t+1$ is typically a function of the state at time t and some random noise:

$$\text{state}(t+1) = f(\text{state}(t), \text{noise})$$

2 Discrete-Time Finite State Markov Chains

2.1 Model Specification

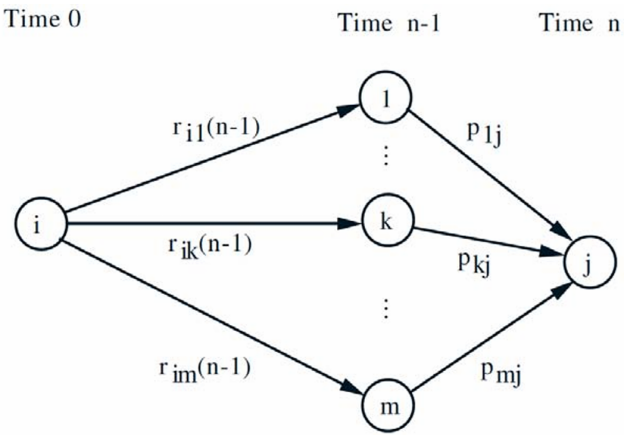
A discrete-time Markov chain is specified by identifying:

1. The set of **states** $\{1, 2, \ldots, m\}$, which is finite.
2. The **initial state** X_0, which may be given deterministically or defined by an initial probability distribution.
3. The **transition probabilities**.

The core probabilistic rule is the **one-step transition probability**, denoted p_{ij}:

$$p_{ij} = P(X_{n+1} = j \mid X_n = i)$$

If the chain is **time-homogeneous**, p_{ij} does not depend on the time step n.



2.2 Checkout Counter Example

The queue at a single checkout counter is a classic example of a DTMC.

- **Time**: Discrete time $n = 0, 1, 2, \ldots$ (e.g., minutes or service completions).
- **State** X_n: The number of customers in the system (waiting or being served) at time n. The state space is $S = \{0, 1, 2, \ldots, M\}$, where M is the maximum capacity (often large or infinite, but we consider a finite space for a finite state Markov chain).
- **Assumptions**:
  - Customer arrivals are a Bernoulli(p) process (one arrival with probability p, zero arrivals with probability $1 - p$).
  - Customer service times are Geometric(q) (one service completion with probability q, no completion with probability $1 - q$).
- **One-Step Transitions** $X_{n+1} \mid X_n = i$:
  - $\mathbf{X_{n+1} = i + 1}$: Arrival and no service completion. Occurs with probability $p(1 - q)$.
  - $\mathbf{X_{n+1} = i - 1}$: Service completion and no arrival. Occurs with probability $(1 - p)q$. (Requires $i \geq 1$).
  - $\mathbf{X_{n+1} = i}$: Either both an arrival and a service completion, OR neither event occurs. Occurs with probability $pq + (1 - p)(1 - q)$.

  (A similar but simpler analysis applies when the transitions are only ± 1 or 0).

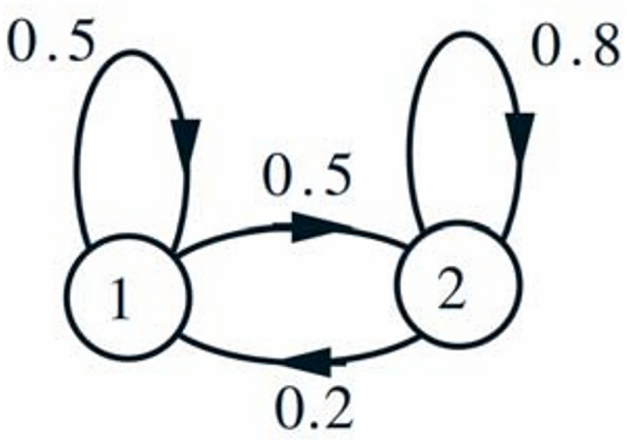
3 n-Step Transition Probabilities

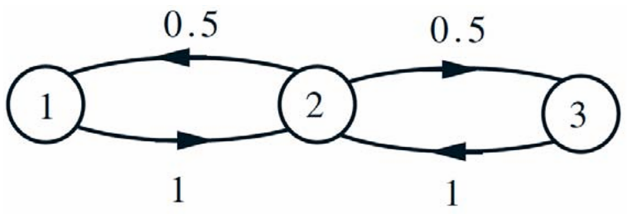
3.1 Definition and Recursion

The **n-step transition probability**, $r_{ij}(n)$, is the probability of moving from state i to state j in exactly n steps.

$$r_{ij}(n) = P(X_n = j | X_0 = i) = P(X_{n+s} = j | X_s = i)$$

The calculation of $r_{ij}(n)$ for $n > 1$ is achieved through the **Chapman-Kolmogorov equation** (a key recursion relation). To go from state i to state j in n steps, the chain must pass through some





intermediate state k after $n-1$ steps, followed by a single transition from k to j.

The recursive formula is:

$$r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1)p_{kj}$$

3.2 Unconditional State Probabilities

If the initial state X_0 is random with a known probability distribution $P(X_0 = i)$, the unconditional probability of the chain being in state j at time n is:

$$P(X_n = j) = \sum_{i=1}^{m} P(X_0 = i)r_{ij}(n)$$

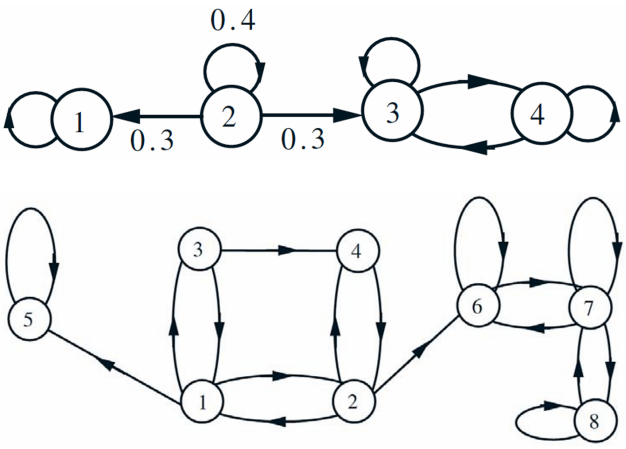
3.3 Example: Two-State Markov Chain Calculation

Consider a two-state Markov chain with the following transition probabilities:

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 0.5 & 0.5 \\ 0.8 & 0.2 \end{pmatrix}$$

We calculate the n-step transition probabilities $r_{ij}(n)$ for small n:

- $\mathbf{n = 0}$ (Initial State): $r_{11}(0) = 1, r_{12}(0) = 0, r_{21}(0) = 0, r_{22}(0) = 1$.
- $\mathbf{n = 1}$ (One Step): $r_{ij}(1) = p_{ij}$. Thus $r_{11}(1) = 0.5, r_{12}(1) = 0.5, r_{21}(1) = 0.8, r_{22}(1) = 0.2$.
- $\mathbf{n = 2}$ (Two Steps): Using the recursion $r_{ij}(2) = \sum_{k=1}^{2} r_{ik}(1)p_{kj}$:
  - $r_{11}(2) = r_{11}(1)p_{11} + r_{12}(1)p_{21} = (0.5)(0.5) + (0.5)(0.8) = 0.25 + 0.40 = 0.65$.



