Data Science for Mathematicians
Lecture 3: Probabilistic Foundations of Modeling

Abstract

This lecture transitions from the deterministic, geometric framework of Ordinary Least Squares (OLS) to the probabilistic foundations essential for modern data science. We begin by motivating the need for a probabilistic perspective to handle uncertainty and noise inherent in real-world data. The lecture provides a rigorous, measure-theoretic definition of random variables and introduces key probability distributions, including the Bernoulli distribution for binary classification tasks and the Gaussian distribution, whose ubiquity is justified by the Central Limit Theorem. We then develop the mathematical tools for summarizing these distributions through their moments: expectation, variance, and covariance, establishing a crucial link between the statistical concept of covariance and the geometric dot product. The core of the lecture introduces the principle of Maximum Likelihood Estimation (MLE) as a primary method for parameter estimation, culminating in a proof that OLS is a special case of MLE under the assumption of Gaussian errors. Finally, we present an alternative paradigm of statistical learning through an introduction to conditional probability and Bayes' Theorem, which forms the basis for Bayesian inference. This lecture establishes the foundational probabilistic language required for subsequent topics in model evaluation, statistical inference, and the bias-variance tradeoff.

Contents

1 Random Variables and Key Distributions

In our previous discussions, we approached data from a deterministic, geometric perspective. We conceptualized data points as vectors in $\mathbb{R}^p$, models as subspaces, and fitting as an act of orthogonal projection. This geometric framework is powerful, providing a clear and intuitive solution to the Ordinary Least Squares (OLS) problem. However, it operates under an implicit assumption of perfect, noise-free measurements. The real world, as we know, is rife with uncertainty, measurement error, and inherent randomness. To build models that are not only descriptive but also inferential—models that can quantify their own uncertainty and make predictions about unseen data—we must move beyond deterministic geometry and embrace the language of probability theory.

This section serves as the bridge. We will introduce the foundational concept of a **random variable**, which allows us to treat data not as fixed points, but as outcomes of a random process. We will then explore several key probability distributions that form the building blocks for a vast array of data science models. Our goal is to formalize the sources of uncertainty and lay the mathematical groundwork for estimating model parameters and evaluating their performance in a principled, statistical manner.

1.1 Measure-Theoretic Foundations

Before we can rigorously define random variables and probability spaces, we must first establish the mathematical machinery of **measure theory**. This framework, developed by Henri Lebesgue and others in the early 20th century, provides the foundation for modern probability theory and allows us to assign *sizes* to sets in a consistent and mathematically rigorous way.

1.1.1 σ-Algebras

The first concept we need is a way to specify which subsets of a given set are *measurable*—that is, which subsets we can meaningfully assign a measure (or probability) to. This is captured by the notion of a σ-algebra.

Definition 1.1 (σ-Algebra)**.** Let Ω be a non-empty set. A collection $\mathcal{F}$ of subsets of Ω is called a **σ-algebra** (or **σ-field**) on Ω if it satisfies the following three properties:

1. **Contains the whole space:** $\Omega \in \mathcal{F}$.
2. **Closed under complementation:** If $A \in \mathcal{F}$, then $A^c = \Omega \setminus A \in \mathcal{F}$.
3. **Closed under countable unions:** If $A_1, A_2, A_3, \ldots \in \mathcal{F}$, then $\displaystyle\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Remark 1.2. From these axioms, several important consequences follow:

- The empty set $\emptyset \in \mathcal{F}$ (since $\emptyset = \Omega^c$).

- $\mathcal{F}$ is closed under countable intersections (by De Morgan's laws: $\bigcap_{i=1}^{\infty} A_i = (\bigcup_{i=1}^{\infty} A_i^c)^c$).
- $\mathcal{F}$ is closed under finite unions and intersections.
- $\mathcal{F}$ is closed under set differences: if $A, B \in \mathcal{F}$, then $A \setminus B = A \cap B^c \in \mathcal{F}$.

Example 1.3 (Trivial and Discrete σ-Algebras)**.** For any non-empty set Ω:

- The **trivial σ-algebra** $\mathcal{F} = \{\emptyset, \Omega\}$ is the smallest σ-algebra on Ω.
- The **power set** $\mathcal{F} = 2^{\Omega}$ (the collection of all subsets of Ω) is the largest σ-algebra on Ω, often called the **discrete σ-algebra**.

Example 1.4 (A Non-trivial σ-Algebra)**.** Let $\Omega = \{1, 2, 3, 4\}$ and let $A = \{1, 2\}$. The σ-algebra generated by A is

$$\mathcal{F} = \{\emptyset, \{1, 2\}, \{3, 4\}, \{1, 2, 3, 4\}\}$$

This is the smallest σ-algebra containing A. Note that while $\{1\} \subset \Omega$, we have $\{1\} \notin \mathcal{F}$—not every subset needs to be in the σ-algebra.

1.1.2 Borel Sets and the Borel σ-Algebra

When working with the real numbers $\mathbb{R}$, we need a σ-algebra that contains all the *reasonable* sets we might want to measure. The power set of $\mathbb{R}$ is too large and leads to pathological constructions (such as non-measurable sets). Instead, we use the **Borel σ-algebra**.

Definition 1.5 (Borel σ-Algebra)**.** The **Borel σ-algebra** on $\mathbb{R}$, denoted $\mathcal{B}(\mathbb{R})$, is the smallest σ-algebra on $\mathbb{R}$ that contains all open intervals (a, b) for $a < b$. Equivalently, it is the σ-algebra generated by the open subsets of $\mathbb{R}$.

Definition 1.6 (Borel Set)**.** Any set belonging to the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ is called a **Borel set**.

Remark 1.7. The Borel σ-algebra $\mathcal{B}(\mathbb{R})$ contains:

- All open sets and all closed sets in $\mathbb{R}$.
- All intervals: open (a, b), closed $[a, b]$, half-open $[a, b)$ and $(a, b]$, and rays $(-\infty, a)$, (a, ∞), etc.
- All countable sets (including singletons $\{x\}$, the integers $\mathbb{Z}$, and the rationals $\mathbb{Q}$).
- Countable unions, countable intersections, and complements of all the above.

In practice, virtually every subset of $\mathbb{R}$ that arises in applications is a Borel set.

Example 1.8 (Common Borel Sets)**.** The following are all Borel sets:

- The interval $[0, 1]$ (closed set).
- The set of rational numbers $\mathbb{Q} = \bigcup_{q \in \mathbb{Q}} \{q\}$ (countable union of singletons).
- The Cantor set (constructed via countable intersections of closed sets).
- The set $\{x \in \mathbb{R} : \sin(x) > 0\}$ (a countable union of open intervals).

1.1.3 Measures and the Lebesgue Measure

With a σ-algebra specifying which sets are measurable, we can now define a **measure**—a function that assigns a non-negative *size* to each measurable set.

Definition 1.9 (Measure)**.** Let Ω be a set and $\mathcal{F}$ a σ-algebra on Ω. A function $\mu : \mathcal{F} \to [0, \infty]$ is called a **measure** if it satisfies:

1. **Non-negativity:** $\mu(A) \geq 0$ for all $A \in \mathcal{F}$.
2. **Null empty set:** $\mu(\emptyset) = 0$.
3. **Countable additivity (σ-additivity):** If $A_1, A_2, A_3, \ldots \in \mathcal{F}$ are pairwise disjoint (i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$), then
$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

The triple $(\Omega, \mathcal{F}, \mu)$ is called a **measure space**.

The most important measure on $\mathbb{R}$ is the **Lebesgue measure**, which generalizes the intuitive notion of *length* to a much larger class of sets.

Definition 1.10 (Lebesgue Measure)**.** The **Lebesgue measure** λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is the unique measure satisfying:

1. For any interval $I = (a, b)$, $[a, b]$, $[a, b)$ or $(a, b]$ with $a \leq b$,
$$\lambda(I) = b - a.$$
2. λ is **translation-invariant**, i.e., for any Borel set A and any $t \in \mathbb{R}$,
$$\lambda(A + t) = \lambda(A), \quad \text{where } A + t = \{a + t : a \in A\}.$$

Example 1.11 (Lebesgue Measure of Common Sets)**.** Consider the following

- $\lambda([2, 5]) = 5 - 2 = 3$ (the *length* of the interval).
- $\lambda(\{x\}) = 0$ for any singleton (a point has zero length).
- $\lambda(\mathbb{Q} \cap [0, 1]) = 0$ (countable sets have measure zero).
- $\lambda([0, 1] \setminus \mathbb{Q}) = 1$ (the irrationals in $[0, 1]$ have full measure).

1.1.4 The Lebesgue Integral

Having defined the Lebesgue measure, we now turn to the **Lebesgue integral**—a generalization of the Riemann integral that is essential for modern probability theory. The Lebesgue integral allows us to integrate a broader class of functions and provides the rigorous foundation for defining expectation.

The key conceptual difference between the two approaches lies in how they partition the domain:

- The **Riemann integral** partitions the *domain* (the x-axis) into small intervals and approximates the function by its values on each interval.
- The **Lebesgue integral** partitions the *range* (the y-axis) and asks: "For each value y, how much of the domain maps to approximately y?"

To build up the Lebesgue integral, we start with simple functions.

Definition 1.12 (Simple Function)**.** A function $\phi : \Omega \to \mathbb{R}$ is called a **simple function** if it takes only finitely many values. Any simple function can be written as

$$\phi = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}$$

where $a_1, \ldots, a_n \in \mathbb{R}$ are the distinct values of ϕ, the sets $A_i = \phi^{-1}(\{a_i\}) = \{\omega \in \Omega : \phi(\omega) = a_i\}$ are pairwise disjoint and cover Ω, and $\mathbf{1}_{A_i}$ is the **indicator function** (or characteristic function) of A_i

$$\mathbf{1}_{A_i}(\omega) = \begin{cases} 1 & \text{if } \omega \in A_i, \\ 0 & \text{if } \omega \notin A_i. \end{cases}$$

Definition 1.13 (Lebesgue Integral of a Simple Function)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $\phi = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}$ be a non-negative simple function where each $A_i \in \mathcal{F}$. The **Lebesgue integral** of ϕ with respect to μ is defined as

$$\int_\Omega \phi \, \mathrm{d}\mu = \sum_{i=1}^{n} a_i \cdot \mu(A_i).$$

Example 1.14 (Integral of a Simple Function)**.** Let $\Omega = [0, 3]$ with Lebesgue measure λ, and define the simple function:

$$\phi(x) = \begin{cases} 2 & \text{if } x \in [0, 1), \\ 5 & \text{if } x \in [1, 2), \\ 1 & \text{if } x \in [2, 3]. \end{cases}$$

Then $\phi = 2 \cdot \mathbf{1}_{[0,1)} + 5 \cdot \mathbf{1}_{[1,2)} + 1 \cdot \mathbf{1}_{[2,3]}$, and

$$\int_{[0,3]} \phi \, \mathrm{d}\lambda = 2 \cdot \lambda([0, 1)) + 5 \cdot \lambda([1, 2)) + 1 \cdot \lambda([2, 3]) = 2 \cdot 1 + 5 \cdot 1 + 1 \cdot 1 = 8.$$

This matches the *area under the step function*, which the Riemann integral would also give.

For general non-negative measurable functions, we define the integral as a supremum over simple function approximations.

Definition 1.15 (Lebesgue Integral of a Non-negative Function)**.** Let $f : \Omega \to [0, \infty]$ be a measurable function. The **Lebesgue integral** of f is

$$\int_\Omega f \, \mathrm{d}\mu = \sup \left\{ \int_\Omega \phi \, \mathrm{d}\mu : \phi \text{ is simple, } 0 \leq \phi \leq f \right\}.$$

For functions that can take both positive and negative values, we decompose into positive and negative parts.

Definition 1.16 (Lebesgue Integral of a General Function)**.** For a measurable function $f : \Omega \to \mathbb{R}$, define its **positive part** $f^+ = \max(f, 0)$ and **negative part** $f^- = \max(-f, 0)$. Note that $f = f^+ - f^-$ and $|f| = f^+ + f^-$. If at least one of $\int_\Omega f^+ \, \mathrm{d}\mu$ or $\int_\Omega f^- \, \mathrm{d}\mu$ is finite, we define:

$$\int_\Omega f \, \mathrm{d}\mu = \int_\Omega f^+ \, \mathrm{d}\mu - \int_\Omega f^- \, \mathrm{d}\mu$$

If both integrals are finite (equivalently, if $\int_\Omega |f| \, \mathrm{d}\mu < \infty$), we say f is **Lebesgue integrable** (or $f \in L^1(\mu)$).

Theorem 1.17 (Riemann vs. Lebesgue Integration)**.** *Let $f : [a, b] \to \mathbb{R}$ be a bounded function. Then:*

1. *If f is Riemann integrable on $[a, b]$, then f is Lebesgue integrable, and the two integrals agree:*

$$\int_a^b f(x)\,\mathrm{d}x = \int_{[a,b]} f\,\mathrm{d}\lambda$$

2. *f is Riemann integrable if and only if f is continuous almost everywhere (i.e., the set of discontinuities has Lebesgue measure zero).*

Thus, the Lebesgue integral extends *the Riemann integral to a strictly larger class of functions.*

Example 1.18 (A Function That Is Lebesgue but Not Riemann Integrable)**.** The **Dirichlet function $\mathbf{1}_{\mathbb{Q}}$** on $[0, 1]$ is defined by:

$$\mathbf{1}_{\mathbb{Q}}(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases}$$

This function is discontinuous at every point, so it is *not* Riemann integrable. However, it is Lebesgue integrable:

$$\int_{[0,1]} \mathbf{1}_{\mathbb{Q}}\,\mathrm{d}\lambda = 1 \cdot \lambda(\mathbb{Q} \cap [0, 1]) + 0 \cdot \lambda([0, 1] \setminus \mathbb{Q}) = 1 \cdot 0 + 0 \cdot 1 = 0$$

The integral is zero because the rationals, despite being dense, have measure zero.

Example 1.19 (Computing a Lebesgue Integral via Riemann Integration)**.** For "nice" functions (continuous or piecewise continuous), we can compute Lebesgue integrals using familiar Riemann techniques. Consider $f(x) = x^2$ on $[0, 2]$:

$$\int_{[0,2]} x^2\,\mathrm{d}\lambda = \int_0^2 x^2\,\mathrm{d}x = \left[\frac{x^3}{3}\right]_0^2 = \frac{8}{3}$$

The Lebesgue framework does not change how we compute integrals of well-behaved functions; it simply provides a more robust theoretical foundation.

Example 1.20 (Lebesgue Integral with Respect to a Discrete Measure)**.** Let $\Omega = \{1, 2, 3\}$ and define the counting measure μ by $\mu(\{k\}) = 1$ for each k. For any function $f : \Omega \to \mathbb{R}$, it yields

$$\int_\Omega f\,\mathrm{d}\mu = f(1) \cdot \mu(\{1\}) + f(2) \cdot \mu(\{2\}) + f(3) \cdot \mu(\{3\}) = f(1) + f(2) + f(3)$$

This shows that *summation is a special case of Lebesgue integration* with respect to a discrete measure.

Remark 1.21 (Connection to Probability)**.** A **probability measure** is simply a measure $\mathbb{P}$ with the additional property that $\mathbb{P}(\Omega) = 1$. This normalization constraint is what distinguishes probability theory from general measure theory. All the properties of measures (countable additivity, monotonicity, continuity from below/above) carry over to probability measures.

With this measure-theoretic foundation in place, we are now equipped to rigorously define probability spaces and random variables.

1.2 A Rigorous View of Random Variables

At a high level, a random variable is a variable whose value is a numerical outcome of a random phenomenon. For a mathematician, this description lacks precision. To be rigorous, we must define it as a mapping from a sample space to the real numbers, a mapping that preserves the structure of the underlying probability space.

Definition 1.22 (Probability Space)**.** A **probability space** is a measure space $(\Omega, \mathcal{F}, \mathbb{P})$ where:

- Ω is the **sample space**, the set of all possible outcomes of an experiment.
- $\mathcal{F}$ is a **σ-algebra** (or Borel field) on Ω, which is a collection of subsets of Ω (called events) that is closed under complement, countable union, and countable intersection. It represents the set of all events to which we can assign a probability.
- $\mathbb{P}$ is a **probability measure**, a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ that assigns a probability to each event, satisfying the Kolmogorov axioms.

To make this abstract definition concrete, consider the following three examples. In each case, we identify the sample space Ω, the collection of events $\mathcal{F}$, and the probability measure $\mathbb{P}$.