- $r_{12}(2) = r_{11}(1)p_{12} + r_{12}(1)p_{22} = (0.5)(0.5) + (0.5)(0.2) = 0.25 + 0.10 = 0.35.$
- $r_{21}(2) = r_{21}(1)p_{11} + r_{22}(1)p_{21} = (0.8)(0.5) + (0.2)(0.8) = 0.40 + 0.16 = 0.56.$
- $r_{22}(2) = r_{21}(1)p_{12} + r_{22}(1)p_{22} = (0.8)(0.5) + (0.2)(0.2) = 0.40 + 0.04 = 0.44.$

This calculation can be easily represented using matrix multiplication: $\mathbf{R}(n) = \mathbf{P}^n$.

4 Long-Term Behavior and Classification of States

4.1 Generic Convergence Questions

As the number of steps n approaches infinity, we analyze the long-term behavior of the chain.

- **Convergence**: Does $r_{ij}(n)$ converge to a limit as $n \to \infty$?
- **Stationarity**: Does the limit depend on the initial state i?
- **Example 1 (Non-convergence due to Periodicity)**: Consider a two-state chain where $p_{11} = 0, p_{12} = 1, p_{21} = 1, p_{22} = 0$. The chain oscillates between states 1 and 2, and $r_{11}(n)$ does not converge. For n odd, $r_{11}(n) = 0$; for n even, $r_{11}(n) = 1$.
- **Example 2 (Dependence on Initial State)**: Consider a chain with states $\{1, 2\}$ forming one isolated set and $\{3, 4\}$ forming another. If the chain starts in 1, it will never reach 3 or 4. The long-term distribution depends entirely on which set of states the chain begins in. For instance, $\lim_{n\to\infty} r_{13}(n) = 0$, but $\lim_{n\to\infty} r_{33}(n) > 0$.

4.2 Recurrent and Transient States

The long-term behavior of a Markov chain is determined by the classification of its states.

- **Recurrent State** i: A state i is recurrent if, once the process enters i, the probability of returning to i at some future time is 1. This means the process will return to i infinitely often.
- **Transient State** i: A state i is transient if there is a non-zero probability that the process, once having left i, will never return.

- **Recurrent Class**: A collection of recurrent states that communicate only among themselves. Once the process enters a recurrent class, it remains within that class forever.
- **Communication**: State i communicates with state j if $r_{ij}(n) > 0$ for some $n \geq 1$.

In the provided diagram:

- **Recurrent States/Classes**: State $\{5\}$ is a recurrent (absorbing) class. States $\{3, 4\}$ form a recurrent class. States $\{6, 7, 8\}$ form a recurrent class.
- **Transient States**: States 1 and 2 are transient. A process starting in state 1 or 2 will eventually transition to one of the recurrent classes and never return to 1 or 2.

Lecture 28: Markov Chains II

Instructor: Prof. Abolfazl Hashemi

1 Review and Warm-up

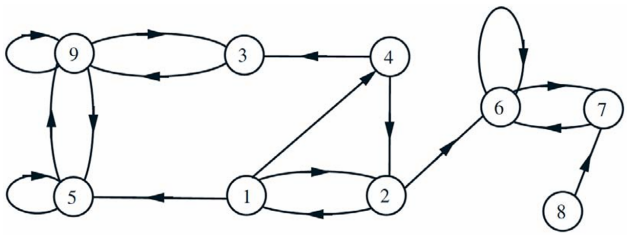
This lecture continues the study of Discrete-Time Markov Chains (DTMCs), focusing on long-term behavior, specifically the existence and calculation of steady-state probabilities.

1.1 Review of DTMC Fundamentals

A DTMC is defined by a discrete time index ($n = 0, 1, \ldots$), a discrete state space, and the **time-homogeneous** assumption, meaning the transition probabilities do not depend on the current time step n.

- **Transition Probabilities**: $p_{ij} = P(X_{n+1} = j \mid X_n = i)$.
- **Markov Property**: The future state X_{n+1} depends only on the current state X_n.
- **n-Step Transition Probability**: $r_{ij}(n) = P(X_n = j \mid X_0 = i)$.
- **Key Recursion (Chapman-Kolmogorov Equation)**: Calculates the n-step probability based on the $(n-1)$-step probability and the one-step transitions.

$$r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1)p_{kj}$$



1.2 Warm-up: Calculating Probabilities of Trajectories

Consider the Markov chain represented by the state diagram below. We assume the one-step transition probabilities p_{ij} are known.

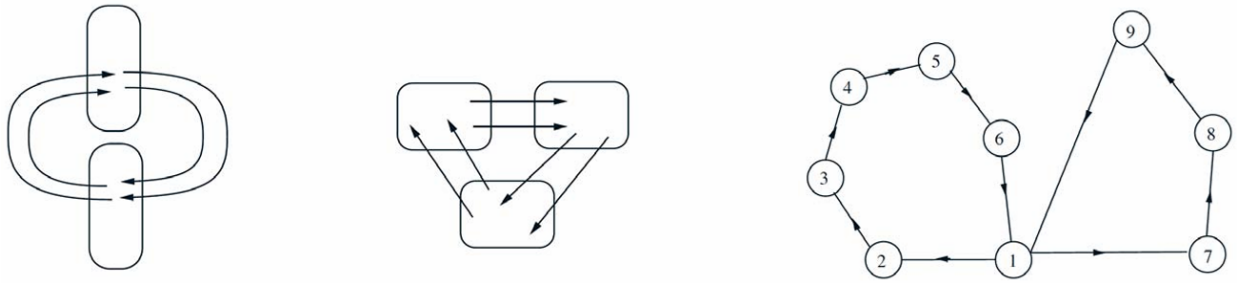
Probability of a Specific Trajectory The probability of a specific sequence of states over N steps, $X_0, X_1, \ldots, X_N$, is found by multiplying the probabilities of successive one-step transitions, utilizing the Markov property.

$$P(X_1 = 2, X_2 = 6, X_3 = 7 \mid X_0 = 1) = P(X_1 = 2 \mid X_0 = 1)P(X_2 = 6 \mid X_1 = 2)P(X_3 = 7 \mid X_2 = 6)$$

$$P(X_1 = 2, X_2 = 6, X_3 = 7 \mid X_0 = 1) = p_{12}p_{26}p_{67}$$

n-Step Probability The probability of reaching state $j = 7$ after $n = 4$ steps, starting from $i = 2$, is the n-step transition probability $r_{27}(4)$:

$$P(X_4 = 7 \mid X_0 = 2) = r_{27}(4)$$



2 Classification of States: Recurrent, Transient, and Periodic

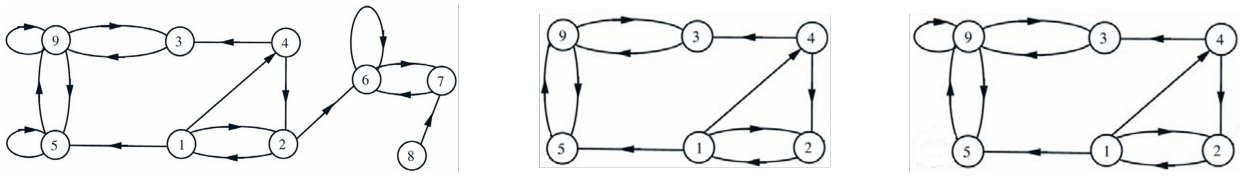
The classification of states is crucial for understanding the long-term behavior and for determining if a steady-state distribution exists.