Example 1.23 (Fair Coin Toss)**.** Consider flipping a fair coin once. The sample space is $\Omega = \{H, T\}$, representing heads and tails. Since Ω is finite, we take $\mathcal{F}$ to be the **power set** of Ω (the set of all subsets):

$$\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$$

Each element of $\mathcal{F}$ is an event we can ask about. For instance, $\{H\}$ is the event "the coin lands heads." The probability measure assigns: $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\{H\}) = 0.5$, $\mathbb{P}(\{T\}) = 0.5$, and $\mathbb{P}(\{H, T\}) = 1$. Notice that $\mathbb{P}(\Omega) = \mathbb{P}(\{H, T\}) = 1$, as required—something must happen.

Example 1.24 (Rolling a Six-Sided Die)**.** When rolling a fair die, the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. Again, we use the power set as our collection of events, so $\mathcal{F}$ contains $2^6 = 64$ subsets. Some example events include:

- $\{6\}$: "rolling a six," with $\mathbb{P}(\{6\}) = 1/6$.
- $\{2, 4, 6\}$: "rolling an even number," with $\mathbb{P}(\{2, 4, 6\}) = 3/6 = 1/2$.
- $\{1, 2, 3\}$: "rolling at most three," with $\mathbb{P}(\{1, 2, 3\}) = 1/2$.

For a fair die, the probability measure is uniform: each singleton outcome has probability 1/6, and compound events have probabilities equal to the sum of their constituent outcomes.

Example 1.25 (Drawing a Card from a Standard Deck)**.** Consider drawing one card from a well-shuffled standard deck of 52 cards. The sample space Ω consists of all 52 cards, and $\mathcal{F}$ is the power set with 2^{52} possible events. We can define events such as:

- $A = \{\text{all hearts}\}$: "drawing a heart," with $\mathbb{P}(A) = 13/52 = 1/4$.
- $B = \{\text{all face cards}\}$: "drawing a face card (J, Q, K)," with $\mathbb{P}(B) = 12/52 = 3/13$.
- $A \cap B$: "drawing a heart that is also a face card," with $\mathbb{P}(A \cap B) = 3/52$.

This example illustrates how $\mathcal{F}$ allows us to define compound events through set operations (unions, intersections, complements), and the probability measure respects these operations in a consistent way.

The key insight from these examples is that $\mathcal{F}$ defines precisely which questions we are allowed to ask about the probability of outcomes, and $\mathbb{P}$ provides consistent answers to all such questions.

With this machinery, we can formally define a random variable.

Definition 1.26 (Random Variable)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A **random variable** X is a function $X : \Omega \to \mathbb{R}$ such that for every Borel set $B \subset \mathbb{R}$, the set $\{\omega \in \Omega \mid X(\omega) \in B\}$ is an event in $\mathcal{F}$. This property is known as **measurability**.

The measurability condition is technically vital. It ensures that the inverse image of any well-behaved set in our target space ($\mathbb{R}$) corresponds to a set in our source space (Ω) for which we have a defined probability. In essence, it allows us to *pull back* questions about the numerical value of X (e.g., "What is the probability that $X \leq x$?") and answer them using the probability measure $\mathbb{P}$ on the original sample space Ω. This gives rise to the notion of the **distribution** of a random variable, which describes the probability that X will take on a certain value or fall within a certain range of values.

The following examples illustrate how random variables serve as the bridge between abstract outcomes and numerical values we can analyze.

Example 1.27 (Coin Toss as a Random Variable)**.** Recall our coin toss with $\Omega = \{H, T\}$. We can define a random variable $X : \Omega \to \mathbb{R}$ by assigning numerical values to each outcome:

$$X(H) = 1, \quad X(T) = 0$$

This converts the abstract outcome "heads" into the number 1 and "tails" into 0. Now we can compute probabilities in terms of X: for instance, $\mathbb{P}(X = 1) = \mathbb{P}(\{H\}) = 0.5$. This encoding is precisely how we model binary outcomes (success/failure) in data science, and it leads directly to the Bernoulli distribution.

Example 1.28 (Sum of Two Dice)**.** Consider rolling two fair dice. The sample space is $\Omega = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}$, containing 36 equally likely outcomes. We can define a random variable $S : \Omega \to \mathbb{R}$ as the sum of the two faces:

$$S((i, j)) = i + j$$

For instance, $S((3, 4)) = 7$ and $S((1, 1)) = 2$. Notice that S is not one-to-one: multiple outcomes map to the same value. For example, $S = 7$ can occur via $(1, 6), (2, 5), (3, 4), (4, 3), (5, 2)$, or $(6, 1)$. Thus:

$$\mathbb{P}(S = 7) = \frac{6}{36} = \frac{1}{6}$$

This illustrates how a random variable can *collapse* many outcomes into a single numerical summary.

Example 1.29 (Winnings in a Card Game)**.** Suppose you draw one card from a standard deck and receive a payout based on the card: \$10 for an Ace, \$5 for a face card (J, Q, K), and \$0 otherwise. We define the random variable $W : \Omega \to \mathbb{R}$ representing your winnings:

$$W(\omega) = \begin{cases} 10 & \text{if } \omega \text{ is an Ace} \\ 5 & \text{if } \omega \text{ is a face card} \\ 0 & \text{otherwise} \end{cases}$$

Now we can ask probabilistic questions about W. There are 4 Aces and 12 face cards in the deck, so:

$$\mathbb{P}(W = 10) = \frac{4}{52} = \frac{1}{13}, \quad \mathbb{P}(W = 5) = \frac{12}{52} = \frac{3}{13}, \quad \mathbb{P}(W = 0) = \frac{36}{52} = \frac{9}{13}.$$

This example shows how random variables translate real-world quantities (money, measurements, counts) into a mathematical framework for analysis.

We primarily distinguish between two types of random variables based on the nature of their range:

- A **discrete random variable** has a range that is finite or countably infinite (e.g., the integers $\{0, 1, 2, \ldots\}$). We can meaningfully ask for the probability of each specific outcome.
- A **continuous random variable** has a range that is uncountably infinite, typically an interval on the real line. For such variables, the probability of any single exact value is zero; instead, we speak of probabilities over intervals.

To characterize the distributions of these two types of random variables, we introduce the Probability Mass Function and the Probability Density Function.

Definition 1.30 (Probability Mass Function)**.** Let X be a discrete random variable taking values in a countable set $S = \{x_1, x_2, \ldots\}$. The **Probability Mass Function (PMF)** of X, denoted $p_X : S \to [0, 1]$, is the function that assigns to each possible value $x \in S$ the probability that X equals x:

$$p_X(x) = \mathbb{P}(X = x)$$

A valid PMF must satisfy:

1. Non-negativity: $p_X(x) \geq 0$ for all $x \in S$.
2. Normalization: $\sum_{x \in S} p_X(x) = 1$.

Example 1.31 (Number of Defective Items)**.** A quality control inspector randomly selects 3 items from a production line. Let X denote the number of defective items found. Based on historical data, the PMF of X is given by:

$$p_X(x) = \begin{cases} 0.70 & \text{if } x = 0 \\ 0.20 & \text{if } x = 1 \\ 0.08 & \text{if } x = 2 \\ 0.02 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

We can verify this is a valid PMF: all values are non-negative, and $0.70 + 0.20 + 0.08 + 0.02 = 1$. Using this PMF, we can compute probabilities such as $\mathbb{P}(X \geq 2) = p_X(2) + p_X(3) = 0.08 + 0.02 = 0.10$.

Example 1.32 (Customer Arrivals)**.** A small coffee shop records the number of customers N arriving during each 5-minute interval. After analyzing many such intervals, the shop determines the following PMF:

$$p_N(n) = \begin{cases} 0.10 & \text{if } n = 0 \\ 0.25 & \text{if } n = 1 \\ 0.30 & \text{if } n = 2 \\ 0.20 & \text{if } n = 3 \\ 0.10 & \text{if } n = 4 \\ 0.05 & \text{if } n = 5 \\ 0 & \text{otherwise} \end{cases}$$

This is a valid PMF since all probabilities are non-negative and $0.10 + 0.25 + 0.30 + 0.20 + 0.10 + 0.05 = 1$. We can answer questions such as: What is the probability of having at least 3 customers? $\mathbb{P}(N \geq 3) = p_N(3) + p_N(4) + p_N(5) = 0.20 + 0.10 + 0.05 = 0.35$.

Definition 1.33 (Probability Density Function)**.** Let X be a continuous random variable. The **Probability Density Function (PDF)** of X, denoted $f_X : \mathbb{R} \to [0, \infty)$, is a function such that the probability of X falling within any interval $[a, b]$ is given by:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x$$

A valid PDF must satisfy:

1. Non-negativity: $f_X(x) \geq 0$ for all $x \in \mathbb{R}$.
2. Normalization: $\displaystyle\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$.

Note that $f_X(x)$ itself is not a probability; it is a density. Consequently, $f_X(x)$ can exceed 1.

Example 1.34 (Continuous Uniform Distribution)**.** Suppose the waiting time T (in minutes) for a bus is uniformly distributed between 0 and 10 minutes. The PDF of T is:

$$f_T(t) = \begin{cases} \frac{1}{10} = 0.1 & \text{if } 0 \leq t \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

This is a valid PDF since $f_T(t) \geq 0$ everywhere and $\int_{-\infty}^{\infty} f_T(t)\,\mathrm{d}t = \int_0^{10} 0.1\,\mathrm{d}t = 1$. To find the probability that a passenger waits between 2 and 5 minutes:

$$\mathbb{P}(2 \leq T \leq 5) = \int_2^5 0.1\,\mathrm{d}t = 0.1 \times (5 - 2) = 0.3$$

Note that the probability of waiting *exactly* 3 minutes is $\mathbb{P}(T = 3) = \int_3^3 f_T(t)\,\mathrm{d}t = 0$, illustrating that for continuous random variables, point probabilities are zero.

Example 1.35 (Exponential Distribution)**.** The lifetime L (in years) of a certain electronic component is modeled by an exponential distribution with rate parameter $\lambda = 0.5$. The PDF of L is

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

We verify normalization

$$\int_0^{\infty} 0.5e^{-0.5\ell}\,d\ell = \left[-e^{-0.5\ell}\right]_0^{\infty} = 0 - (-1) = 1.$$

To find the probability that the component lasts between 1 and 3 years:

$$\mathbb{P}(1 \leq L \leq 3) = \int_1^3 0.5e^{-0.5\ell}\,d\ell = \left[-e^{-0.5\ell}\right]_1^3 = -e^{-1.5} + e^{-0.5} \approx 0.384$$

This example also illustrates that $f_L(0) = 0.5 < 1$, but for distributions concentrated on small intervals, the PDF can exceed 1 at some points.

1.3 Core Discrete Distributions for Modeling

Discrete distributions are the natural choice for modeling count data or categorical outcomes. In data science, they are the foundation of classification models.

Before introducing specific distributions, we define a fundamental concept that appears throughout probability theory and statistics.

Definition 1.36 (Independent and Identically Distributed (i.i.d.))**.** A collection of random variables $X_1, X_2, \ldots, X_n$ is said to be **independent and identically distributed**, abbreviated **i.i.d.**, if the following two conditions hold:

1. **(Independence)** The random variables are mutually independent, meaning the outcome of any subset does not affect the probability distribution of the others. Formally, for any collection of measurable sets $A_1, \ldots, A_n$, it follows that

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).$$

2. **(Identically Distributed)** All random variables share the same probability distribution. That is, $X_1, X_2, \ldots, X_n$ all have the same CDF:

$$F_{X_1}(x) = F_{X_2}(x) = \cdots = F_{X_n}(x) \quad \text{for all } x.$$

Example 1.37 (i.i.d. Random Variables)**.** Consider rolling a fair six-sided die n times. Let X_i denote the outcome of the i-th roll. The sequence $X_1, X_2, \ldots, X_n$ is i.i.d. because

(1) each roll is physically independent of the others, and

(2) each X_i follows the same discrete uniform distribution on $\{1, 2, 3, 4, 5, 6\}$.

In contrast, consider drawing cards from a standard deck *without replacement.* If X_i denotes the value of the i-th card drawn, then $X_1, X_2, \ldots$ are *not* independent—knowing X_1 changes the probabilities for X_2. Thus, this sequence is not i.i.d.

The i.i.d. assumption is pervasive in statistics and machine learning. It simplifies analysis because it allows us to express joint probabilities as products and to apply powerful limit theorems such as the law of large numbers and the central limit theorem.

1.3.1 The Bernoulli Distribution

The simplest, yet arguably most important, discrete distribution is the Bernoulli distribution. It models a single trial with exactly two possible outcomes: success (1) or failure (0).

Definition 1.38 (Bernoulli Distribution)**.** A random variable X follows a **Bernoulli distribution** with parameter $p \in [0, 1]$, denoted $X \sim \text{Bernoulli}(p)$, if its PMF is given by:

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

Here, p represents the probability of success ($X = 1$).

Example 1.39 (Biased Coin)**.** Consider a biased coin that lands heads with probability $p = 0.7$. Let X denote the outcome of a single flip, where $X = 1$ represents heads and $X = 0$ represents tails. Then $X \sim \text{Bernoulli}(0.7)$, and the PMF is:

$$\mathbb{P}(X = 1) = 0.7, \quad \mathbb{P}(X = 0) = 0.3$$

Using the general formula: $\mathbb{P}(X = 1) = 0.7^1 \cdot 0.3^0 = 0.7$ and $\mathbb{P}(X = 0) = 0.7^0 \cdot 0.3^1 = 0.3$.

Example 1.40 (Free Throw Success)**.** A basketball player has a free throw success rate of 80%. Let X represent the outcome of a single free throw attempt, where $X = 1$ indicates a successful shot and $X = 0$ indicates a miss. Then $X \sim \text{Bernoulli}(0.8)$. We can compute:

$$\mathbb{P}(\text{success}) = \mathbb{P}(X = 1) = 0.8, \quad \mathbb{P}(\text{miss}) = \mathbb{P}(X = 0) = 0.2$$

The expected value is $\mathbb{E}[X] = 0 \cdot 0.2 + 1 \cdot 0.8 = 0.8 = p$, confirming that the expectation of a Bernoulli random variable equals its success probability.

Example 1.41 (Titanic Survival)**.** Recall the Titanic dataset. Let X_i be a random variable representing the survival status of passenger i. We can model this as $X_i \sim \text{Bernoulli}(p)$, where $X_i = 1$ if the passenger survived and $X_i = 0$ otherwise. The parameter p is the underlying probability of survival for any given passenger. A naive model might estimate p as the overall survival rate in the dataset. A more sophisticated logistic regression model (which we will see in a later lesson) aims to model p_i as a function of passenger features (e.g., class, sex, age). Each passenger's survival is a draw from a Bernoulli distribution with a potentially unique success parameter.

1.3.2 The Binomial Distribution

The Binomial distribution is a direct extension of the Bernoulli. It describes the number of successes in a fixed number of i.i.d. Bernoulli trials.

Definition 1.42 (Binomial Distribution)**.** Let $X_1, X_2, \dots, X_n$ be n independent Bernoulli trials, each with success probability p. Let $Y = \sum_{i=1}^n X_i$. Then Y follows a **binomial distribution** with parameters n and p, denoted $Y \sim \text{Binomial}(n, p)$. Its PMF is given by the probability of observing exactly k successes in n trials:

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

The term $\displaystyle\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient, which counts the number of ways to arrange k successes among n trials.

Example 1.43 (Coin Flips)**.** Suppose we flip a fair coin $n = 5$ times. Let Y denote the number of heads obtained. Since each flip is an independent Bernoulli trial with $p = 0.5$, we have $Y \sim \text{Binomial}(5, 0.5)$. The probability of getting exactly 3 heads is:

$$\mathbb{P}(Y = 3) = \binom{5}{3} (0.5)^3 (0.5)^2 = 10 \cdot 0.125 \cdot 0.25 = 0.3125$$

Similarly, $\mathbb{P}(Y = 0) = \binom{5}{0}(0.5)^5 = 0.03125$ and $\mathbb{P}(Y = 5) = \binom{5}{5}(0.5)^5 = 0.03125$.

Example 1.44 (Multiple Free Throws)**.** Recall the basketball player with an 80% free throw success rate. If she attempts $n = 10$ free throws, let Y be the total number of successful shots. Then $Y \sim \text{Binomial}(10, 0.8)$. The probability of making exactly 8 shots is:

$$\mathbb{P}(Y = 8) = \binom{10}{8} (0.8)^8 (0.2)^2 = 45 \cdot 0.1678 \cdot 0.04 \approx 0.302$$

We can also compute

$$\mathbb{P}(Y \geq 9) = \mathbb{P}(Y = 9) + \mathbb{P}(Y = 10) = \binom{10}{9}(0.8)^9(0.2) + \binom{10}{10}(0.8)^{10} \approx 0.268 + 0.107 = 0.375.$$

Example 1.45 (Click-Through Rate)**.** Imagine an A/B test where we show $n = 1000$ users a new website design (Version B) and we want to model the number of users who click a specific button. Let Y be the number of clicks. We can model this as $Y \sim \text{Binomial}(1000, p_B)$, where p_B is the unknown click-through probability for Version B. If we observe $k = 50$ clicks, our data is $Y = 50$. The task of statistical inference is to use this observation to estimate p_B and determine if it is significantly different from the click-through rate of the old design, p_A.