2.1 Recurrent and Transient States

- **Recurrent State** i: A state i is recurrent if, starting from i, the probability of returning to i at some future time is one. In other words, the process will return to i infinitely often.
- **Transient State** i: A state i is transient if it is not recurrent. The process, once having left i, may never return.
- **Recurrent Class**: A collection of recurrent states that only communicate with each other. Communication means that a state i is reachable from state j, and vice versa. Once the process enters a recurrent class, it cannot leave.

2.2 Periodic States

A further classification of recurrent states involves periodicity, which impacts convergence to steady-state.



- **Periodic States**: The states in a recurrent class are **periodic** with period $d > 1$ if they can be partitioned into d groups, such that all transitions from one group lead deterministically to the next group, forming a cycle of length d.
- **Aperiodic State**: A state is aperiodic if the greatest common divisor of all possible return times to that state is 1. If a recurrent class contains a self-loop (a transition $p_{ii} > 0$), then the class is aperiodic.
- **Example 1 (Period $d = 2$)**: The top and bottom groups form a chain-link structure. The process must transition from the top group to the bottom group, and then back, meaning it returns to a state in the top group only in even numbers of steps.
- **Example 2 (Cycle)**: The sequence of states $1 \to 2 \to 3 \to \cdots \to 9 \to 1$ forms a cycle. If no self-loops exist, the period is $d = 9$.

3 Steady-State Behavior and Balance Equations

3.1 Convergence Theorem

The existence and uniqueness of a steady-state distribution, π_j, depends critically on the structural properties reviewed above.

- **Steady-State Probability** π_j: This is the long-run probability of finding the system in state j, regardless of the starting state i.

$$\pi_j = \lim_{n\to\infty} r_{ij}(n)$$

- **Theorem for Convergence**: The n-step transition probability $r_{ij}(n)$ converges to a unique limit π_j (i.e., a steady-state distribution exists and is independent of the initial state i) if the Markov chain satisfies two conditions:

  1. The recurrent states are all contained within a **single class** (i.e., the chain is irreducible and has no transient states).
  2. The single recurrent class is **not periodic** (i.e., the chain is aperiodic).

3.2 Balance Equations

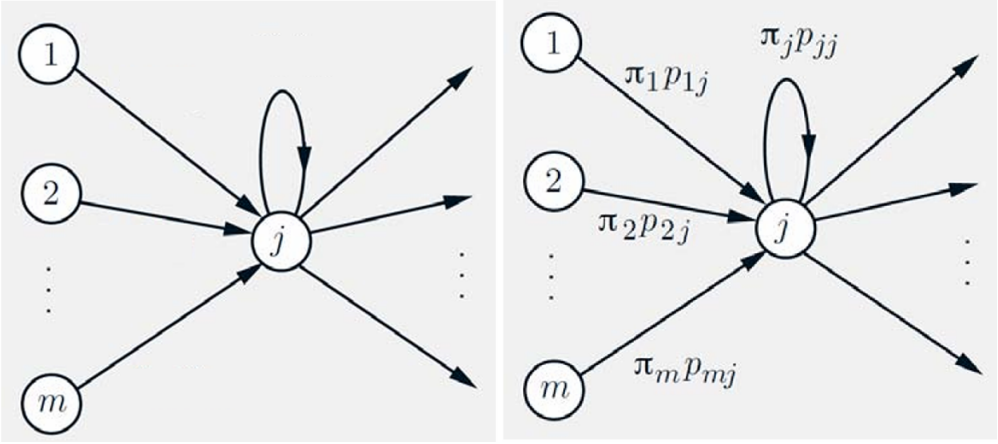
Assuming the chain converges to a steady-state distribution $\{\pi_j\}$, this distribution must satisfy the **balance equations** and the **normalization condition**.

Derivation from the Chapman-Kolmogorov Equation The key recursion equation, $r_{ij}(n) = \sum_k r_{ik}(n-1)p_{kj}$, holds for all n. By taking the limit as $n \to \infty$ on both sides:

$$\lim_{n\to\infty} r_{ij}(n) = \sum_k \left(\lim_{n\to\infty} r_{ik}(n-1)\right) p_{kj}$$

Substituting $\pi_j = \lim_{n\to\infty} r_{ij}(n)$:

$$\pi_j = \sum_k \pi_k p_{kj}$$



Interpretation of Balance Equations The balance equation $\pi_j = \sum_k \pi_k p_{kj}$ has a powerful physical interpretation based on frequency.

- π_j: The long-run frequency (or probability) of being in state j.
- $\sum_k \pi_k p_{kj}$: The long-run frequency of transitions **into** state j.
- $\pi_j \cdot 1$: The long-run frequency of transitions **out of** state j. (Since $\sum_k p_{jk} = 1$).

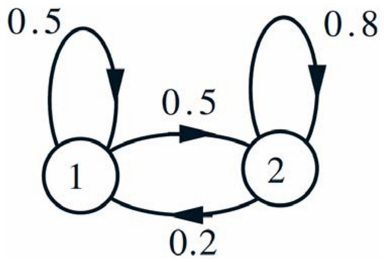
The equation states that in steady-state, the long-run frequency of entering state j must precisely equal the long-run frequency of leaving state j.

The Normalization Condition The set of balance equations provides m linear equations, but only $m-1$ of them are linearly independent. The final necessary equation is the **normalization condition**:

$$\sum_{j=1}^{m} \pi_j = 1$$

3.3 Example: Steady-State Calculation for Two-State Chain

Consider the two-state chain with transition matrix $P = \begin{pmatrix} 0.5 & 0.5 \\ 0.2 & 0.8 \end{pmatrix}$.



The chain is irreducible and aperiodic (since $p_{11} = 0.5 > 0$ and $p_{22} = 0.8 > 0$), so a steady-state distribution $\{\pi_1, \pi_2\}$ exists.

1. **Balance Equations** $\pi_j = \sum_k \pi_k p_{kj}$:

$$\pi_1 = \pi_1 p_{11} + \pi_2 p_{21} \implies \pi_1 = 0.5\pi_1 + 0.2\pi_2$$
$$\pi_2 = \pi_1 p_{12} + \pi_2 p_{22} \implies \pi_2 = 0.5\pi_1 + 0.8\pi_2$$

2. **Normalization Condition**:

$$\pi_1 + \pi_2 = 1$$

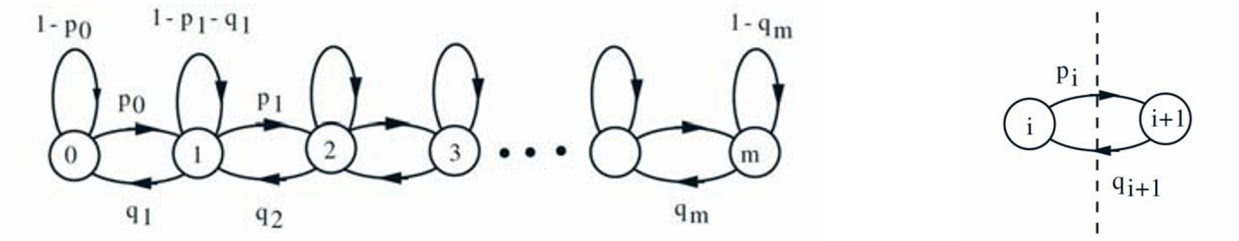
3. **Solving the System**: From the first balance equation:

$$0.5\pi_1 = 0.2\pi_2 \implies \pi_2 = \frac{0.5}{0.2}\pi_1 = 2.5\pi_1$$

Substitute $\pi_2 = 2.5\pi_1$ into the normalization equation:

$$\pi_1 + 2.5\pi_1 = 1 \implies 3.5\pi_1 = 1$$
$$\pi_1 = \frac{1}{3.5} = \frac{2}{7}$$
$$\pi_2 = 1 - \pi_1 = 1 - \frac{2}{7} = \frac{5}{7}$$

The steady-state distribution is $\{\pi_1 = 2/7, \pi_2 = 5/7\}$.



4 Birth-Death Processes

4.1 Definition and Local Balance

A **birth-death process** is a special case of a Markov chain where transitions are only allowed between adjacent states. In the context of queues, a "birth" is an arrival (transition $i \to i+1$), and a "death" is a service completion (transition $i \to i-1$).

The process is defined by state-dependent birth probabilities p_i (from $i \to i+1$) and death probabilities q_i (from $i \to i-1$). The total transition probability out of state i is $p_i + q_i + P(\text{stay in } i) = 1$.

For birth-death processes, the global balance equations $\pi_j = \sum_k \pi_k p_{kj}$ simplify to a system of **local balance equations** between adjacent states i and $i+1$:

$$\pi_i p_i = \pi_{i+1} q_{i+1} \text{ for } i = 0, 1, \ldots, m-1$$

This means the long-run frequency of transitions from state i to $i+1$ ($\pi_i p_i$) must equal the long-run frequency of transitions from state $i+1$ to i ($\pi_{i+1} q_{i+1}$).

4.2 Solving the Steady-State Distribution

The local balance equations allow for a recursive solution for all π_i in terms of π_0.

$$\pi_{i+1} = \pi_i \frac{p_i}{q_{i+1}}$$

Special Case: Constant Rates $p_i = p, q_i = q$ If the birth and death rates are constant (e.g., in a simple queue), let $\rho = p/q$. The relationship becomes $\pi_{i+1} = \pi_i \frac{p}{q} = \pi_i \rho$. This yields a **geometric distribution** for the steady-state probabilities:

$$\pi_i = \pi_0 \rho^i \text{ for } i = 0, 1, \ldots, m$$

For the case where the state space is infinite ($m \approx \infty$), a steady-state distribution only exists if the death rate exceeds the birth rate, $p < q$ (or $\rho < 1$).

- The initial probability π_0 is found by normalizing the infinite geometric series $\sum_{i=0}^{\infty} \pi_i = 1$:

$$\pi_0 = 1 - \rho$$

- The expected number of customers in the system in steady-state is calculated as the mean of the geometric distribution:

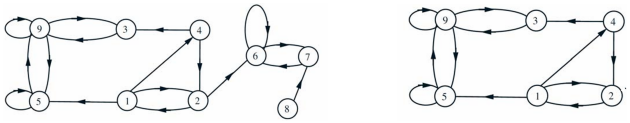
$$E[X_n] = \sum_{i=0}^{\infty} i\pi_i = \frac{\rho}{1-\rho}$$

Lecture 29: Markov Chains III

Instructor: Prof. Abolfazl Hashemi

1 Review of Steady-State Behavior

This lecture begins with a review of the conditions under which a Markov chain reaches a stationary, or steady-state, distribution, followed by an exploration of absorption probabilities and expected passage times, concepts critical for analyzing transient behavior.



1.1 Steady-State Convergence

A time-homogeneous, discrete-time Markov chain converges to a unique steady-state distribution $\{\pi_j\}$ if two conditions are met:

1. The recurrent states must form a single class (i.e., the chain must be irreducible among its recurrent states).
2. The single recurrent class must be aperiodic (i.e., the greatest common divisor of all possible return times to any state must be one).

When these conditions are satisfied, the n-step transition probability $r_{ij}(n)$ converges to the steady-state probability π_j, regardless of the initial state i:

$$\lim_{n\to\infty} r_{ij}(n) = \pi_j, \quad \forall i$$

The steady-state probabilities $\{\pi_j\}$ are found as the unique solution to the system of linear equations consisting of the **balance equations** and the **normalization condition**:

- **Balance Equations**: $\pi_j = \sum_k \pi_k p_{kj}, \quad j = 1, \ldots, m.$
- **Normalization**: $\sum_j \pi_j = 1.$

1.2 Use of Steady-State Probabilities in Long-Term Analysis

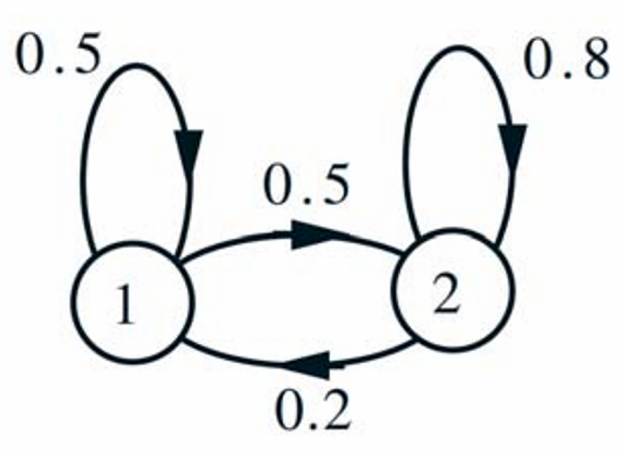
Once the steady-state probabilities $\{\pi_j\}$ are known, they allow for the calculation of long-term and conditional probabilities involving transitions far into the future. Consider the two-state chain example with steady-state probabilities $\pi_1 = 2/7$ and $\pi_2 = 5/7$ and one-step transition probabilities $p_{11} = 0.5$ and $p_{12} = 0.5$.