1.4 The Gaussian Distribution

While discrete distributions are fundamental, many real-world phenomena are measured on a continuous scale: height, weight, temperature, sensor readings. Among all continuous distributions, the Gaussian distribution holds a place of special prominence. Its ubiquity is not accidental; it is a consequence of one of the most profound results in probability theory, the Central Limit Theorem.

Definition 1.46 (Gaussian Distribution)**.** A continuous random variable X follows a **Gaussian (or Normal) distribution** with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, if its PDF is given by:

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

The parameter μ controls the center (location) of the distribution, and σ (the standard deviation) controls its spread. The case where $\mu = 0$ and $\sigma^2 = 1$ is called the **standard normal distribution**.

The mathematical form of the Gaussian PDF has several important properties, which we now state and prove formally.

Theorem 1.47 (Symmetry of the Gaussian Distribution)**.** *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. The PDF $f_X(x)$ is symmetric about the mean μ. That is, for all $h \in \mathbb{R}$*

$$f_X(\mu + h) = f_X(\mu - h).$$

Proof. Let $h \in \mathbb{R}$ be arbitrary. We compute

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

Similarly

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

Since both expressions are equal, we have $f_X(\mu + h) = f_X(\mu - h)$. □

Theorem 1.48 (Unimodality of the Gaussian Distribution)**.** *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. The PDF $f_X(x)$ has a unique global maximum at $x = \mu$.*

Proof. To find critical points, we compute the first derivative. Let $g(x) = -\frac{(x-\mu)^2}{2\sigma^2}$, so

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

Then

$$f_X'(x) = f_X(x) \cdot g'(x) = f_X(x) \cdot \left(-\frac{x-\mu}{\sigma^2}\right).$$

Setting $f_X'(x) = 0$ and noting that $f_X(x) > 0$ for all x, we require $x - \mu = 0$, giving $x = \mu$ as the unique critical point. To confirm this is a maximum, observe that $f_X'(x) > 0$ for $x < \mu$ and $f_X'(x) < 0$ for $x > \mu$. Thus f_X is increasing on $(-\infty, \mu)$ and decreasing on (μ, ∞), confirming that $x = \mu$ is the unique global maximum. □

Theorem 1.49 (Asymptotic Behavior of the Gaussian Distribution)**.** *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. The PDF satisfies $\lim_{x \to \pm\infty} f_X(x) = 0$.*

Proof. As $x \to \pm\infty$, we have $(x-\mu)^2 \to \infty$. Therefore

$$-\frac{(x-\mu)^2}{2\sigma^2} \to -\infty.$$

Since the exponential function satisfies $\lim_{t\to-\infty} e^t = 0$, we conclude:

$$\lim_{x\to\pm\infty} f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \lim_{x\to\pm\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot 0 = 0.$$

□

Theorem 1.50 (Inflection Points of the Gaussian Distribution)**.** *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. The PDF $f_X(x)$ has exactly two inflection points, located at $x = \mu - \sigma$ and $x = \mu + \sigma$.*

Proof. Inflection points occur where the second derivative changes sign, i.e., where $f_X''(x) = 0$. From the proof of unimodality,

$$f_X'(x) = -\frac{x-\mu}{\sigma^2} f_X(x).$$

Using the product rule

$$\begin{aligned} f_X''(x) &= -\frac{1}{\sigma^2} f_X(x) + \left(-\frac{x-\mu}{\sigma^2}\right) f_X'(x) \\ &= -\frac{1}{\sigma^2} f_X(x) + \left(-\frac{x-\mu}{\sigma^2}\right)\left(-\frac{x-\mu}{\sigma^2}\right) f_X(x) \\ &= f_X(x)\left[-\frac{1}{\sigma^2} + \frac{(x-\mu)^2}{\sigma^4}\right] = \frac{f_X(x)}{\sigma^4}\left[(x-\mu)^2 - \sigma^2\right] \end{aligned}$$

Setting $f_X''(x) = 0$ and noting $f_X(x) > 0$, we require $(x-\mu)^2 = \sigma^2$, giving $x - \mu = \pm\sigma$. Thus the inflection points are $x = \mu \pm \sigma$. □

Example 1.51 (Standard Normal Distribution)**.** Consider the standard normal distribution $Z \sim \mathcal{N}(0,1)$ with $\mu = 0$ and $\sigma = 1$. By the theorems above,

- the PDF is symmetric about 0: $f_Z(-2) = f_Z(2)$.
- the maximum value occurs at $z = 0$: $f_Z(0) = \frac{1}{\sqrt{2\pi}} \approx 0.3989$.
- the inflection points are at $z = -1$ and $z = 1$, where $f_Z(\pm 1) = \frac{1}{\sqrt{2\pi}} e^{-1/2} \approx 0.2420$.

These properties explain the familiar *bell curve* shape centered at zero with the steepest descent occurring one standard deviation from the mean.

Example 1.52 (Comparing Two Gaussian Distributions)**.** Let $X \sim \mathcal{N}(100, 25)$ (so $\mu = 100$, $\sigma = 5$) and $Y \sim \mathcal{N}(100, 100)$ (so $\mu = 100$, $\sigma = 10$). Both distributions are centered at $\mu = 100$ and symmetric about this value. However,

- for X, the maximum density is $f_X(100) = \frac{1}{\sqrt{2\pi \cdot 25}} = \frac{1}{5\sqrt{2\pi}} \approx 0.0798$, with inflection points at $x = 95$ and $x = 105$.
- for Y, the maximum density is $f_Y(100) = \frac{1}{\sqrt{2\pi \cdot 100}} = \frac{1}{10\sqrt{2\pi}} \approx 0.0399$, with inflection points at $y = 90$ and $y = 110$.

The distribution with smaller variance (X) has a taller, narrower peak, while the distribution with larger variance (Y) is shorter and more spread out. Both integrate to 1, so the areas under the curves are equal.

By defining random variables and exploring these fundamental distributions, we have established the vocabulary needed to describe data generation processes. The Bernoulli and Binomial distributions provide a framework for classification problems, while the Gaussian distribution provides a model for continuous phenomena. However, to fully appreciate why the Gaussian distribution is so central to statistics and data science, we must first develop the tools to characterize distributions through their *moments*—expectation and variance. Once these concepts are established, we will state the Central Limit Theorem, which provides the theoretical justification for the ubiquity of the Gaussian distribution.

2 Moments

Having defined random variables and their distributions, our next task is to summarize them. While a PDF or PMF provides a complete description of a random variable, it is often too detailed for practical comparison or analysis. We need concise numerical summaries that capture the most salient features of a distribution. These summaries are known as **moments**. In this section, we will rigorously define the first two moments—expectation and variance—which describe the center and spread of a distribution, respectively. We will then generalize these concepts to multiple dimensions through covariance and the covariance matrix, forging a critical link between the statistical properties of data and the geometric framework of dot products and vector spaces established in the first two weeks.

2.1 The First Moment: Mathematical Expectation

The expectation of a random variable is its theoretical mean, a probability-weighted average of all its possible values. It represents the *center of mass* of the distribution.

Definition 2.1 (Expectation)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a random variable. The **expectation** (or **expected value**) of X, denoted $\mathbb{E}[X]$, is defined as the Lebesgue integral of X with respect to the probability measure $\mathbb{P}$:

$$\mathbb{E}[X] = \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega).$$

We say X is **integrable** if $\mathbb{E}[|X|] < \infty$, and write $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$.

Remark 2.2. The expectation $\mathbb{E}[X]$ is well-defined only when X is integrable. Not all random variables satisfy this condition—the Cauchy distribution is a classical counterexample. Throughout this course, we assume integrability whenever we write $\mathbb{E}[X]$.

Remark 2.3 (Induced Probability Measure). Given a random variable $X : \Omega \to \mathbb{R}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we can define a new probability measure $\mathbb{P}_X$ on $\mathbb{R}$ called the **induced** (or **pushforward**) **measure**. For any Borel set $B \subseteq \mathbb{R}$, we define:

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}).$$

This is written concisely as $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$. The measure $\mathbb{P}_X$ lives on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-algebra on $\mathbb{R}$ (the σ-algebra generated by open intervals). The induced measure $\mathbb{P}_X$ captures the **distribution** of X: it tells us the probability that X takes values in any given subset of $\mathbb{R}$. The PMF and PDF are simply convenient representations of this induced measure for discrete and continuous random variables, respectively.

Proposition 2.4 (Computation of Expectation)**.** *Let X be an integrable random variable. Then the expectation can be computed via the induced distribution as follows:*

- *If X is **discrete** with PMF $p_X(x)$ and support $S \subseteq \mathbb{R}$, then:*

$$\mathbb{E}[X] = \sum_{x \in S} x \cdot p_X(x).$$

- *If X is **continuous** with PDF $f_X(x)$, then:*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) \, \mathrm{d}x.$$

Proof. By the change of variables formula for Lebesgue integrals, we have

$$\mathbb{E}[X] = \int_{\Omega} X(\omega) \, \mathrm{d}\mathbb{P}(\omega) = \int_{\mathbb{R}} x \, \mathrm{d}\mathbb{P}_X(x),$$

where $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$ is the induced probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For discrete X, this measure is a sum of point masses, yielding the summation formula. For continuous X with density f_X, we have $\mathrm{d}\mathbb{P}_X(x) = f_X(x) \, \mathrm{d}x$, yielding the integral formula. □

The following examples illustrate how to compute expectations for both discrete and continuous random variables.

Example 2.5 (Expectation of a Fair Die Roll)**.** Let X be the outcome of rolling a fair six-sided die. The PMF is $p_X(k) = \frac{1}{6}$ for $k \in \{1, 2, 3, 4, 5, 6\}$. Using the discrete expectation formula:

$$\begin{aligned} \mathbb{E}[X] &= \sum_{k=1}^{6} k \cdot p_X(k) \\ &= \sum_{k=1}^{6} k \cdot \frac{1}{6} \\ &= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3.5 \end{aligned}$$

Note that $\mathbb{E}[X] = 3.5$ is not a possible outcome of the die—the expectation need not be a value the random variable can actually take.

Example 2.6 (Expectation of a Bernoulli Random Variable)**.** Let $X \sim \text{Bernoulli}(p)$, so $X = 1$ with probability p and $X = 0$ with probability $1 - p$. Then:

$$\mathbb{E}[X] = 0 \cdot (1 - p) + 1 \cdot p = p$$

This elegant result shows that the expectation of an indicator random variable equals the probability of the event it indicates. For instance, if X represents whether a coin lands heads with $p = 0.7$, then $\mathbb{E}[X] = 0.7$.

Example 2.7 (Expectation of a Continuous Uniform Distribution)**.** Let $X \sim \text{Uniform}(a, b)$ with PDF $f_X(x) = \frac{1}{b-a}$ for $x \in [a, b]$. Using the continuous expectation formula:

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x) \, \mathrm{d}x = \int_{a}^{b} x \cdot \frac{1}{b-a} \, \mathrm{d}x \\ &= \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b = \frac{1}{b-a} \cdot \frac{b^2 - a^2}{2} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2} \end{aligned}$$

As expected intuitively, the mean of a uniform distribution is the midpoint of its support. For example, if $X \sim \text{Uniform}(0, 10)$, then $\mathbb{E}[X] = 5$.

Example 2.8 (Expectation of an Exponential Distribution)**.** Let $X \sim \text{Exponential}(\lambda)$ with PDF $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. We compute:

$$\mathbb{E}[X] = \int_0^\infty x \cdot \lambda e^{-\lambda x} \, \mathrm{d}x$$

Using integration by parts with $u = x$ and $\mathrm{d}v = \lambda e^{-\lambda x} \, \mathrm{d}x$, we get $\mathrm{d}u = \mathrm{d}x$ and $v = -e^{-\lambda x}$:

$$\begin{aligned} \mathbb{E}[X] &= \Big[-xe^{-\lambda x}\Big]_0^\infty + \int_0^\infty e^{-\lambda x} \, \mathrm{d}x \\ &= 0 + \left[-\frac{1}{\lambda} e^{-\lambda x}\right]_0^\infty = 0 - \left(-\frac{1}{\lambda}\right) = \frac{1}{\lambda} \end{aligned}$$

Thus, for an exponential distribution with rate λ, the expected value is $\frac{1}{\lambda}$. If a light bulb's lifetime follows Exponential(0.1) (measured in years), its expected lifetime is $\mathbb{E}[X] = 10$ years.

The expectation can be interpreted as the long-run average value of the random variable if we were to repeat the underlying experiment an infinite number of times. It is the single number that best summarizes the central tendency of the distribution.

A crucial property of expectation is its linearity. This property allows us to manipulate expectations of combinations of random variables with ease, a tool we will use constantly throughout the course.

Theorem 2.9 (Linearity of Expectation)**.** *For any two random variables X and Y (not necessarily independent) and any constants $a, b \in \mathbb{R}$,*

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

Proof for the continuous case. Let $f_{X,Y}(x, y)$ be the joint PDF of X and Y.

$$\begin{aligned} \mathbb{E}[aX + bY] &= \int_{-\infty}^\infty \int_{-\infty}^\infty (ax + by) f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y \\ &= \int_{-\infty}^\infty \int_{-\infty}^\infty ax f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y + \int_{-\infty}^\infty \int_{-\infty}^\infty by f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y \\ &= a \int_{-\infty}^\infty x \left(\int_{-\infty}^\infty f_{X,Y}(x, y) \, \mathrm{d}y\right) \mathrm{d}x + b \int_{-\infty}^\infty y \left(\int_{-\infty}^\infty f_{X,Y}(x, y) \, \mathrm{d}x\right) \mathrm{d}y \\ &= a \int_{-\infty}^\infty x f_X(x) \, \mathrm{d}x + b \int_{-\infty}^\infty y f_Y(y) \, \mathrm{d}y \\ &= a\mathbb{E}[X] + b\mathbb{E}[Y] \end{aligned}$$

The step from the third to the fourth line uses the definition of marginal densities: $f_X(x) = \int_{-\infty}^\infty f_{X,Y}(x, y) \, \mathrm{d}y$ and $f_Y(y) = \int_{-\infty}^\infty f_{X,Y}(x, y) \, \mathrm{d}x$. The proof for the discrete case is analogous, with sums replacing integrals. □

The power of linearity of expectation lies in its ability to simplify seemingly complex calculations. The following examples demonstrate how linearity allows us to bypass difficult computations entirely.

Example 2.10 (Expected Number of Heads in n Coin Flips)**.** Suppose we flip a biased coin n times, where each flip lands heads with probability p. Let X denote the total number of heads. To compute $\mathbb{E}[X]$ directly, we would need to use the binomial distribution:

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

This sum is tedious to evaluate. Instead, we use the *indicator variable trick*. Define $X_i = \mathbf{1}_{\{\text{flip } i \text{ is heads}\}}$ for $i = 1, \ldots, n$. Then $X = X_1 + X_2 + \cdots + X_n$, and by linearity:

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n] = p + p + \cdots + p = np$$

We obtained the answer $\mathbb{E}[X] = np$ without summing over all $n + 1$ terms of the binomial distribution. This technique—decomposing a count into a sum of indicators—is one of the most useful applications of linearity.

Example 2.11 (Expected Sum of Two Dice Without Joint Distribution)**.** Let $S = X_1 + X_2$ be the sum of two fair six-sided dice. A direct calculation would require enumerating all 36 equally likely outcomes (i, j) for $i, j \in \{1, \ldots, 6\}$, i.e.,

$$\mathbb{E}[S] = \frac{1}{36} \sum_{i=1}^{6} \sum_{j=1}^{6} (i + j)$$

However, linearity of expectation gives us the answer immediately:

$$\mathbb{E}[S] = \mathbb{E}[X_1] + \mathbb{E}[X_2] = 3.5 + 3.5 = 7.$$

Crucially, this approach works *regardless of whether X_1 and X_2 are independent.* Even if the dice were somehow dependent (e.g., weighted so that matching outcomes are more likely), linearity would still hold. This universality—linearity requires no assumptions about independence—is what makes it such a powerful tool.

Example 2.12 (Expected Value of a Sample Mean)**.** Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables from *any* distribution with mean μ. The sample mean is defined as

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

To find $\mathbb{E}[\bar{X}]$ directly would require knowing the distribution of $\bar{X}$, which can be complicated (e.g., the sum of uniforms yields a piecewise polynomial distribution). However, linearity gives us the answer immediately

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \cdot n\mu = \mu.$$

This elegant result—the expected value of the sample mean equals the population mean—holds for *any* distribution, and we derived it without computing a single integral.

Before presenting our next example, we introduce a continuous distribution that is fundamental in modeling waiting times and lifetimes.

Example 2.13 (Total Service Time as Sum of Exponentials)**.** A customer at a bank must complete two independent tasks: first, identity verification taking time $T_1 \sim \text{Exponential}(\lambda_1)$, then transaction processing taking time $T_2 \sim \text{Exponential}(\lambda_2)$. What is the expected total service time $\mathbb{E}[T_1 + T_2]$?

The distribution of $T_1 + T_2$ is a **hypoexponential distribution**, whose PDF involves convolution integrals. Computing $\mathbb{E}[T_1 + T_2]$ directly from this PDF would be tedious. Instead, linearity yields

$$\mathbb{E}[T_1 + T_2] = \mathbb{E}[T_1] + \mathbb{E}[T_2] = \frac{1}{\lambda_1} + \frac{1}{\lambda_2}.$$

For instance, if verification takes on average 2 minutes ($\lambda_1 = 0.5$) and processing takes on average 5 minutes ($\lambda_2 = 0.2$), then $\mathbb{E}[T_1 + T_2] = 2 + 5 = 7$ minutes. Linearity allows us to bypass the complex convolution entirely.

2.2 The Second Central Moment: Variance

While expectation tells us about the center of a distribution, it tells us nothing about its spread or scale. For example, two Gaussian distributions $\mathcal{N}(0,1)$ and $\mathcal{N}(0,100)$ have the same mean, but are dramatically different in their dispersion. Variance is the measure that quantifies this spread.

Definition 2.14 (Variance)**.** Let X be a random variable with mean $\mu = \mathbb{E}[X]$. The **variance** of X, denoted $\text{Var}(X)$ or σ_X^2, is the expected value of its squared deviation from the mean:

$$\text{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right]$$

The **standard deviation**, σ_X, is the positive square root of the variance, $\sigma_X = \sqrt{\text{Var}(X)}$. It has the advantage of being in the same units as the random variable itself.

The variance is called the second *central* moment because it is the expectation of a function of the *centered* random variable $(X - \mu)$. A more convenient formula for computation can be derived directly from the definition.

Theorem 2.15 (Computational Formula for Variance)**.** *For a random variable X with mean $\mu = \mathbb{E}[X]$,*

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

Proof. Starting from the definition and expanding the square:

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}[(X-\mu)^2] \\
&= \mathbb{E}[X^2 - 2X\mu + \mu^2] \\
&= \mathbb{E}[X^2] - \mathbb{E}[2X\mu] + \mathbb{E}[\mu^2] \quad \text{(by linearity of expectation)} \\
&= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 \quad \text{(since } \mu \text{ is a constant)} \\
&= \mathbb{E}[X^2] - 2\mu^2 + \mu^2 \\
&= \mathbb{E}[X^2] - \mu^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.
\end{aligned}$$

□

This result is immensely useful. It states that the variance is the expected value of the square of the variable minus the square of its expected value.

Example 2.16 (Variance of a Fair Die – Discrete Uniform Distribution)**.** Let X be the outcome of rolling a fair six-sided die, so $X \in \{1,2,3,4,5,6\}$ each with probability $1/6$. This is an example of a **discrete uniform distribution**, where each of the $n = 6$ equally-spaced outcomes has the same probability $1/n$. We first compute $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$:

$$\mathbb{E}[X] = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = \frac{7}{2}$$

$$\mathbb{E}[X^2] = \frac{1}{6}(1^2+2^2+3^2+4^2+5^2+6^2) = \frac{1+4+9+16+25+36}{6} = \frac{91}{6}$$

Applying the computational formula:

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{91}{6} - \frac{49}{4} = \frac{182-147}{12} = \frac{35}{12} \approx 2.92$$

The standard deviation is $\sigma_X = \sqrt{35/12} \approx 1.71$.

Example 2.17 (Variance of a Uniform Distribution)**.** Let $X \sim \text{Uniform}(a, b)$ with PDF $f(x) = \dfrac{1}{b-a}$ for $x \in [a, b]$. We compute

$$\mathbb{E}[X] = \int_a^b x \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a} \cdot \frac{x^2}{2}\Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2},$$

$$\mathbb{E}[X^2] = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a} \cdot \frac{x^3}{3}\Big|_a^b = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}.$$

Applying the computational formula yields

$$\text{Var}(X) = \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{a^2 + ab + b^2}{3} - \frac{a^2 + 2ab + b^2}{4} = \frac{(b-a)^2}{12}.$$

For example, if $X \sim \text{Uniform}(0, 1)$, then $\text{Var}(X) = 1/12 \approx 0.083$ and $\sigma_X = 1/\sqrt{12} \approx 0.289$.

Example 2.18 (Variance of an Exponential Distribution)**.** Let $X \sim \text{Exponential}(\lambda)$ with PDF $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. Using integration by parts, we compute

$$\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x}\,dx = \frac{1}{\lambda},$$

$$\mathbb{E}[X^2] = \int_0^\infty x^2\lambda e^{-\lambda x}\,dx = \frac{2}{\lambda^2}.$$

Applying the computational formula:

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Thus $\sigma_X = 1/\lambda$, which equals the mean. This property—that the standard deviation equals the mean—is characteristic of the exponential distribution.

Unlike expectation, variance is not linear. Its scaling property is important:

Theorem 2.19 (Properties of Variance)**.** *For a random variable X and constants $a, b \in \mathbb{R}$, it follows that*

$$\text{Var}(aX + b) = a^2 \,\text{Var}(X).$$

Proof.

$$\begin{aligned}
\text{Var}(aX + b) &= \mathbb{E}[((aX + b) - \mathbb{E}[aX + b])^2] \\
&= \mathbb{E}[(aX + b - (a\mathbb{E}[X] + b))^2] \\
&= \mathbb{E}[(aX - a\mathbb{E}[X])^2] \\
&= \mathbb{E}[a^2(X - \mathbb{E}[X])^2] \\
&= a^2\mathbb{E}[(X - \mathbb{E}[X])^2] = a^2 \,\text{Var}(X)
\end{aligned}$$

□

Notice that adding a constant b shifts the distribution but does not change its spread, hence b disappears from the final expression.

Example 2.20 (Scaling a Die Roll)**.** Recall from a previous example that for a fair six-sided die X, we have $\text{Var}(X) = 35/12$. Suppose we define a new random variable $Y = 2X + 3$ (e.g., doubling the outcome and adding 3). By the theorem, it yield

$$\text{Var}(Y) = \text{Var}(2X + 3) = 2^2 \cdot \text{Var}(X) = 4 \cdot \frac{35}{12} = \frac{35}{3} \approx 11.67.$$

Notice that the additive constant $+3$ does not affect the variance at all—it only shifts the distribution. The scaling factor 2, however, increases the variance by a factor of $2^2 = 4$.

Example 2.21 (Temperature Conversion: Celsius to Fahrenheit)**.** Let C be a random variable representing temperature in Celsius with $\mathbb{E}[C] = 20$ and $\text{Var}(C) = 25$ (so $\sigma_C = 5$ degrees Celsius). The conversion to Fahrenheit is given by $F = \frac{9}{5}C + 32$. Using the theorem yields

$$\text{Var}(F) = \text{Var}\left(\frac{9}{5}C + 32\right) = \left(\frac{9}{5}\right)^2 \text{Var}(C) = \frac{81}{25} \cdot 25 = 81.$$

Thus $\sigma_F = 9$ degrees Fahrenheit. The additive constant 32 (the offset between the two scales) does not contribute to the variance, while the scaling factor 9/5 causes the standard deviation to scale from 5 to $5 \times (9/5) = 9$. This illustrates why temperature variability *looks larger* when expressed in Fahrenheit than in Celsius.

2.3 Covariance: Re-establishing the Geometric Link

We now turn to the relationship between two random variables, X and Y. Covariance measures the degree to which they vary together.

Definition 2.22 (Covariance)**.** Let X and Y be two random variables with means $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$, respectively. The **covariance** between X and Y is:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

Similar to variance, a useful computational formula is $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

The sign of the covariance indicates the direction of the linear relationship:

- $\text{Cov}(X, Y) > 0$: X and Y tend to move in the same direction. When X is above its mean, Y is likely to be above its mean.
- $\text{Cov}(X, Y) < 0$: X and Y tend to move in opposite directions.
- $\text{Cov}(X, Y) = 0$: X and Y are **uncorrelated**. Note that independence implies uncorrelatedness, but the converse is not true in general.

Now we arrive at a critical insight that unifies the probabilistic view of data with the geometric view from the first lesson. In our geometric framework, we used the dot product to measure the similarity (or alignment) of two vectors. Let's see how this relates to covariance.

Consider a dataset of n observations for two variables, represented as vectors $\mathbf{x} = (x_1, \dots, x_n)^T$ and $\mathbf{y} = (y_1, \dots, y_n)^T$ in $\mathbb{R}^n$. We can think of these observations as realizations of random variables X and Y. The *sample mean* is the natural estimator for the expectation:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx \mathbb{E}[X] \quad \text{and} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \approx \mathbb{E}[Y]$$

The *sample covariance* is the estimator for the theoretical covariance

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \approx \text{Cov}(X, Y).$$

(Note: The use of $n - 1$ instead of n is for an unbiased estimator, a detail we will revisit.)

Let's analyze the summation term. First, we center the data vectors by subtracting their respective means as follows.

$$\tilde{\mathbf{x}} = \mathbf{x} - \bar{x}\mathbf{1} = (x_1 - \bar{x}, \dots, x_n - \bar{x})^T,$$

$$\tilde{\mathbf{y}} = \mathbf{y} - \bar{y}\mathbf{1} = (y_1 - \bar{y}, \dots, y_n - \bar{y})^T.$$

The sum in the sample covariance formula is precisely the dot product of these centered vectors

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \tilde{\mathbf{x}} \cdot \tilde{\mathbf{y}}.$$

Therefore, we have established a direct proportionality

$$s_{xy} = \frac{1}{n-1}(\tilde{\mathbf{x}} \cdot \tilde{\mathbf{y}}).$$

This is a profound connection. The **sample covariance**, our primary statistical measure of the linear relationship between two variables, is simply a scaled version of the **dot product** of their centered data vectors. The geometric notion of alignment that we used to define similarity is mathematically synonymous with the statistical notion of covariance. A large positive dot product between centered vectors means they point in similar directions in $\mathbb{R}^n$, which in turn means the variables have a large positive covariance.

2.4 The Covariance Matrix

The final step is to generalize from two variables to p variables. We collect p random variables into a random vector $\mathbf{X} = [X_1, X_2, \ldots, X_p]^T$. The variance-covariance structure of this vector is captured by a single mathematical object: the covariance matrix.

Definition 2.23 (Covariance Matrix)**.** For a random vector $\mathbf{X} \in \mathbb{R}^p$ with mean vector $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$, the **covariance matrix** Σ is a $p \times p$ matrix defined as

$$\Sigma = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T].$$

The entry in the i-th row and j-th column of Σ is $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$.

Unpacking this definition, we see:

$$\Sigma = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \cdots & \mathrm{Var}(X_p) \end{pmatrix}.$$

Key properties of the covariance matrix Σ include the following.

1. **Symmetry:** Since $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$, we have $\Sigma = \Sigma^T$.
2. **Positive Semi-Definite:** For any constant vector $\mathbf{a} \in \mathbb{R}^p$, $\mathrm{Var}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\Sigma\mathbf{a} \geq 0$. This property is fundamental and connects covariance matrices to the theory of quadratic forms from linear algebra.
3. **Diagonal Entries:** The diagonal entries are the variances of the individual random variables, $\Sigma_{ii} = \mathrm{Var}(X_i)$.

The sample covariance matrix, denoted $\mathbf{S}$, is the empirical counterpart, computed from a data matrix D of size $n \times p$.

The covariance matrix is the key mathematical object for describing the second-order structure of a multivariate distribution. It generalizes the concept of variance to multiple dimensions. For example, the **Multivariate Normal Distribution** is defined entirely by its mean vector $\boldsymbol{\mu}$ and its covariance matrix Σ, denoted $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. The covariance matrix Σ describes the shape, orientation, and scale of the elliptical contours of the distribution's density function. Methods like Principal Component Analysis (PCA), which we will study in later lessons, are based entirely on finding the eigenvectors and eigenvalues of the sample covariance matrix to identify the directions of maximal variance in the data. This is a direct application of the spectral theorem from linear algebra to solve a statistical problem.

2.5 The Law of Large Numbers

We now present one of the most fundamental results in probability theory: the Law of Large Numbers (LLN). This theorem formalizes the intuitive notion that averaging more observations yields increasingly accurate estimates of the true mean. Before stating the LLN, we establish a useful probabilistic inequality.

Lemma 2.24 (Chebyshev's Inequality). *Let X be a random variable with finite mean $\mu = \mathbb{E}[X]$ and finite variance $\sigma^2 = \mathrm{Var}(X)$. Then for any $\epsilon > 0$, it follows that*

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

Proof. Define the indicator random variable $\mathbf{1}_{|X-\mu|\geq\epsilon}$, which equals 1 when $|X - \mu| \geq \epsilon$ and 0 otherwise. Observe that when $|X - \mu| \geq \epsilon$, we have $(X - \mu)^2 \geq \epsilon^2$, so

$$\epsilon^2 \cdot \mathbf{1}_{|X-\mu|\geq\epsilon} \leq (X - \mu)^2.$$

Taking expectations of both sides yields

$$\epsilon^2 \cdot \mathbb{E}[\mathbf{1}_{|X-\mu|\geq\epsilon}] \leq \mathbb{E}[(X - \mu)^2] = \sigma^2.$$

Since $\mathbb{E}[\mathbf{1}_{|X-\mu|\geq\epsilon}] = \mathbb{P}(|X - \mu| \geq \epsilon)$, dividing by ϵ^2 yields the result. □

Chebyshev's inequality provides a universal bound on how much probability mass can lie far from the mean, using only the variance. It holds for *any* distribution with finite variance, making it a powerful tool despite often being a loose bound.

Theorem 2.25 (Weak Law of Large Numbers). *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite mean $\mu = \mathbb{E}[X_i]$ and finite variance $\sigma^2 = \mathrm{Var}(X_i)$. Define the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for any $\epsilon > 0$:*

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) = 0.$$

*In other words, $\bar{X}_n$ **converges in probability** to μ, denoted $\bar{X}_n \xrightarrow{p} \mu$.*

Proof. We apply Chebyshev's inequality to the sample mean $\bar{X}_n$. First, we compute the mean and variance of $\bar{X}_n$.

By linearity of expectation, we have

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \cdot n\mu = \mu.$$

Since the X_i are independent,

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

Applying Chebyshev's inequality to $\bar{X}_n$ yields

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

As $n \to \infty$, the right-hand side converges to 0, so

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) = 0.$$ □

Example 2.26 (Estimating a Population Mean)**.** Suppose we want to estimate the average height μ of adults in a country. We take a random sample of n individuals and compute the sample mean $\bar{X}_n$. The weak law of large numbers guarantees that as our sample size increases, the probability that our estimate $\bar{X}_n$ differs from the true mean μ by more than any specified tolerance ϵ becomes arbitrarily small. For instance, with $\epsilon = 0.1$ cm, $\sigma = 10$ cm, and $n = 10000$:

$$\mathbb{P}(|\bar{X}_n - \mu| \geq 0.1) \leq \frac{100}{10000 \times 0.01} = 0.1.$$

This bound (at most 10% chance of being off by more than 0.1 cm) is actually quite conservative—the true probability is typically much smaller.

Remark 2.27 (Weak vs. Strong Law of Large Numbers)*.* The **weak law** (proven above) establishes convergence in probability: $\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \to 0$. The **strong law of large numbers** establishes almost sure convergence: $\mathbb{P}(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$. The strong law is a more powerful result (almost sure convergence implies convergence in probability) but requires more sophisticated proof techniques. For most applications in data science, convergence in probability is sufficient.

LLN is the theoretical foundation for statistical inference. It justifies why sample statistics (like the sample mean) are useful estimators of population parameters: with enough data, our estimates will be arbitrarily close to the true values with high probability.

2.6 The Central Limit Theorem

With the concepts of expectation and variance now established, we can formally state the central limit theorem (CLT), which explains why the Gaussian distribution is so central to statistics and data science.