The analysis utilizes the Markov property and the convergence property:

- $P(X_1 = 1$ **and** $X_{100} = 1 \mid X_0 = 1)$: The process starts at state 1, goes to 1 in the first step, and returns to 1 at time 100.

$$P(X_1 = 1 \mid X_0 = 1) \cdot P(X_{100} = 1 \mid X_1 = 1) = p_{11} \cdot r_{11}(99) \approx p_{11}\pi_1$$

$$P(X_1 = 1 \text{ and } X_{100} = 1 \mid X_0 = 1) \approx 0.5 \cdot \frac{2}{7} = \frac{1}{7}$$



- $P(X_{100} = 1$ **and** $X_{101} = 2 \mid X_0 = 1)$: The chain is in state 1 at time 100, and transitions to state 2 at time 101.

$$P(X_{100} = 1 \mid X_0 = 1) \cdot P(X_{101} = 2 \mid X_{100} = 1) \approx \pi_1 \cdot p_{12}$$

$$P(X_{100} = 1 \text{ and } X_{101} = 2 \mid X_0 = 1) \approx \frac{2}{7} \cdot 0.5 = \frac{1}{7}$$

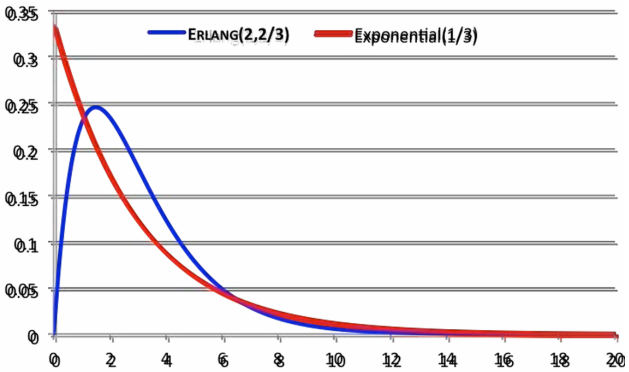
- $P(X_{100} = 1$ **and** $X_{200} = 1 \mid X_0 = 1)$: The chain is in state 1 at time 100, and returns to 1 at time 200 (100 steps later).

$$P(X_{100} = 1 \mid X_0 = 1) \cdot P(X_{200} = 1 \mid X_{100} = 1) \approx \pi_1 \cdot \pi_1$$

$$P(X_{100} = 1 \text{ and } X_{200} = 1 \mid X_0 = 1) \approx \frac{2}{7} \cdot \frac{2}{7} = \frac{4}{49}$$

2 Application: Design of a Phone System (Erlang B)

A practical application of Markov chains is modeling the capacity of a phone system, often referred to as the Erlang B model, which is used to determine the probability of a call being blocked.



2.1 Model and Discrete-Time Approximation

The system is modeled as a Birth-Death process, where the state i is the number of active calls, and the total capacity is B.

- **Call Arrivals**: Modeled as a Poisson process with rate λ.
- **Call Durations**: Modeled as independent Exponential random variables with parameter μ.

Using a discrete-time approximation with small time slots δ:

- The probability of a new call arriving is approximated as $\lambda\delta$.
- If there are i active calls, the probability of a call ending (a departure) is the superposition of i independent Exponential distributions, approximated as $i\mu\delta$.

The resulting DTMC has states $i \in \{0, 1, \ldots, B\}$, where state B is the maximum capacity.

2.2 Steady-State and Blocking Probability

The steady-state distribution $\{\pi_i\}$ satisfies the local balance equations, equating the flow of transitions between adjacent states $i-1 \leftrightarrow i$:

$$\pi_i \cdot (i\mu) = \pi_{i-1} \cdot \lambda \quad \text{for } i = 1, \ldots, B$$

This leads to the following expression for π_i in terms of π_0:

$$\pi_i = \pi_0 \frac{\lambda}{\mu} \frac{\lambda}{(2\mu)} \cdots \frac{\lambda}{(i\mu)} = \pi_0 \frac{(\lambda/\mu)^i}{i!} = \pi_0 \frac{\rho^i}{i!}$$

where $\rho = \lambda/\mu$ is the traffic intensity (measured in Erlangs).

The normalization constant π_0 is found by summing all π_i to 1:

$$\pi_0 = \left(\sum_{i=0}^{B} \frac{\rho^i}{i!} \right)^{-1}$$

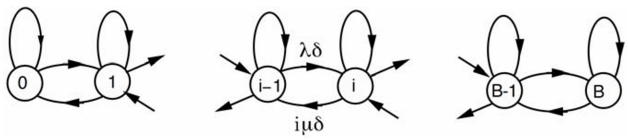
The key design parameter is the probability that an arriving customer finds the system busy, which occurs when all B lines are occupied (state B):

$$P(\text{arriving customer finds busy system}) = \pi_B = \frac{\frac{\rho^B}{B!}}{\sum_{i=0}^{B} \frac{\rho^i}{i!}}$$

This formula is known as the **Erlang B formula** and is used to determine the necessary capacity B for a target blocking probability.

3 Calculating Absorption Probabilities

Absorption probabilities are calculated for Markov chains with one or more **absorbing states**. An absorbing state k is a recurrent state where $p_{kk} = 1$, meaning once the process enters k, it never leaves.



3.1 Absorption Probability a_i

Let a_i be the probability that the chain eventually reaches a specific absorbing state k, given it started in transient state i.

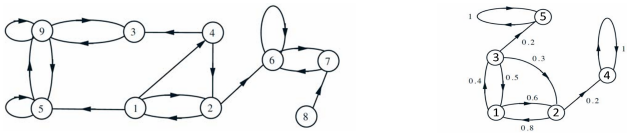
- **Boundary Conditions**: For any absorbing state k':

$$a_k = 1 \quad \text{(Target state)}$$

$$a_{k'} = 0 \quad \text{(Other absorbing states)}$$

- **Governing Equation (for transient states i)**: For any transient state i, the chain must transition to some state j in the next step, and from there, the probability of reaching k is a_j. The a_i values form a system of linear equations:

$$a_i = \sum_j p_{ij} a_j$$



3.2 Example Application of Absorption Probability

Consider a chain with absorbing states 4 and 5 (i.e., $p_{44} = 1$ and $p_{55} = 1$), and let a_i be the probability of absorption into state 4.

- $a_4 = 1$.
- $a_5 = 0$.

The equations for any transient states (which would be the remaining states, say 1, 2, 3, etc., if $p_{ii} < 1$ for those) are set up and solved simultaneously with the boundary conditions to find the unique solution.

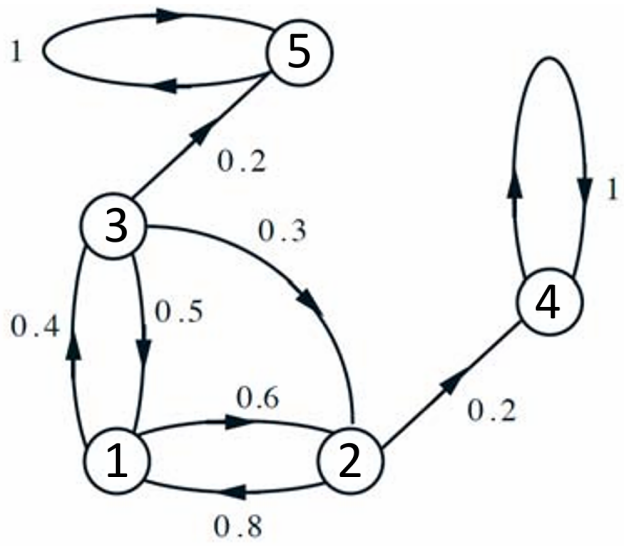
4 Expected Time to Absorption and Passage Times