Theorem 2.28 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite mean $\mu = \mathbb{E}[X_i]$ and finite variance $\sigma^2 = \mathrm{Var}(X_i) > 0$. Define the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then, as $n \to \infty$, the standardized sample mean converges in distribution to the standard normal:*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

Equivalently, for large n, we have the approximation $\bar{X}_n \overset{approx}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Proof. The proof uses **characteristic functions**. The characteristic function of a random variable X is defined as $\varphi_X(t) = \mathbb{E}[e^{itX}]$, where $i = \sqrt{-1}$. We rely on **Lévy's Continuity Theorem**: if a sequence of characteristic functions $\varphi_n(t) \to \varphi(t)$ pointwise for all t, and φ is continuous at $t = 0$, then the corresponding distributions converge in distribution.

Step 1: Standardize the variables. Without loss of generality, assume $\mu = 0$ and $\sigma^2 = 1$ (we can always work with $Y_i = (X_i - \mu)/\sigma$). Define the standardized sum:

$$Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i = \frac{\bar{X}_n - 0}{1/\sqrt{n}}.$$

Step 2: Compute the characteristic function of Z_n. Since the X_i are i.i.d., the characteristic function of Z_n is:

$$\varphi_{Z_n}(t) = \mathbb{E}\left[e^{itZ_n}\right] = \mathbb{E}\left[\exp\left(i\frac{t}{\sqrt{n}}\sum_{i=1}^n X_i\right)\right] = \prod_{i=1}^n \mathbb{E}\left[e^{i(t/\sqrt{n})X_i}\right] = \left[\varphi_X\left(\frac{t}{\sqrt{n}}\right)\right]^n.$$

Step 3: Taylor expand the characteristic function. Since $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$, we expand $\varphi_X(s)$ around $s = 0$:

$$\varphi_X(s) = \mathbb{E}[e^{isX}] = \mathbb{E}\left[1 + isX + \frac{(isX)^2}{2!} + O(s^3)\right] = 1+is\mathbb{E}[X]-\frac{s^2}{2}\mathbb{E}[X^2]+O(s^3) = 1-\frac{s^2}{2}+O(s^3).$$

Step 4: Take the limit. Substituting $s = t/\sqrt{n}$:

$$\varphi_{Z_n}(t) = \left[1 - \frac{t^2}{2n} + O\left(\frac{1}{n^{3/2}}\right)\right]^n .$$

Using the fact that $\lim_{n\to\infty}(1 + a/n)^n = e^a$, we obtain:

$$\lim_{n\to\infty} \varphi_{Z_n}(t) = e^{-t^2/2}.$$

Step 5: Identify the limit. The function $e^{-t^2/2}$ is the characteristic function of the standard normal distribution $\mathcal{N}(0, 1)$. By Lévy's Continuity Theorem, $Z_n \xrightarrow{d} \mathcal{N}(0, 1)$. □

The CLT is a profound result. It states that the distribution of sample means approaches a Gaussian distribution regardless of the underlying distribution of the individual observations, provided they have finite mean and variance. This explains why the Gaussian distribution appears so frequently in nature, as many phenomena can be viewed as the cumulative effect of numerous small, independent random processes. For data scientists, the CLT means that sample means from non-normal populations will have an approximately normal sampling distribution, a fact that underpins much of classical hypothesis testing.

Example 2.29 (Averaging Dice Rolls)**.** Consider rolling a fair six-sided die. Each roll X_i has mean $\mu = \mathbb{E}[X_i] = \frac{1+2+3+4+5+6}{6} = 3.5$ and variance $\sigma^2 = \mathrm{Var}(X_i) = \frac{35}{12} \approx 2.917$. The distribution of a single roll is discrete and uniform—far from Gaussian. However, by the CLT, if we roll the die $n = 100$ times and compute the sample mean $\bar{X}_{100}$, this average is approximately normally distributed:

$$\bar{X}_{100} \overset{\text{approx}}{\sim} \mathcal{N}\left(3.5, \frac{2.917}{100}\right) = \mathcal{N}(3.5, 0.02917).$$

The standard deviation of the sample mean is $\sigma/\sqrt{n} \approx 0.171$. Thus, approximately 95% of the time, the average of 100 dice rolls will fall within $3.5 \pm 1.96 \times 0.171 \approx [3.16, 3.84]$.

3 Parameter Estimation

In the previous sections, we established a probabilistic vocabulary for describing data. We can now model observations as draws from distributions like the Bernoulli or Gaussian, which are governed by a set of parameters (e.g., p or μ, σ^2). This leads to a fundamental question in statistics and machine learning: Given a set of observed data and a chosen parametric model, how do we determine the *best* values for these parameters?

For example, if we model coin flips as Bernoulli trials, how do we use a sequence of observed flips to estimate the coin's bias, p? If we assume the errors in a linear model are Gaussian, how do we use the data to estimate the regression coefficients β? The geometric approach of OLS provided one answer through projection. Here, we introduce a more general and powerful principle for parameter estimation that is rooted entirely in probability: the principle of **maximum likelihood estimation (MLE)**.

3.1 The Likelihood Function

The central object in MLE is the likelihood function. It is essential to distinguish this from a probability distribution function, as they represent fundamentally different concepts despite their mathematical similarity.

Suppose we have a dataset $D = \{x_1, x_2, \ldots, x_n\}$ that we assume are i.i.d. realizations of a random variable X. Let's say we have chosen a parametric model for the distribution of X, described by a PDF or PMF $f(x|\theta)$, where θ is the vector of parameters for the model.

Definition 3.1 (Likelihood Function)**.** Given observed data $D = \{x_1, \ldots, x_n\}$, the **likelihood function**, denoted $L(\theta|D)$, is the joint probability of observing this specific data, viewed as a function of the parameters θ. Due to the i.i.d. assumption, this joint probability is the product of the individual probabilities:

$$L(\theta|D) = \mathbb{P}(D|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Let's dissect this definition carefully.

- The expression $f(x|\theta)$ is the probability (or density) of a single data point x given a fixed value of θ.
- The likelihood function $L(\theta|D)$ takes this same mathematical expression but flips the perspective. The data D is now considered fixed (it's what we observed), and the parameters θ are the variables of the function.
- The value of $L(\theta|D)$ is not the probability that θ is the true parameter. Instead, it answers the question: "If the true parameter vector were θ, what would be the probability of observing the data D that we actually collected?" It measures how *reasonable* a particular parameter vector θ is, in light of our data.

Example 3.2 (Bernoulli Likelihood)**.** Suppose we flip a coin 3 times and observe the sequence $D = \{H, T, H\}$, which we code as $\{1, 0, 1\}$. We model each flip as an independent draw from a Bernoulli(p) distribution. The PMF is $f(x|p) = p^x(1-p)^{1-x}$. The likelihood function is:

$$\begin{aligned} L(p|D = \{1,0,1\}) &= f(1|p) \cdot f(0|p) \cdot f(1|p) \\ &= (p^1(1-p)^0) \cdot (p^0(1-p)^1) \cdot (p^1(1-p)^0) \\ &= p \cdot (1-p) \cdot p = p^2(1-p) \end{aligned}$$

This is a function of p. For $p = 0.5$, $L = 0.125$. For $p = 0.7$, $L = 0.7^2(0.3) = 0.147$. Our observed data is more likely under the hypothesis that $p = 0.7$ than under the hypothesis that $p = 0.5$.

Example 3.3 (Poisson Likelihood)**.** A store records the number of customers arriving each hour. Over 5 hours, they observe $D = \{3, 7, 4, 5, 6\}$ customers. We model each count as an independent draw from a Poisson(λ) distribution, where λ is the average arrival rate. The PMF is $f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$. The likelihood function is:

$$L(\lambda|D) = \prod_{i=1}^{5} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{3+7+4+5+6} e^{-5\lambda}}{3! \cdot 7! \cdot 4! \cdot 5! \cdot 6!} = \frac{\lambda^{25} e^{-5\lambda}}{3! \cdot 7! \cdot 4! \cdot 5! \cdot 6!}$$

The denominator is a constant (it depends only on the data, not on λ), so the likelihood is proportional to $\lambda^{25}e^{-5\lambda}$. Evaluating at $\lambda = 4$: $L \propto 4^{25}e^{-20} \approx 2.32 \times 10^6$. At $\lambda = 5$: $L \propto 5^{25}e^{-25} \approx 4.14 \times 10^6$. The sample mean is $\bar{x} = 25/5 = 5$, and indeed $\lambda = 5$ yields a higher likelihood than $\lambda = 4$.

Example 3.4 (Gaussian Likelihood)**.** Suppose we measure the weights (in kg) of 4 objects from a production line: $D = \{10.2, 9.8, 10.1, 10.3\}$. We model each measurement as an independent draw from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with known variance $\sigma^2 = 0.04$. The PDF is

$$f(x|\mu) = \frac{1}{\sqrt{2\pi \cdot 0.04}} \exp\left(-\frac{(x-\mu)^2}{2 \cdot 0.04}\right) = \frac{1}{0.4\sqrt{\pi}} \exp\left(-\frac{(x-\mu)^2}{0.08}\right).$$

The likelihood function is

$$L(\mu|D) = \prod_{i=1}^{4} f(x_i|\mu) \propto \exp\left(-\frac{1}{0.08}\sum_{i=1}^{4}(x_i-\mu)^2\right).$$

The likelihood is maximized when $\sum_{i=1}^{4}(x_i - \mu)^2$ is minimized, which occurs at $\mu = \bar{x} = (10.2 + 9.8 + 10.1 + 10.3)/4 = 10.1$. This foreshadows a key result: for Gaussian data, the maximum likelihood estimate of the mean is the sample mean.

3.2 The Principle of Maximum Likelihood Estimation

The example above naturally leads to the core idea of MLE. If the likelihood function measures how well a set of parameters explains the observed data, then a sensible strategy for choosing the parameters is to find the set that makes our data *most likely.*

Definition 3.5 (Maximum Likelihood Estimation (MLE))**.** The **maximum likelihood estimate** (MLE) of a parameter vector θ is the value $\hat{\theta}_{MLE}$ that maximizes the likelihood function $L(\theta|D)$.

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta|D) = \arg\max_{\theta} \prod_{i=1}^{n} f(x_i|\theta)$$

Directly maximizing the likelihood product can be mathematically cumbersome. The product of many small probabilities can also lead to numerical underflow issues in computation. We can simplify the problem significantly by maximizing the **log-likelihood** instead. Since the natural logarithm is a strictly increasing (monotonically increasing) function, maximizing $\log(L(\theta|D))$ is equivalent to maximizing $L(\theta|D)$—the maximum will occur at the same value of θ.

The log-likelihood function, denoted $\ell(\theta|D)$, is

$$\ell(\theta|D) = \log L(\theta|D) = \log\left(\prod_{i=1}^{n} f(x_i|\theta)\right) = \sum_{i=1}^{n} \log f(x_i|\theta).$$

This transformation converts the product into a sum, which is far easier to handle, particularly when using calculus to find the maximum. The optimization problem becomes

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i|\theta).$$

For differentiable log-likelihood functions, a standard approach to finding the maximum is to take the derivative (or gradient, in the multivariate case) with respect to θ, set it to zero, and solve for θ. This is known as the *score function.*

Example 3.6 (MLE for the Bernoulli Parameter)**.** Let's formalize the coin flip example. Suppose we have a dataset $D = \{x_1, \ldots, x_n\}$ of n i.i.d. Bernoulli trials, where $x_i \in \{0, 1\}$. Let n_1 be the number of successes (1s) and n_0 be the number of failures (0s), so $n_1 + n_0 = n$.

1. Write down the likelihood function: The PMF is $f(x_i|p) = p^{x_i}(1-p)^{1-x_i}$. The likelihood is the product:

$$L(p|D) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{\sum(1-x_i)} = p^{n_1}(1-p)^{n_0}.$$

2. Compute the log-likelihood:

$$\ell(p|D) = \log(p^{n_1}(1-p)^{n_0}) = n_1 \log(p) + n_0 \log(1-p).$$

3. Differentiate with respect to the parameter (p):

$$\frac{d\ell}{dp} = \frac{d}{dp}[n_1 \log(p) + n_0 \log(1-p)] = \frac{n_1}{p} - \frac{n_0}{1-p}.$$

4. Set the derivative to zero and solve for p:

$$\begin{aligned} \frac{n_1}{\hat{p}} - \frac{n_0}{1-\hat{p}} &= 0, \\ \frac{n_1}{\hat{p}} &= \frac{n_0}{1-\hat{p}}, \\ n_1(1-\hat{p}) &= n_0\hat{p}, \\ n_1 - n_1\hat{p} &= n_0\hat{p}, \\ n_1 &= (n_0 + n_1)\hat{p}, \\ n_1 &= n\hat{p}, \\ \hat{p} &= \frac{n_1}{n}. \end{aligned}$$

The maximum likelihood estimate for the Bernoulli parameter p is simply the sample proportion of successes. This result is both intuitive and reassuring; the formal principle of MLE has recovered the obvious empirical estimate.

3.3 Connecting OLS and MLE

We now arrive at a crucial synthesis of the course material. The ordinary least squares solution from the previous lesson was derived from a purely geometric standpoint of minimizing the length of a residual vector. We can now show that this is mathematically identical to the MLE solution under a specific and very common probabilistic assumption: that the errors are i.i.d. Gaussian.

Consider the standard linear model:

$$y_i = \mathbf{x}_i^T \beta + \epsilon_i.$$

where $\mathbf{x}_i$ is the vector of predictors for observation i, β is the vector of coefficients, and ϵ_i is the error term.

Assume that the errors are i.i.d. draws from a zero-mean normal distribution with variance σ^2. That is, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

This assumption about the errors implies a distribution for our target variable y_i. Since $y_i = \mathbf{x}_i^T \beta + \epsilon_i$ is the sum of a constant (given $\mathbf{x}_i$ and β) and a normal random variable (ϵ_i), y_i is also normally distributed. Specifically, y_i follows a normal distribution with:

- **Mean** $\mathbf{x}_i^T \beta$ (the predicted value from the regression line)
- **Variance** σ^2 (inherited from the error term)

We write this concisely using conditional distribution notation as

$$y_i \mid \mathbf{x}_i, \beta, \sigma^2 \sim \mathcal{N}(\mathbf{x}_i^T \beta, \sigma^2).$$

(Given the predictors $\mathbf{x}_i$, coefficients β, and variance σ^2, the response y_i is normally distributed with mean $\mathbf{x}_i^T \beta$ and variance σ^2.) Now we can use MLE to find the best estimate for β.

The PDF for a single observation y_i is

$$f(y_i|\mathbf{x}_i, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T \beta)^2}{2\sigma^2}\right).$$

The log-likelihood for the entire dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is the sum of the individual log-PDFs:

$$\begin{aligned}
\ell(\beta, \sigma^2|D) &= \sum_{i=1}^n \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T \beta)^2}{2\sigma^2}\right)\right] \\
&= \sum_{i=1}^n \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_i - \mathbf{x}_i^T \beta)^2}{2\sigma^2}\right] \\
&= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2.
\end{aligned}$$

To find $\hat{\beta}_{MLE}$, we need to maximize $\ell(\beta, \sigma^2|D)$. Notice that the first term, $-\frac{n}{2}\log(2\pi\sigma^2)$, does not depend on β. Therefore, maximizing the entire expression with respect to β is equivalent to maximizing just the second term.

$$\hat{\beta}_{\text{MLE}} = \arg\max_{\beta} \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2\right].$$

Since $2\sigma^2$ is a positive constant, maximizing this is equivalent to minimizing its negation:

$$\hat{\beta}_{\text{MLE}} = \arg\min_{\beta} \left[\sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2\right].$$

The expression inside the minimization is precisely the **residual sum of squares (RSS)**, $\|y - X\beta\|_2^2$. This is exactly the quantity that ordinary least squares seeks to minimize.

This is a profound result. It shows that the deterministic, geometric method of finding the orthogonal projection of $\mathbf{y}$ onto the column space of X is identical to the probabilistic method of finding the maximum likelihood estimate for the model parameters, provided we assume normally distributed errors. This equivalence gives the OLS solution a powerful statistical justification and allows us to use the machinery of probability theory to analyze its properties, such as constructing confidence intervals and performing hypothesis tests on the coefficients, which will be a focus of upcoming lectures.

4 Conditional Probability and Bayes' Theorem

The principle of MLE provides a powerful frequentist framework for parameter estimation. It seeks a single point estimate, $\hat{\theta}_{\text{MLE}}$, that best explains the data we have observed. However, there is an alternative and profoundly influential paradigm in statistics: the Bayesian approach. Instead of viewing parameters as fixed, unknown constants, Bayesian inference treats them as random variables about which we can have beliefs that are updated as we collect data. The mathematical engine that drives this process of belief updating is Bayes' Theorem, which is built upon the fundamental concept of conditional probability.

4.1 Conditional Probability

In many real-world scenarios, our knowledge about the occurrence of one event can alter our assessment of the probability of another. This is the intuition behind conditional probability.

Definition 4.1 (Conditional Probability)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $A, B \in \mathcal{F}$ be two events with $\mathbb{P}(B) > 0$. The **conditional probability** of event A occurring, given that event B has occurred, is defined as

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

where $\mathbb{P}(A \cap B)$ is the probability of the intersection of A and B, i.e., the probability that both events occur.

Example 4.2 (Rolling a Die)**.** Suppose we roll a fair six-sided die. Let A be the event "the outcome is even" (i.e., $A = \{2, 4, 6\}$) and let B be the event "the outcome is greater than 3" (i.e., $B = \{4, 5, 6\}$). We have

- $\mathbb{P}(A) = 3/6 = 1/2$,
- $\mathbb{P}(B) = 3/6 = 1/2$,
- $A \cap B = \{4, 6\}$, so $\mathbb{P}(A \cap B) = 2/6 = 1/3$.

The conditional probability of rolling an even number, given that we rolled greater than 3, is

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

This makes intuitive sense: among the outcomes greater than 3 (namely 4, 5, 6), two of them (4 and 6) are even.

Example 4.3 (Drawing Cards)**.** Consider drawing a single card from a standard 52-card deck. Let A be the event "the card is a King" and let B be the event "the card is a face card" (Jack, Queen, or King). We have

- $\mathbb{P}(A) = 4/52 = 1/13$ (there are 4 Kings),
- $\mathbb{P}(B) = 12/52 = 3/13$ (there are 12 face cards: 4 Jacks + 4 Queens + 4 Kings),
- $A \cap B = A$ (every King is a face card), so $\mathbb{P}(A \cap B) = 4/52 = 1/13$.

The conditional probability of drawing a King, given that we drew a face card, is:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/13}{3/13} = \frac{1}{3}.$$

Once we know the card is a face card, we have effectively reduced our sample space from 52 cards to just 12 face cards, of which 4 are Kings.

Example 4.4 (Medical Testing)**.** A medical test for a disease has the following characteristics: if a patient has the disease, the test correctly returns positive 95% of the time (sensitivity). If a patient does not have the disease, the test correctly returns negative 90% of the time (specificity). Suppose 1% of the population has the disease.

Let D be the event "patient has the disease" and T^+ be the event "test is positive." We are given

- $\mathbb{P}(D) = 0.01$ (prevalence),

- $\mathbb{P}(T^+|D) = 0.95$ (sensitivity: positive test given disease),
- $\mathbb{P}(T^-|\neg D) = 0.90$, which means $\mathbb{P}(T^+|\neg D) = 0.10$ (false positive rate).

We can compute $\mathbb{P}(D \cap T^+)$, the probability of having the disease *and* testing positive:

$$\mathbb{P}(D \cap T^+) = \mathbb{P}(T^+|D) \cdot \mathbb{P}(D) = 0.95 \times 0.01 = 0.0095.$$

This example sets up the classic application of Bayes' theorem, which we will explore in the next section to answer: "If a patient tests positive, what is the probability they actually have the disease?"

This definition is essentially a re-normalization. By conditioning on B, we are restricting our sample space from Ω to the subset B. The formula then re-scales the probability of the part of A that lies within B (i.e., $A \cap B$) by the probability of this new, smaller sample space.

Rearranging the definition gives the **product rule** (or chain rule) of probability, which is extremely useful for calculating the probability of an intersection:

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

Similarly, if $\mathbb{P}(A) > 0$, we have $\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$.

Example 4.5 (Drawing Two Cards Without Replacement)**.** What is the probability of drawing two Aces in a row from a standard 52-card deck (without replacement)? Let A_1 be the event "first card is an Ace" and A_2 be the event "second card is an Ace." We want $\mathbb{P}(A_1 \cap A_2)$.

Using the product rule:

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_1) = \frac{3}{51} \times \frac{4}{52} = \frac{12}{2652} = \frac{1}{221} \approx 0.0045.$$

Here, $\mathbb{P}(A_1) = 4/52$ (4 Aces among 52 cards), and $\mathbb{P}(A_2|A_1) = 3/51$ (after drawing one Ace, 3 Aces remain among 51 cards).

Example 4.6 (Weather and Commuting)**.** Suppose that on any given day, the probability of rain is $\mathbb{P}(R) = 0.3$. If it rains, the probability of heavy traffic is $\mathbb{P}(T|R) = 0.8$. What is the probability that it both rains and there is heavy traffic?

Using the product rule:

$$\mathbb{P}(R \cap T) = \mathbb{P}(T|R) \cdot \mathbb{P}(R) = 0.8 \times 0.3 = 0.24.$$

There is a 24% chance of experiencing both rain and heavy traffic on any given day.

Example 4.7 (Manufacturing Quality Control)**.** A factory produces electronic components in two stages. In Stage 1, 95% of components pass inspection. Among those that pass Stage 1, 90% pass Stage 2 inspection. What is the probability that a randomly selected component passes both stages?

Let S_1 be "passes Stage 1" and S_2 be "passes Stage 2." We are given $\mathbb{P}(S_1) = 0.95$ and $\mathbb{P}(S_2|S_1) = 0.90$. Using the product rule:

$$\mathbb{P}(S_1 \cap S_2) = \mathbb{P}(S_2|S_1) \cdot \mathbb{P}(S_1) = 0.90 \times 0.95 = 0.855.$$

Thus, 85.5% of all components pass both inspection stages. This can be extended: if there were a Stage 3 with $\mathbb{P}(S_3|S_1 \cap S_2) = 0.92$, the probability of passing all three stages would be $0.855 \times 0.92 = 0.787$.

4.2 The Engine of Inference

Bayes' theorem is a direct and elegant consequence of the definition of conditional probability. It provides a mechanism for *inverting* conditional probabilities—that is, if we know $\mathbb{P}(B|A)$, Bayes' theorem allows us to calculate $\mathbb{P}(A|B)$. In the context of data science, this inversion is the key to statistical inference.

Theorem 4.8 (Bayes' Theorem)**.** *Let A and B be events with $\mathbb{P}(A), \mathbb{P}(B) > 0$. Then*

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

Proof. From the product rule, we have two expressions for $\mathbb{P}(A \cap B)$:

1. $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$.
2. $\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$.

Equating these two gives

$$\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A).$$

Dividing by $\mathbb{P}(B)$ (which is non-zero by assumption) yields the theorem

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

□

Example 4.9 (Medical Testing Revisited)**.** Recall the medical testing scenario from earlier: a test has 95% sensitivity ($\mathbb{P}(T^+|D) = 0.95$) and 90% specificity ($\mathbb{P}(T^-|\neg D) = 0.90$), with disease prevalence of 1% ($\mathbb{P}(D) = 0.01$). A patient tests positive. What is the probability they actually have the disease?

We want $\mathbb{P}(D|T^+)$. First, we compute $\mathbb{P}(T^+)$ using the law of total probability:

$$\begin{aligned}\mathbb{P}(T^+) &= \mathbb{P}(T^+|D)\mathbb{P}(D) + \mathbb{P}(T^+|\neg D)\mathbb{P}(\neg D)\\ &= 0.95 \times 0.01 + 0.10 \times 0.99\\ &= 0.0095 + 0.099 = 0.1085\end{aligned}$$

Now applying Bayes' theorem:

$$\mathbb{P}(D|T^+) = \frac{\mathbb{P}(T^+|D)\mathbb{P}(D)}{\mathbb{P}(T^+)} = \frac{0.95 \times 0.01}{0.1085} = \frac{0.0095}{0.1085} \approx 0.0876$$

Surprisingly, even with a positive test result, there is only about an 8.8% chance the patient has the disease! This counterintuitive result arises because the disease is rare (1% prevalence), so the false positives from the healthy population (10% of 99%) overwhelm the true positives.

Example 4.10 (Spam Email Filtering)**.** An email filter classifies messages as spam or not spam. Historical data shows:

- 30% of all emails are spam: $\mathbb{P}(S) = 0.30$
- The word "free" appears in 80% of spam emails: $\mathbb{P}(\text{“free”}|S) = 0.80$
- The word "free" appears in 10% of legitimate emails: $\mathbb{P}(\text{“free”}|\neg S) = 0.10$

If an email contains the word "free," what is the probability it is spam?

First, compute $\mathbb{P}(\text{“free”})$:

$$\mathbb{P}(\text{“free”}) = 0.80 \times 0.30 + 0.10 \times 0.70 = 0.24 + 0.07 = 0.31$$

Applying Bayes' theorem:

$$\mathbb{P}(S|\text{“free”}) = \frac{\mathbb{P}(\text{“free”}|S)\mathbb{P}(S)}{\mathbb{P}(\text{“free”}} = \frac{0.80 \times 0.30}{0.31} = \frac{0.24}{0.31} \approx 0.774$$

An email containing "free" has about a 77.4% probability of being spam. This is the foundation of Naive Bayes classifiers, which extend this idea to multiple features.

Example 4.11 (Quality Control: Identifying Defective Sources)**.** A factory receives components from two suppliers. Supplier A provides 60% of components, and Supplier B provides 40%. The defect rates are: 2% for Supplier A and 5% for Supplier B. A randomly selected component is found to be defective. What is the probability it came from Supplier B?

Let B denote "from Supplier B" and Def denote "defective." We want $\mathbb{P}(B|\text{Def})$.

First, compute $\mathbb{P}(\text{Def})$:

$$\mathbb{P}(\text{Def}) = \mathbb{P}(\text{Def}|A)\mathbb{P}(A) + \mathbb{P}(\text{Def}|B)\mathbb{P}(B) = 0.02 \times 0.60 + 0.05 \times 0.40 = 0.012 + 0.020 = 0.032$$

Applying Bayes' theorem:

$$\mathbb{P}(B|\text{Def}) = \frac{\mathbb{P}(\text{Def}|B)\mathbb{P}(B)}{\mathbb{P}(\text{Def})} = \frac{0.05 \times 0.40}{0.032} = \frac{0.020}{0.032} = 0.625$$

Given that a component is defective, there is a 62.5% probability it came from Supplier B, even though Supplier B only provides 40% of total components. The higher defect rate of Supplier B shifts the posterior probability.

While mathematically simple, the conceptual power of Bayes's theorem is immense when we re-cast it in the language of models and data. Let's replace the abstract event A with our model parameters, θ, and the event B with our observed data, D. This reframing gives us Bayes' theorem for statistical inference:

$$\underbrace{\mathbb{P}(\theta|D)}_{\text{Posterior}} = \frac{\overbrace{\mathbb{P}(D|\theta)}^{\text{Likelihood}}\ \overbrace{\mathbb{P}(\theta)}^{\text{Prior}}}{\underbrace{\mathbb{P}(D)}_{\text{Evidence}}}.$$

Let's carefully define each component in this formulation:

- **Prior Probability**, $\mathbb{P}(\theta)$: This term represents our belief about the parameters θ *before* we observe any data. It is our *prior* knowledge. This could be based on previous experiments, domain expertise, or it could be chosen to be "uninformative" if we have no strong prior beliefs.

- **Likelihood**, $\mathbb{P}(D|\theta)$: This is exactly the likelihood function we defined in Section 3. It is the probability of observing our data D for a given set of parameters θ. It quantifies how well the data supports a particular parameter value.

- **Posterior Probability**, $\mathbb{P}(\theta|D)$: This is the quantity we want to compute. It represents our updated belief about the parameters θ *after* having observed the data D. It is a full probability distribution for the parameters, not just a point estimate.

- **Evidence** (or Marginal Likelihood), $\mathbb{P}(D)$: This is the probability of observing the data, averaged over all possible values of the parameters. For continuous θ, it is calculated by integrating the likelihood times the prior over the entire parameter space:

$$\mathbb{P}(D) = \int \mathbb{P}(D|\theta)\mathbb{P}(\theta)\, d\theta.$$

  The evidence acts as a normalization constant. It ensures that the posterior distribution integrates to 1, making it a valid probability distribution. In many applications, the evidence is computationally very difficult to calculate. However, for the purpose of finding the most probable parameter value, we can often ignore it, as it does not depend on θ:

$$\mathbb{P}(\theta|D) \propto \mathbb{P}(D|\theta)\mathbb{P}(\theta).$$

  This relationship states that the posterior is proportional to the likelihood times the prior. The parameter value that maximizes the posterior probability is known as the **Maximum A Posteriori (MAP)** estimate.

Bayesian inference is thus a process:

Prior Belief + Observed Data (via Likelihood) → Updated Posterior Belief.

This provides a natural and coherent framework for learning from data by continuously updating our knowledge.

Example 4.12 (Bayesian Coin Flipping)**.** Suppose we have a coin and want to estimate its probability of landing heads, θ. Before flipping, we express our prior belief using a **Beta distribution**: $\theta \sim \text{Beta}(\alpha, \beta)$, which has PDF:

$$\mathbb{P}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \theta \in [0,1].$$

A common choice is $\alpha = \beta = 1$, which gives a uniform prior (no preference for any value of θ).

We flip the coin n times and observe h heads and $t = n - h$ tails. The likelihood is Binomial:

$$\mathbb{P}(D|\theta) = \binom{n}{h}\theta^h(1-\theta)^t \propto \theta^h(1-\theta)^t.$$

Using Bayes' theorem, the posterior is:

$$\mathbb{P}(\theta|D) \propto \mathbb{P}(D|\theta)\mathbb{P}(\theta) \propto \theta^h(1-\theta)^t \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{h+\alpha-1}(1-\theta)^{t+\beta-1}.$$

This is another Beta distribution: $\theta|D \sim \text{Beta}(\alpha + h, \beta + t)$.

Numerical example: Start with a uniform prior $\text{Beta}(1,1)$. After observing 7 heads in 10 flips:

- **Prior:** $\text{Beta}(1,1)$ with mean $= 0.5$
- **Posterior:** $\text{Beta}(1+7, 1+3) = \text{Beta}(8,4)$ with mean $= 8/12 \approx 0.667$
- **MAP estimate:** $\hat{\theta}_{\text{MAP}} = (8-1)/(8+4-2) = 7/10 = 0.7$
- **MLE estimate:** $\hat{\theta}_{\text{MLE}} = 7/10 = 0.7$

With a uniform prior, MAP equals MLE. However, with an informative prior like $\text{Beta}(5,5)$ (believing the coin is likely fair), the posterior would be $\text{Beta}(12,8)$ with mean 0.6—the prior *pulls* our estimate toward 0.5.

Example 4.13 (Bayesian Estimation of a Normal Mean)**.** Suppose we observe data $D = \{x_1, \ldots, x_n\}$ from a normal distribution with unknown mean μ and known variance σ^2. We place a normal prior on μ:

$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2),$$

where μ_0 is our prior belief about the mean and τ_0^2 reflects our uncertainty.

The likelihood for the observed data is:

$$\mathbb{P}(D|\mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

After applying Bayes' theorem and completing the square, the posterior is also normal:

$$\mu|D \sim \mathcal{N}(\mu_n, \tau_n^2),$$

where the posterior parameters are:

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \qquad \tau_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)^{-1}.$$

Numerical example: Suppose we believe the average height of adults is around $\mu_0 = 170$ cm with prior uncertainty $\tau_0 = 5$ cm. We collect $n = 25$ measurements with sample mean $\bar{x} = 175$ cm and known $\sigma = 10$ cm.

- **Prior precision:** $1/\tau_0^2 = 1/25 = 0.04$
- **Data precision:** $n/\sigma^2 = 25/100 = 0.25$
- **Posterior precision:** $1/\tau_n^2 = 0.04 + 0.25 = 0.29$, so $\tau_n^2 \approx 3.45$
- **Posterior mean:** $\mu_n = \frac{0.04\times170+0.25\times175}{0.29} = \frac{6.8+43.75}{0.29} \approx 174.3$ cm

The posterior mean (174.3 cm) is a precision-weighted average of the prior mean (170 cm) and the sample mean (175 cm). Since the data has higher precision (more informative), the posterior is pulled closer to $\bar{x}$. As $n \to \infty$, the posterior converges to the MLE.

4.3 The Naive Bayes Classifier

To make these ideas concrete, let's see how Bayes' theorem directly leads to a simple but surprisingly effective classification algorithm that you will implement in the workshop: the Naive Bayes classifier.

The goal of a classifier is to predict a class label C_k (from a set of K possible classes) for a given observation, which is described by a feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$. The Bayesian approach is to calculate the posterior probability of each class given the features, $\mathbb{P}(C_k|\mathbf{x})$, and then predict the class with the highest posterior probability (this is the MAP decision rule).