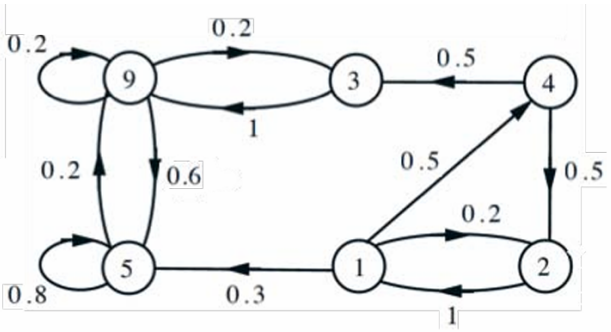
4.1 Expected Time to Absorption (μ_i)

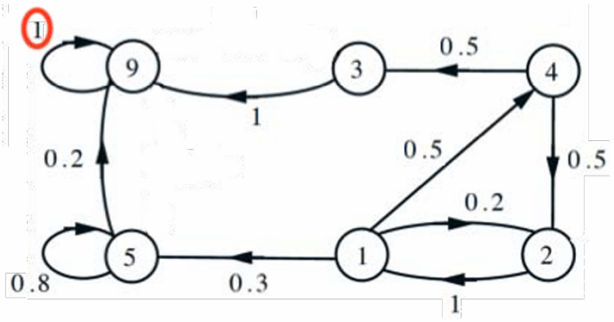
Let μ_i be the expected number of transitions until the chain reaches an absorbing state k, given the chain started in state i.

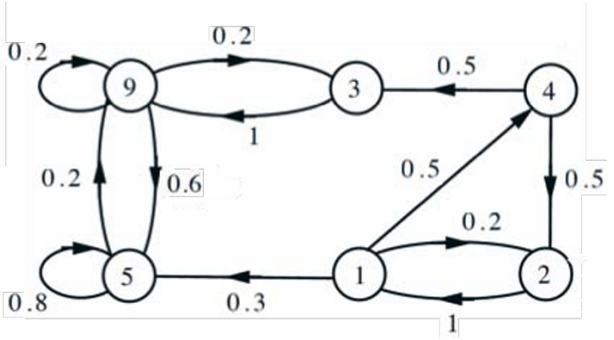
- **Boundary Conditions**: $\mu_k = 0$ (expected time is zero if already in the absorbing state).
- **Governing Equation (for transient states i)**: The time is 1 (for the first step) plus the expected future time, which depends on the state j entered after the first step:

$$\mu_i = 1 + \sum_j p_{ij} \mu_j$$


1
5
0.2
3
0.3
1
4
0.4
0.5
0.6
0.2
1
2
0.8


0.2
0.2
9
3
0.5
4
1
0.2
0.6
0.5
0.5
0.2
5
1
2
0.8
0.3
1


1
9
3
0.5
4
1
0.2
0.5
0.5
0.2
5
1
2
0.8
0.3
1


0.2
0.2
9
3
0.5
4
1
0.2
0.6
0.5
0.5
0.2
5
1
2
0.8
0.3
1

4.2 Mean First Passage and Recurrence Times

The concepts extend to recurrent chains that do not have explicit absorbing states.

- **Mean First Passage Time (t_i from i to s)**: The expected number of steps until the process reaches state s for the first time, given $X_0 = i$. This calculation treats the target state s as an absorbing state for the purpose of setting up the boundary condition ($t_s = 0$).
  - $t_s = 0$.
  - $t_i = 1 + \sum_j p_{ij} t_j, \quad \text{for all } i \neq s.$
- **Mean Recurrence Time (t_s^* of s)**: The expected number of steps until the process returns to state s, given $X_0 = s$. This is the mean first passage time from s back to s.

$$t_s^* = E[\min\{n \geq 1 \text{ such that } X_n = s\} | X_0 = s]$$

$$t_s^* = 1 + \sum_j p_{sj} t_j$$

For an irreducible and aperiodic Markov chain, the mean recurrence time t_s^* is the reciprocal of the steady-state probability π_s:

$$\pi_s = \frac{1}{t_s^*}$$

5 The Gambler's Ruin Example

The classic **Gambler's Ruin** problem provides an excellent context for applying both absorption probability and expected time to absorption calculations.

A gambler starts with i dollars and bets \$1 in a fair game (win or lose with probability 0.5) until she reaches 0 dollars (ruin/loss) or n dollars (win).

- **States**: $\{0, 1, \ldots, n\}$.
- **Absorbing States**: 0 (loss) and n (win).
- **Transition Probabilities (Fair Game)**: $p_{i,i+1} = 0.5$ and $p_{i,i-1} = 0.5$ for $0 < i < n$.

5.1 Absorption Probability (a_i: Win)

Let a_i be the probability that the gambler ends up with n dollars (wins the game), starting from i dollars.

- **Boundary Conditions**: $a_0 = 0$ and $a_n = 1$.
- **Governing Equation**: $a_i = 0.5a_{i-1} + 0.5a_{i+1}$ for $0 < i < n$.
- **Solution (Fair Game)**: The probability of winning is simply proportional to the initial stake relative to the target:

$$a_i = \frac{i}{n}$$

The expected wealth at the end of the game is $0 \cdot (1 - a_i) + n \cdot a_i = n \cdot (i/n) = i$. This demonstrates the principle that in a fair game, the expected final wealth equals the initial wealth.

5.2 Expected Time in the Game (μ_i)

Let μ_i be the expected number of plays until the game ends (absorption into state 0 or n), starting from i dollars.

- **Boundary Conditions**: $\mu_0 = 0$ and $\mu_n = 0$.
- **Governing Equation**: $\mu_i = 1 + 0.5\mu_{i-1} + 0.5\mu_{i+1}$ for $0 < i < n$.
- **Solution (Fair Game)**: The expected number of plays is:

$$\mu_i = i(n - i)$$

Lecture 30: Stationarity and Ergodicity

Instructor: Prof. Abolfazl Hashemi

1 Introduction to Random Processes

A **Random Process** (or stochastic process), denoted $X(t)$, is a family or collection of random variables indexed by time t. Unlike a single random variable, a random process is defined across an entire domain of time.

- For a fixed time t_0, $X(t_0)$ is a single random variable.
- A specific realization of the process, $x(t)$, is called a **sample function** or a time series.

The complete probabilistic description of a random process is given by its set of all **finite-dimensional joint distributions** for any set of time instants $t_1, t_2, \ldots, t_n$:

$$F_{X(t_1),\ldots,X(t_n)}(x_1, \ldots, x_n) = P(X(t_1) \leq x_1, \ldots, X(t_n) \leq x_n)$$

2 Stationarity: Time Invariance of Statistics

Stationarity refers to the property that the statistical characteristics of a random process do not change over time. Different degrees of time invariance lead to different definitions of stationarity.

2.1 Strict-Sense Stationarity (SSS)

A random process $X(t)$ is **Strict-Sense Stationary (SSS)** if its finite-dimensional joint distributions are invariant to any shift in time.