Using Bayes' theorem, the posterior for class C_k is:

$$\mathbb{P}(C_k|\mathbf{x}) = \frac{\mathbb{P}(\mathbf{x}|C_k)\mathbb{P}(C_k)}{\mathbb{P}(\mathbf{x})}$$

To make a prediction, we compute this for every class $k = 1, \ldots, K$ and choose the class that maximizes it:

$$\hat{C} = \arg\max_{k\in\{1,\ldots,K\}} \frac{\mathbb{P}(\mathbf{x}|C_k)\mathbb{P}(C_k)}{\mathbb{P}(\mathbf{x})}$$

As noted before, the evidence term $\mathbb{P}(\mathbf{x})$ is the same for all classes, so it doesn't affect the argmax. We can therefore simplify the decision rule to:

$$\hat{C} = \arg\max_{k\in\{1,\ldots,K\}} \mathbb{P}(\mathbf{x}|C_k)\mathbb{P}(C_k)$$

This requires us to estimate two quantities from our training data:

1. $\mathbb{P}(C_k)$: The class prior. This is easy to estimate as the proportion of training examples belonging to class C_k.
2. $\mathbb{P}(\mathbf{x}|C_k)$: The class-conditional likelihood of the features. This is the probability of observing the feature vector $\mathbf{x}$ for an item in class C_k.

Estimating $\mathbb{P}(\mathbf{x}|C_k)$ is the hard part. It is a joint probability distribution over a p-dimensional feature space. If the features are continuous, this requires estimating a p-dimensional density function, which is notoriously difficult and requires a huge amount of data (this is related to the *curse of dimensionality*).

This is where the *naive* assumption comes in. The naive Bayes classifier makes a bold and simplifying assumption of *conditional independence* of the features, given the class.

Definition 4.14 (Conditional Independence Assumption)**.** The features $x_1, x_2, \ldots, x_p$ are assumed to be conditionally independent given the class C_k. Mathematically, this means

$$\mathbb{P}(\mathbf{x}|C_k) = \mathbb{P}(x_1, x_2, \ldots, x_p|C_k) = \prod_{j=1}^{p} \mathbb{P}(x_j|C_k)$$

This assumption is *naive* because in most real-world datasets, features are rarely truly independent (e.g., in a medical dataset, height and weight are correlated). However, this simplification is what makes the algorithm so efficient. It decomposes the hard problem of estimating one high-dimensional joint distribution into the much easier problem of estimating p separate one-dimensional distributions, $\mathbb{P}(x_j|C_k)$.

Substituting this assumption back into our decision rule gives the final form of the naive Bayes classifier

$$\hat{C} = \arg \max_{k\in\{1,\ldots,K\}} \mathbb{P}(C_k) \prod_{j=1}^{p} \mathbb{P}(x_j|C_k).$$

In practice, to avoid numerical underflow from multiplying many small probabilities, we often work with the log of this expression:

$$\hat{C} = \arg \max_{k\in\{1,\ldots,K\}} \left(\log \mathbb{P}(C_k) + \sum_{j=1}^{p} \log \mathbb{P}(x_j|C_k) \right).$$

To implement this, we simply need to choose a plausible distribution for each feature x_j within each class C_k (e.g., a Gaussian for continuous features, or a Bernoulli/Multinomial for discrete features), estimate its parameters from the training data, and then use the formula above to classify new, unseen data points. This is exactly the task you will undertake in the workshop.

Example 4.15 (Binary Classification: Email Spam Detection)**.** Consider a spam filter with two classes: C_1 = Spam and C_2 = Not Spam. Each email is represented by binary features indicating the presence of certain words: $\mathbf{x} = (x_1, x_2, x_3)$ where $x_1 = 1$ if "free" appears, $x_2 = 1$ if "meeting" appears, and $x_3 = 1$ if "winner" appears.

From training data, we estimate the following probabilities:

	Spam (C_1)	Not Spam (C_2)
Class prior $\mathbb{P}(C_k)$	0.40	0.60
$\mathbb{P}(x_1 = 1|C_k)$ ("free")	0.80	0.10
$\mathbb{P}(x_2 = 1|C_k)$ ("meeting")	0.10	0.70
$\mathbb{P}(x_3 = 1|C_k)$ ("winner")	0.70	0.05

Now suppose we receive a new email containing "free" and "winner" but not "meeting": $\mathbf{x} = (1, 0, 1)$. We compute the (unnormalized) posterior for each class:

$$\begin{aligned}\mathbb{P}(C_1)\prod_{j=1}^{3}\mathbb{P}(x_j|C_1) &= 0.40 \times 0.80 \times (1 - 0.10) \times 0.70 \\ &= 0.40 \times 0.80 \times 0.90 \times 0.70 = 0.2016\end{aligned}$$

$$\begin{aligned}\mathbb{P}(C_2)\prod_{j=1}^{3}\mathbb{P}(x_j|C_2) &= 0.60 \times 0.10 \times (1 - 0.70) \times 0.05 \\ &= 0.60 \times 0.10 \times 0.30 \times 0.05 = 0.0009\end{aligned}$$

Since $0.2016 \gg 0.0009$, we classify this email as **Spam**. To compute the actual posterior probability:

$$\mathbb{P}(\text{Spam}|\mathbf{x}) = \frac{0.2016}{0.2016 + 0.0009} = \frac{0.2016}{0.2025} \approx 0.996$$

The classifier is 99.6% confident this email is spam.

Example 4.16 (Multi-class Classification: Iris Flower Species)**.** The classic Iris dataset involves classifying flowers into three species: C_1 = Setosa, C_2 = Versicolor, and C_3 = Virginica, based on four continuous features: sepal length (x_1), sepal width (x_2), petal length (x_3), and petal width (x_4), all measured in centimeters.

For continuous features, we use **Gaussian Naive Bayes**, assuming each feature follows a normal distribution within each class:

$$\mathbb{P}(x_j|C_k) = \frac{1}{\sqrt{2\pi\sigma_{jk}^2}} \exp\left(-\frac{(x_j - \mu_{jk})^2}{2\sigma_{jk}^2}\right)$$

where μ_{jk} and σ_{jk}^2 are the mean and variance of feature j for class k, estimated from training data.

Suppose from training data we estimate (showing only petal measurements for brevity):

	Setosa (C_1)	Versicolor (C_2)	Virginica (C_3)
Class prior $\mathbb{P}(C_k)$	0.33	0.33	0.34
Petal length: μ_{3k}	1.46 cm	4.26 cm	5.55 cm
Petal length: σ_{3k}	0.17 cm	0.47 cm	0.55 cm
Petal width: μ_{4k}	0.24 cm	1.33 cm	2.03 cm
Petal width: σ_{4k}	0.11 cm	0.20 cm	0.27 cm

Now suppose we observe a new flower with petal length $x_3 = 4.5$ cm and petal width $x_4 = 1.5$ cm. Using only these two features:

$$\begin{aligned}\mathbb{P}(x_3 = 4.5|C_1) &= \frac{1}{\sqrt{2\pi(0.17)^2}} \exp\left(-\frac{(4.5 - 1.46)^2}{2(0.17)^2}\right) \approx 0 \text{ (essentially zero)} \\ \mathbb{P}(x_3 = 4.5|C_2) &= \frac{1}{\sqrt{2\pi(0.47)^2}} \exp\left(-\frac{(4.5 - 4.26)^2}{2(0.47)^2}\right) \approx 0.801 \\ \mathbb{P}(x_3 = 4.5|C_3) &= \frac{1}{\sqrt{2\pi(0.55)^2}} \exp\left(-\frac{(4.5 - 5.55)^2}{2(0.55)^2}\right) \approx 0.199\end{aligned}$$

Similarly for petal width:

$$\mathbb{P}(x_4 = 1.5|C_1) \approx 0, \quad \mathbb{P}(x_4 = 1.5|C_2) \approx 1.494, \quad \mathbb{P}(x_4 = 1.5|C_3) \approx 0.540$$

Computing the unnormalized posteriors (ignoring C_1 since its likelihood is essentially zero):

$$\mathbb{P}(C_2) \cdot \mathbb{P}(x_3|C_2) \cdot \mathbb{P}(x_4|C_2) = 0.33 \times 0.801 \times 1.494 \approx 0.395$$
$$\mathbb{P}(C_3) \cdot \mathbb{P}(x_3|C_3) \cdot \mathbb{P}(x_4|C_3) = 0.34 \times 0.199 \times 0.540 \approx 0.037$$

Since $0.395 > 0.037$, we classify this flower as **Versicolor**. The normalized posterior probability is:

$$\mathbb{P}(\text{Versicolor}|\mathbf{x}) \approx \frac{0.395}{0.395 + 0.037} \approx 0.914$$

The classifier is about 91.4% confident this flower is Versicolor.

5 Conclusion

This lecture has marked a pivotal transition in our course, moving from the deterministic, geometric world of vector spaces and projections to the uncertain, probabilistic framework required for modern data analysis. We have laid the foundational statistical language necessary to describe, model, and ultimately learn from data in the presence of real-world noise and variability. By embracing probability theory, we have unlocked a far more powerful and realistic perspective on the nature of data modeling.

5.1 Summary of Key Concepts

The journey today has equipped us with three core pillars of probabilistic modeling:

1. **Distributions as Models**: We have established that the fundamental assumption in statistical learning is that our observed data are realizations of random variables drawn from some underlying probability distribution. We have rigorously defined the key distributions that serve as the building blocks for a vast array of models: the Bernoulli distribution for binary outcomes (classification) and the Gaussian (Normal) distribution for continuous variables, whose central role is justified by both its mathematical properties and the Central Limit Theorem.

2. **Parameter Estimation via Maximum Likelihood**: We have introduced a powerful, general principle for fitting models to data: Maximum Likelihood Estimation (MLE). By defining the likelihood function $L(\theta|D)$, we can systematically find the parameter vector $\hat{\theta}_{MLE}$ that makes our observed data most probable. Critically, we demonstrated that the Ordinary Least Squares (OLS) solution, derived geometrically in Week 2, is mathematically equivalent to the MLE solution under the common assumption of i.i.d. Gaussian errors. This result provides a profound statistical justification for the method of least squares.

3. **Quantifying Relationships and Uncertainty**: We have defined the moments of a distribution, focusing on expectation ($\mathbb{E}[X]$) as the measure of central tendency and variance ($\text{Var}(X)$) as the measure of spread or uncertainty. Furthermore, we generalized these concepts to multiple dimensions through the covariance ($\text{Cov}(X, Y)$) and the covariance matrix (Σ). In doing so, we forged a crucial link back to our geometric foundations, showing that the sample covariance is directly proportional to the dot product of centered data vectors, thereby unifying statistical correlation with geometric alignment.

5.2 Bridge to the Workshop: Applying Theory to Practice

The theoretical concepts developed today are not mere abstractions; they are the direct tools you will use to build a functioning machine learning model. In this week's workshop, you will implement the **Naive Bayes classifier**. This algorithm is a beautiful and direct application of the principles we've discussed:

- It is built directly upon **Bayes' Theorem**, using the formula $\mathbb{P}(C_k|\mathbf{x}) \propto \mathbb{P}(\mathbf{x}|C_k)\mathbb{P}(C_k)$ to find the most probable class for a given data point.
- It requires you to model the **distribution** of your features. You will have to make modeling choices, for instance, by assuming that the continuous features in your dataset follow a Gaussian distribution within each class.
- You will use the training data to **estimate the parameters** of these distributions (the means and variances for the Gaussians, and the class priors $\mathbb{P}(C_k)$) in a manner analogous to MLE, which are then used to make predictions on new data.

This exercise will solidify your understanding of how abstract probability theory translates into a concrete, predictive algorithm.

5.3 Bridge to Lesson 4: The Problem of Model Evaluation

We have now developed a principled method for fitting a model to data. However, this raises a deeper and more subtle set of questions that will motivate our next lecture. Now that we can estimate a model parameter ($\hat{\theta}$) and quantify the inherent uncertainty in our data (variance), how do we evaluate the quality of our model? How do we know if our model is good? More specifically, how do we determine if the model is too complex for our dataset, fitting to the noise rather than the signal (overfitting), or too simple, failing to capture the underlying structure (underfitting)?

This line of questioning leads directly to the core concepts of **Model Evaluation and Statistical Inference**. We will formally introduce the **Bias-Variance Tradeoff**, a fundamental dilemma in machine learning that describes the tension between model complexity and its ability to generalize to new, unseen data. We will use the concepts of expectation and variance from this lecture to precisely define bias and variance in a modeling context. Finally, we will introduce a suite of key metrics for quantitatively measuring model performance, such as Mean Squared Error (MSE) and R^2 for regression, and Accuracy, Precision, and Recall for classification, providing us with the tools for rigorous model selection and comparison.

Exercise

1. **(Measure Theory)** Let $\Omega = \{a, b, c, d\}$. Determine whether the following collections of subsets are σ-algebras on Ω. If not, explain which axiom fails.

   (a) $\mathcal{F}_1 = \{\emptyset, \{a\}, \{b, c, d\}, \Omega\}$

   (b) $\mathcal{F}_2 = \{\emptyset, \{a, b\}, \{c, d\}, \{a, c\}, \Omega\}$

   (c) $\mathcal{F}_3 = \{\emptyset, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$

2. **(Lebesgue Integral)** Consider the function $f : [0, 2] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 3 & \text{if } x \in [0, 1) \cap \mathbb{Q}, \\ 1 & \text{if } x \in [0, 1) \setminus \mathbb{Q}, \\ 2 & \text{if } x \in [1, 2]. \end{cases}$$

   Compute the Lebesgue integral $\int_{[0,2]} f \, \mathrm{d}\lambda$, where λ is the Lebesgue measure. Is this function Riemann integrable? Justify your answer.

3. **(Probability Space Construction)** A biased die is rolled, where the probability of rolling face k is proportional to k (i.e., $\mathbb{P}(\{k\}) \propto k$ for $k \in \{1, 2, 3, 4, 5, 6\}$).

   (a) Determine the exact probability of each outcome.

   (b) Define a random variable X that equals 1 if the outcome is prime and 0 otherwise. Find the PMF of X.

   (c) Compute $\mathbb{E}[X]$ and $\mathrm{Var}(X)$.

4. **(Gaussian Properties)** Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Prove that the standardized variable $Z = \dfrac{X - \mu}{\sigma}$ follows a standard normal distribution $\mathcal{N}(0, 1)$. Then use this result to prove that for any $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = \mathbb{P}(-2 \leq Z \leq 2) \approx 0.9545.$$

5. **(Expectation of Functions)** Let X be a continuous random variable with PDF $f_X(x) = 2x$ for $x \in [0, 1]$ and $f_X(x) = 0$ otherwise.

   (a) Verify that f_X is a valid PDF.

   (b) Compute $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, and $\mathrm{Var}(X)$.

   (c) Find $\mathbb{E}[e^X]$.

   (d) If $Y = 3X^2 - 2X + 1$, find $\mathbb{E}[Y]$.

6. **(Binomial Distribution)** In a clinical trial, a new drug has a 70% success rate. If 15 patients are treated independently:

   (a) What is the probability that exactly 12 patients respond successfully?

   (b) What is the probability that at least 10 patients respond successfully?

   (c) Find the expected number of successes and the standard deviation.

   (d) Using the normal approximation with continuity correction, estimate $\mathbb{P}(X \geq 10)$ and compare with the exact value.

7. **(Variance of Sums)** Let $X_1, X_2, \ldots, X_n$ be independent random variables with $\mathbb{E}[X_i] = \mu_i$ and $\text{Var}(X_i) = \sigma_i^2$. Prove that

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \sigma_i^2.$$

Then, show that if the X_i are identically distributed with common variance σ^2, the variance of the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is $\dfrac{\sigma^2}{n}$.

8. **(Covariance Properties)** Let X and Y be random variables. Prove the following properties:

   (a) $\text{Cov}(X, X) = \text{Var}(X)$.

   (b) $\text{Cov}(aX + b, cY + d) = ac \cdot \text{Cov}(X, Y)$ for constants $a, b, c, d \in \mathbb{R}$.

   (c) $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y)$.

   (d) If X and Y are independent, then $\text{Cov}(X, Y) = 0$.

9. **(MLE for Gaussian)** Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations from $\mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown.

   (a) Write down the likelihood function $L(\mu, \sigma^2|\mathbf{x})$.

   (b) Derive the log-likelihood function.

   (c) Show that the MLE for μ is $\hat{\mu}_{MLE} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

   (d) Show that the MLE for σ^2 is $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

   (e) Why is the MLE for variance biased? What is the unbiased estimator?

10. **(MLE for Exponential)** Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. from an exponential distribution with rate parameter $\lambda > 0$, so the PDF is $f(x|\lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$.

    (a) Write down the likelihood and log-likelihood functions.

    (b) Derive the MLE $\hat{\lambda}_{MLE}$.

    (c) If the observed data are $\{1.2, 0.8, 2.1, 1.5, 0.9\}$, compute $\hat{\lambda}_{MLE}$.