2.1.1 Definition

For any number of observations n, any set of time instants $t_1, \ldots, t_n$, and any arbitrary time shift τ:

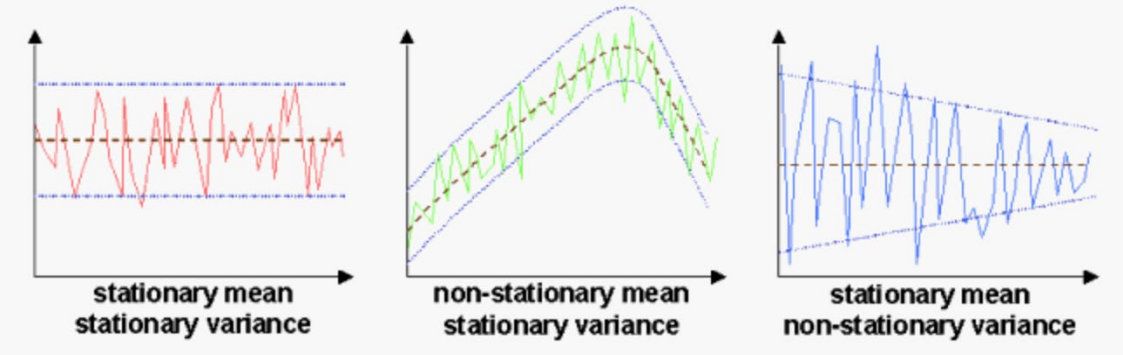
$$F_{X(t_1),\ldots,X(t_n)}(x_1, \ldots, x_n) = F_{X(t_1+\tau),\ldots,X(t_n+\tau)}(x_1, \ldots, x_n)$$

This is the strongest form of stationarity, implying that all statistical moments (mean, variance, skewness, etc.) are time-invariant.

2.1.2 Necessary Conditions Implied by SSS

By setting specific values for n and τ, SSS implies necessary conditions on the moments:

1. **First-order moment** ($n = 1$): The mean of the process must be a constant, independent of time t.
$$E[X(t)] = E[X(t + \tau)] = \mu_X \quad \text{(constant)}$$
2. **Second-order moment** ($n = 2$): The **autocorrelation function** depends only on the time difference $\tau = t_2 - t_1$.
$$R_X(t_1, t_2) = E[X(t_1)X(t_2)] = E[X(t_1 + \tau)X(t_2 + \tau)] = R_X(t_2 - t_1) = R_X(\tau)$$



2.1.3 Numerical Example: Process that is NOT SSS

Consider the process $X(t) = A\sin(\omega_0 t)$, where A is a random variable $A \sim U[0,1]$. We examine the mean $E[X(t)]$:

$$E[X(t)] = E[A\sin(\omega_0 t)] = E[A]E[\sin(\omega_0 t)]$$

The expected value of A is $E[A] = \int_0^1 a \cdot 1 da = 1/2$. Substituting this back yields:

$$E[X(t)] = (1/2)\sin(\omega_0 t)$$

Since the mean $E[X(t)]$ is a sinusoidal function of time t, it is not constant. Because a constant mean is a necessary condition for SSS, the process $X(t)$ is **not SSS**. This example demonstrates that any time-dependent structure in the deterministic components of the process violates stationarity.

2.2 Wide-Sense Stationarity (WSS)

Wide-Sense Stationarity (WSS) is a weaker and more commonly used form of stationarity that only restricts the first two statistical moments.

2.2.1 Definition

A random process $X(t)$ is WSS if and only if:

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 (lag) $\tau = t_2 - t_1$:

$$R_X(t_1, t_2) = R_X(t_2 - t_1) = R_X(\tau)$$

2.2.2 Relationship between SSS and WSS

- SSS $\implies$ WSS, provided the first two moments exist.
- WSS $\not\implies$ SSS, as WSS does not impose constraints on higher-order moments (e.g., the process could have a time-varying third moment).

2.2.3 WSS and Gaussian Processes

There is one critical exception to the WSS $\not\Longrightarrow$ SSS rule: a **Gaussian Process**. A Gaussian process is entirely defined by its first two moments (mean and covariance/autocorrelation). Therefore, for a Gaussian process, the WSS condition is sufficient to guarantee SSS:

$$\text{WSS} \overset{\text{Gaussian}}{\Longleftrightarrow} \text{SSS}$$

2.2.4 Numerical Example: WSS Check

Consider the process $X(t) = A\cos(\omega_0 t) + B\sin(\omega_0 t)$, where A and B are independent random variables distributed as $N(0, \sigma^2)$.

1. **Check Mean ($E[X(t)]$):**

   $$E[X(t)] = E[A]\cos(\omega_0 t) + E[B]\sin(\omega_0 t)$$

   Since A and B are zero-mean ($E[A] = E[B] = 0$):

   $$E[X(t)] = 0 \cdot \cos(\omega_0 t) + 0 \cdot \sin(\omega_0 t) = 0$$

   The mean is constant. **Condition 1 satisfied**.

2. **Check Autocorrelation ($R_X(t_1, t_2)$):**

   $$R_X(t_1, t_2) = E[X(t_1)X(t_2)]$$

   $$R_X(t_1, t_2) = E[(A\cos(\omega_0 t_1) + B\sin(\omega_0 t_1))(A\cos(\omega_0 t_2) + B\sin(\omega_0 t_2))]$$

   Since A and B are independent and uncorrelated ($E[AB] = E[A]E[B] = 0$):

   $$R_X(t_1, t_2) = E[A^2]\cos(\omega_0 t_1)\cos(\omega_0 t_2) + E[B^2]\sin(\omega_0 t_1)\sin(\omega_0 t_2)$$

   Since $E[A^2] = E[B^2] = \text{var}(A) = \sigma^2$:

   $$R_X(t_1, t_2) = \sigma^2[\cos(\omega_0 t_1)\cos(\omega_0 t_2) + \sin(\omega_0 t_1)\sin(\omega_0 t_2)]$$

   Using the trigonometric identity $\cos(a - b) = \cos a \cos b + \sin a \sin b$:

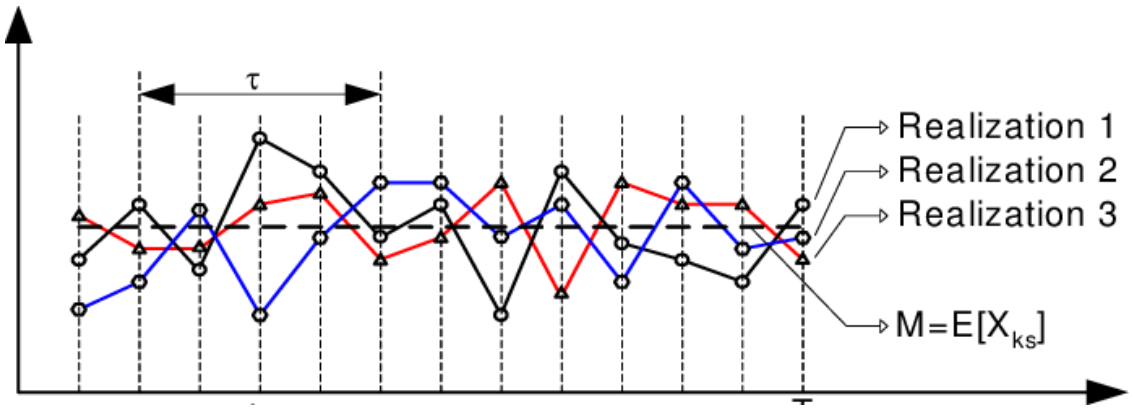
   $$R_X(t_1, t_2) = \sigma^2 \cos(\omega_0(t_2 - t_1))$$

   Letting $\tau = t_2 - t_1$, the autocorrelation depends only on the lag τ: $R_X(\tau) = \sigma^2\cos(\omega_0\tau)$. **Condition 2 satisfied**.

   Conclusion: The process $X(t)$ is **WSS**. Because A and B are Gaussian, $X(t)$ is also a Gaussian process, and thus it is also SSS.

3 Ergodicity: Time Averages vs. Ensemble Averages

Ergodicity addresses the relationship between averaging across multiple realizations of the process at a fixed time (the **ensemble average**) and averaging over a long time for a single realization (the **time average**).



3.1 Definition of Ergodicity in the Mean

A random process $X(t)$ is **Ergodic in the Mean** if the time average of a single realization converges to the ensemble mean (expected value) of the process.

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T X(t)dt = E[X(t)] \quad \text{(in probability or a.s.)}$$

If a process is ergodic, measuring a single, sufficiently long sample function can yield all the required statistical properties of the entire process.

3.1.1 Necessary Conditions for Ergodicity

Ergodicity is a much stronger requirement than stationarity.

- A process must be **WSS** to be ergodic.
- The process must be sufficiently "mixing," meaning the dependence between samples separated by a large time lag must vanish. A common sufficient condition is that the autocorrelation function must decay fast enough, specifically:

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T |R_X(\tau) - \mu_X^2|d\tau = 0$$

3.1.2 Numerical Example: Process that is NOT Ergodic

Consider the process $X(t) = A$, where A is a random variable $A \sim U[0,1]$.

1. **WSS Check**:
   - Mean: $E[X(t)] = E[A] = 1/2$ (Constant).
   - Autocorrelation: $R_X(t_1,t_2) = E[A^2] = \text{var}(A) + (E[A])^2 = 1/12 + 1/4 = 1/3$ (Constant, depends only on $\tau = 0$, so $R_X(\tau) = 1/3$). The process is **WSS**.
2. **Ergodicity Check**: The time average for a specific realization is:

$$\text{Time Average} = \lim_{T\to\infty} \frac{1}{T}\int_0^T A dt = A$$

   Since A is a random variable, the time average A is not necessarily equal to the ensemble mean $E[X(t)] = 1/2$. The time average depends on the specific outcome of the initial randomization

(A). **Conclusion**: Since the time average is a random variable and not equal to the constant ensemble mean, the process is **NOT Ergodic**. The process is not sufficiently mixing because its memory is infinite ($R_X(\tau)$ does not decay to μ_X^2).

4 Applications of Stationarity and Ergodicity

4.1 Application in Signal Processing and Spectral Analysis

4.1.1 WSS Assumption

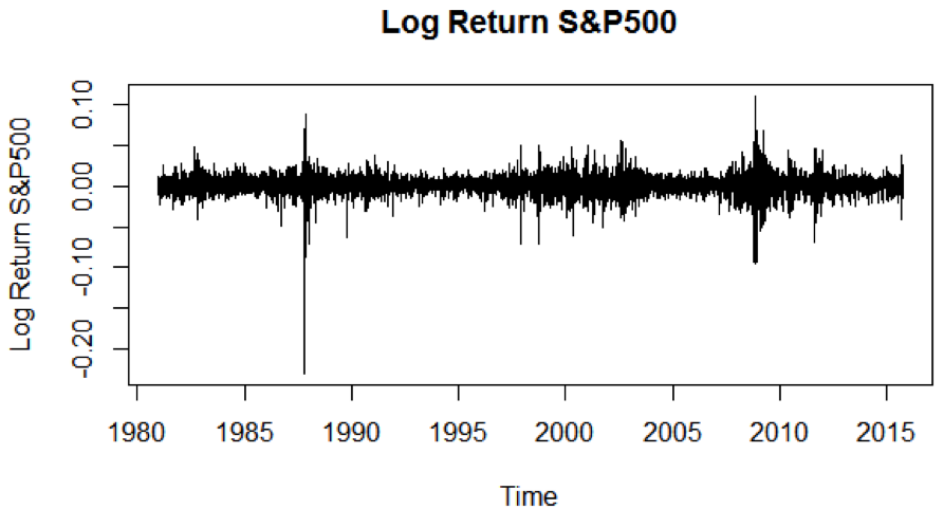
The assumption that a random signal is WSS is foundational in nearly all classical signal processing techniques, including filtering, estimation, and spectral analysis. This assumption greatly simplifies the mathematical models.

4.1.2 Power Spectral Density (PSD)

The concept of the Power Spectral Density (PSD), $S_X(\omega)$, is exclusively defined for WSS processes. The PSD describes how the power of a signal is distributed over frequency. The crucial link between the time domain and frequency domain for WSS processes is given by the Wiener-Khinchin Theorem:

$$S_X(\omega) = \mathcal{F}\{R_X(\tau)\} = \int_{-\infty}^{\infty} R_X(\tau) e^{-j\omega\tau} d\tau$$

Where $R_X(\tau)$ is the autocorrelation function. This theorem allows engineers to analyze the frequency content of stochastic signals purely from their autocorrelation structure.



4.2 Application in Financial Engineering

4.2.1 The Challenge of Non-Stationarity

Financial time series, such as stock returns $R(t)$ or volatility, pose a significant challenge because they exhibit high degrees of non-stationarity in the real world.

- The mean return (low SSS) may drift over long periods.
- The variance (volatility) of returns changes dramatically during periods of crisis or high market activity (low WSS).

Models that assume strict SSS (like the simplistic original random walk models) often fail to capture real market behavior.

4.2.2 Modeling Non-Stationary Volatility

Sophisticated models are required to handle this non-stationary structure. GARCH (Generalized Autoregressive Conditional Heteroskedasticity) models, popular in financial engineering, are designed to capture the clustering of volatility, where large changes in price tend to be followed by large changes, regardless of the sign. These models effectively assume that, while the process itself is non-stationary, the time-series of its innovations (shocks) might satisfy certain weaker stationarity conditions, often WSS, allowing for reliable forecasting of variance.

4.3 Application in Machine Learning (MCMC)

4.3.1 Ergodicity in Sampling

In Machine Learning and Computational Statistics, Markov Chain Monte Carlo (MCMC) methods are widely used to sample from complex, high-dimensional probability distributions (e.g., in Bayesian inference). These algorithms construct a Markov chain whose stationary distribution is the target distribution π.

4.3.2 Requirements for MCMC

For MCMC simulations to produce accurate results, the underlying Markov chain must be **ergodic**. An ergodic Markov chain ensures two critical properties:

1. **Convergence**: The chain must be irreducible (able to reach any state from any other) and aperiodic (not stuck in cycles) to guarantee convergence to the unique stationary distribution π_j.
2. **Averaging**: Ergodicity ensures that the time average computed from a single long run of the simulation (the sample mean of the simulated values) provides a reliable, consistent estimate of the true ensemble expected value under the target distribution π_j.

Without ergodicity, the simulation results would be dominated by the initial state or exhibit long-term biases, rendering them useless for estimating population characteristics.