    (d) Find $\mathbb{E}[\hat{\lambda}_{MLE}]$ and determine whether it is an unbiased estimator of λ.

11. **(Central Limit Theorem Application)** A factory produces bolts with weights that have mean $\mu = 50$ grams and standard deviation $\sigma = 4$ grams (distribution unknown). A random sample of $n = 64$ bolts is selected.

    (a) Using the CLT, what is the approximate distribution of the sample mean $\bar{X}$?

    (b) Find the probability that the sample mean is between 49 and 51 grams.

    (c) Find the value c such that $\mathbb{P}(\bar{X} > c) = 0.05$.

    (d) How large should n be so that $\mathbb{P}(|\bar{X} - 50| < 0.5) \geq 0.99$?

12. **(Bayes' Theorem)** A factory has three machines (A, B, C) producing items. Machine A produces 50% of items with 2% defect rate, Machine B produces 30% with 3% defect rate, and Machine C produces 20% with 5% defect rate.

    (a) What is the probability that a randomly selected item is defective?

    (b) If an item is found to be defective, what is the probability it came from each machine?

    (c) Two items are selected and both are defective. What is the probability that both came from Machine C? (Assume independence.)

13. **(Bayesian Inference)** A coin has unknown probability θ of landing heads. We use a Beta$(2,2)$ prior for θ, which has PDF $f(\theta) = 6\theta(1-\theta)$ for $\theta \in [0,1]$.

    (a) Compute the prior mean and variance of θ.

    (b) After observing 8 heads in 10 flips, derive the posterior distribution of θ.

    (c) Compute the posterior mean and compare it to the MLE.

    (d) Find the MAP estimate and show how it differs from both the posterior mean and MLE.

    (e) Compute a 95% credible interval for θ (you may express this in terms of the Beta distribution quantiles).

14. **(Naive Bayes Classification)** Consider a classification problem with two classes (C_1 and C_2) and two continuous features (x_1, x_2). The class priors are $\mathbb{P}(C_1) = 0.4$ and $\mathbb{P}(C_2) = 0.6$. Within each class, the features are assumed independent and normally distributed:

	μ_1	σ_1	μ_2	σ_2
Class C_1	2	1	5	2
Class C_2	4	1.5	3	1

    (a) For a new observation $\mathbf{x} = (3,4)$, compute the (unnormalized) posterior probability for each class.

    (b) Classify the observation and compute the probability of correct classification.

    (c) Find the decision boundary in the (x_1, x_2) plane (the set of points where both classes are equally likely).

15. **(Comprehensive Problem)** Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli(p) random variables representing whether each of n website visitors clicks an advertisement.

    (a) Write the joint PMF of $(X_1, \ldots, X_n)$ and show it depends on the data only through $T = \sum_{i=1}^{n} X_i$.

    (b) Show that $T \sim$ Binomial(n,p) and derive $\mathbb{E}[T]$ and Var(T).

    (c) Using MLE, show that $\hat{p} = T/n$ and prove that $\hat{p}$ is an unbiased estimator of p.

    (d) Using Chebyshev's inequality, show that for any $\epsilon > 0$:

$$\mathbb{P}(|\hat{p} - p| \geq \epsilon) \leq \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2}.$$

    (e) Apply the CLT to show that for large n, $\hat{p} \overset{\text{approx}}{\sim} \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$.

    (f) If we observe 45 clicks out of 500 visitors, construct an approximate 95% confidence interval for p.

Solution to exercise

1. **(Measure Theory)**

   (a) $\mathcal{F}_1 = \{\emptyset, \{a\}, \{b,c,d\}, \Omega\}$ **is a σ-algebra**. We verify:

   - Contains Ω: Yes, $\Omega = \{a,b,c,d\} \in \mathcal{F}_1$.
   - Closed under complementation: $\{a\}^c = \{b,c,d\} \in \mathcal{F}_1$ and $\{b,c,d\}^c = \{a\} \in \mathcal{F}_1$. Also $\emptyset^c = \Omega$ and $\Omega^c = \emptyset$.
   - Closed under countable unions: All possible unions of elements in $\mathcal{F}_1$ yield sets already in $\mathcal{F}_1$.

   (b) $\mathcal{F}_2 = \{\emptyset, \{a,b\}, \{c,d\}, \{a,c\}, \Omega\}$ **is NOT a σ-algebra**.
   It fails closure under complementation: $\{a,c\}^c = \{b,d\} \notin \mathcal{F}_2$.
   Alternatively, it fails closure under union: $\{a,b\} \cup \{a,c\} = \{a,b,c\} \notin \mathcal{F}_2$.

   (c) $\mathcal{F}_3 = \{\emptyset, \{a,b\}, \{c,d\}, \{a,b,c,d\}\}$ **is a σ-algebra**. This is the σ-algebra generated by the partition $\{\{a,b\},\{c,d\}\}$. We verify:

   - $\Omega = \{a,b,c,d\} \in \mathcal{F}_3$.
   - $\{a,b\}^c = \{c,d\} \in \mathcal{F}_3$ and vice versa.
   - All unions remain in $\mathcal{F}_3$: $\{a,b\} \cup \{c,d\} = \Omega \in \mathcal{F}_3$.

2. **(Lebesgue Integral)**

   The function f takes three values: 3 on $[0,1) \cap \mathbb{Q}$, 1 on $[0,1) \setminus \mathbb{Q}$, and 2 on $[1,2]$.

   Computing the Lebesgue integral:

$$\begin{aligned}\int_{[0,2]} f \, \mathrm{d}\lambda &= 3 \cdot \lambda([0,1) \cap \mathbb{Q}) + 1 \cdot \lambda([0,1) \setminus \mathbb{Q}) + 2 \cdot \lambda([1,2]) \\ &= 3 \cdot 0 + 1 \cdot 1 + 2 \cdot 1 \\ &= 0 + 1 + 2 = \boxed{3}\end{aligned}$$

   Here we used that $\lambda(\mathbb{Q} \cap [0,1)) = 0$ (rationals have measure zero) and $\lambda([0,1) \setminus \mathbb{Q}) = 1 - 0 = 1$.

   Riemann integrability: The function is **NOT Riemann integrable** on $[0,1)$. For any partition of $[0,1)$, each subinterval contains both rationals and irrationals (both sets are dense in $\mathbb{R}$). Therefore, the upper Riemann sum over $[0,1)$ equals $3 \cdot 1 = 3$ while the lower sum equals $1 \cdot 1 = 1$. Since they don't converge to the same value, the Riemann integral doesn't exist on $[0,1)$.

3. **(Probability Space Construction)**

   (a) Since $\mathbb{P}(\{k\}) \propto k$, we have $\mathbb{P}(\{k\}) = \frac{k}{C}$ where $C = 1+2+3+4+5+6 = 21$. Therefore:

$$\mathbb{P}(\{1\}) = \frac{1}{21}, \quad \mathbb{P}(\{2\}) = \frac{2}{21}, \quad \mathbb{P}(\{3\}) = \frac{3}{21} = \frac{1}{7},$$
$$\mathbb{P}(\{4\}) = \frac{4}{21}, \quad \mathbb{P}(\{5\}) = \frac{5}{21}, \quad \mathbb{P}(\{6\}) = \frac{6}{21} = \frac{2}{7}.$$

   (b) The prime outcomes are $\{2,3,5\}$. The random variable X is defined as:

$$X(\omega) = \begin{cases} 1 & \text{if } \omega \in \{2,3,5\}, \\ 0 & \text{otherwise.} \end{cases}$$

The PMF of X is:

$$\mathbb{P}(X=1) = \mathbb{P}(\{2,3,5\}) = \frac{2+3+5}{21} = \frac{10}{21},$$
$$\mathbb{P}(X=0) = 1 - \frac{10}{21} = \frac{11}{21}.$$

(c) Computing the moments:

$$\mathbb{E}[X] = 0 \cdot \frac{11}{21} + 1 \cdot \frac{10}{21} = \frac{10}{21} \approx 0.476.$$

$$\mathbb{E}[X^2] = 0^2 \cdot \frac{11}{21} + 1^2 \cdot \frac{10}{21} = \frac{10}{21}.$$

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{10}{21} - \frac{100}{441} = \frac{210-100}{441} = \frac{110}{441} \approx 0.249.$$

4. **(Gaussian Properties)**

Proof that $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$:

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with PDF $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

For the transformation $Z = g(X) = \frac{X-\mu}{\sigma}$, the inverse is $X = h(Z) = \sigma Z + \mu$, and $\frac{dh}{dZ} = \sigma$.

Using the change of variables formula:

$$\begin{aligned} f_Z(z) &= f_X(h(z)) \cdot \left|\frac{dh}{dz}\right| = f_X(\sigma z + \mu) \cdot \sigma \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\sigma z + \mu - \mu)^2}{2\sigma^2}\right) \cdot \sigma \\ &= \frac{\sigma}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\sigma^2 z^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right). \end{aligned}$$

This is exactly the PDF of $\mathcal{N}(0,1)$. □

Application:

$$\begin{aligned} \mathbb{P}(\mu - 2\sigma \le X \le \mu + 2\sigma) &= \mathbb{P}\left(\frac{\mu - 2\sigma - \mu}{\sigma} \le Z \le \frac{\mu + 2\sigma - \mu}{\sigma}\right) \\ &= \mathbb{P}(-2 \le Z \le 2) \\ &= \Phi(2) - \Phi(-2) = 2\Phi(2) - 1 \\ &\approx 2(0.9772) - 1 = 0.9545. \end{aligned}$$

5. **(Expectation of Functions)**

(a) **Verification:** We check non-negativity and normalization.

- Non-negativity: $f_X(x) = 2x \ge 0$ for $x \in [0,1]$.
- Normalization: $\int_0^1 2x\,dx = [x^2]_0^1 = 1$. ✓

(b) **Computing moments:**

$$\mathbb{E}[X] = \int_0^1 x \cdot 2x\,dx = 2\int_0^1 x^2\,dx = 2 \cdot \frac{x^3}{3}\Big|_0^1 = \frac{2}{3}.$$

$$\mathbb{E}[X^2] = \int_0^1 x^2 \cdot 2x\,dx = 2\int_0^1 x^3\,dx = 2 \cdot \frac{x^4}{4}\Big|_0^1 = \frac{1}{2}.$$

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{1}{2} - \frac{4}{9} = \frac{9-8}{18} = \frac{1}{18}.$$

(c) **Computing $\mathbb{E}[e^X]$:**

$$\mathbb{E}[e^X] = \int_0^1 e^x \cdot 2x\,dx = 2\int_0^1 xe^x\,dx.$$

Using integration by parts with $u = x$, $dv = e^x dx$:

$$\int_0^1 xe^x\,dx = [xe^x]_0^1 - \int_0^1 e^x\,dx = e - [e^x]_0^1 = e - (e-1) = 1.$$

Therefore, $\mathbb{E}[e^X] = 2 \cdot 1 = 2$.

(d) **Computing $\mathbb{E}[Y]$:** For $Y = 3X^2 - 2X + 1$, using linearity:

$$\mathbb{E}[Y] = 3\mathbb{E}[X^2] - 2\mathbb{E}[X] + 1 = 3 \cdot \frac{1}{2} - 2 \cdot \frac{2}{3} + 1 = \frac{3}{2} - \frac{4}{3} + 1 = \frac{9-8+6}{6} = \frac{7}{6}.$$

6. **(Binomial Distribution)**

Let $X \sim \text{Binomial}(15, 0.7)$.

(a) $\mathbb{P}(X = 12) = \binom{15}{12}(0.7)^{12}(0.3)^3 = 455 \cdot 0.01384 \cdot 0.027 \approx \boxed{0.170}$.

(b) $\mathbb{P}(X \geq 10) = \sum_{k=10}^{15} \binom{15}{k}(0.7)^k(0.3)^{15-k}$.
Computing each term:

$$\begin{aligned} &\mathbb{P}(X = 10) \approx 0.206, \quad \mathbb{P}(X = 11) \approx 0.219, \quad \mathbb{P}(X = 12) \approx 0.170, \\ &\mathbb{P}(X = 13) \approx 0.092, \quad \mathbb{P}(X = 14) \approx 0.031, \quad \mathbb{P}(X = 15) \approx 0.005. \end{aligned}$$

Therefore, $\mathbb{P}(X \geq 10) \approx \boxed{0.722}$.

(c) $\mathbb{E}[X] = np = 15 \cdot 0.7 = \boxed{10.5}$.
$\sigma = \sqrt{np(1-p)} = \sqrt{15 \cdot 0.7 \cdot 0.3} = \sqrt{3.15} \approx \boxed{1.775}$.

(d) **Normal approximation with continuity correction:**
$X \overset{\text{approx}}{\sim} \mathcal{N}(10.5, 3.15)$. With continuity correction:

$$\mathbb{P}(X \geq 10) \approx \mathbb{P}(X > 9.5) = \mathbb{P}\left(Z > \frac{9.5 - 10.5}{1.775}\right) = \mathbb{P}(Z > -0.563).$$

$$= 1 - \Phi(-0.563) = \Phi(0.563) \approx 0.713.$$

This is close to the exact value of 0.722, with error about 1.2%.

7. **(Variance of Sums)**

Proof: Let $S = \sum_{i=1}^n X_i$ and $\mu_S = \mathbb{E}[S] = \sum_{i=1}^n \mu_i$ (by linearity).

$$\begin{aligned}\operatorname{Var}(S) = \mathbb{E}[(S-\mu_S)^2] &= \mathbb{E}\left[\left(\sum_{i=1}^{n}(X_i-\mu_i)\right)^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n}(X_i-\mu_i)^2 + 2\sum_{i<j}(X_i-\mu_i)(X_j-\mu_j)\right] \\ &= \sum_{i=1}^{n}\mathbb{E}[(X_i-\mu_i)^2] + 2\sum_{i<j}\mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)] \\ &= \sum_{i=1}^{n}\operatorname{Var}(X_i) + 2\sum_{i<j}\operatorname{Cov}(X_i,X_j).\end{aligned}$$

Since the X_i are independent, $\operatorname{Cov}(X_i, X_j) = 0$ for $i \neq j$. Therefore:

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\sigma_i^2. \quad \square$$

Variance of sample mean: If all $\sigma_i^2 = \sigma^2$, then:

$$\operatorname{Var}(\bar{X}) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}. \quad \square$$

8. **(Covariance Properties)**

   (a) $\operatorname{Cov}(X,X) = \mathbb{E}[(X-\mathbb{E}[X])(X-\mathbb{E}[X])] = \mathbb{E}[(X-\mathbb{E}[X])^2] = \operatorname{Var}(X). \quad \square$

   (b) Let $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. Then $\mathbb{E}[aX+b] = a\mu_X + b$ and $\mathbb{E}[cY+d] = c\mu_Y + d$.

   $$\begin{aligned}\operatorname{Cov}(aX+b, cY+d) &= \mathbb{E}[(aX+b-a\mu_X-b)(cY+d-c\mu_Y-d)] \\ &= \mathbb{E}[a(X-\mu_X)\cdot c(Y-\mu_Y)] \\ &= ac\cdot\mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = ac\cdot\operatorname{Cov}(X,Y). \quad \square\end{aligned}$$

   (c)

   $$\begin{aligned}\operatorname{Var}(X+Y) &= \mathbb{E}[(X+Y-\mathbb{E}[X]-\mathbb{E}[Y])^2] \\ &= \mathbb{E}[((X-\mu_X)+(Y-\mu_Y))^2] \\ &= \mathbb{E}[(X-\mu_X)^2] + \mathbb{E}[(Y-\mu_Y)^2] + 2\mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\ &= \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y). \quad \square\end{aligned}$$

   (d) If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Therefore:

   $$\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0. \quad \square$$

9. **(MLE for Gaussian)**

   (a) **Likelihood:**

   $$L(\mu,\sigma^2|\mathbf{x}) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right).$$

(b) **Log-likelihood:**

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

(c) **MLE for μ:** Taking derivative w.r.t. μ:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right).$$

Setting to zero: $\sum_{i=1}^{n} x_i = n\hat{\mu}$, so $\hat{\mu}_{MLE} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. □

(d) **MLE for σ^2:** Taking derivative w.r.t. σ^2:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Setting to zero and substituting $\hat{\mu} = \bar{x}$:

$$\frac{n}{2\hat{\sigma}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2(\hat{\sigma}^2)^2} \implies \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \quad \square$$

(e) **Bias:** The MLE for variance is biased because:

$$\mathbb{E}[\hat{\sigma}^2_{MLE}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

The unbiased estimator is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

10. **(MLE for Exponential)**

(a) **Likelihood:**

$$L(\lambda|\mathbf{x}) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right).$$

Log-likelihood:

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

(b) **MLE:** Taking derivative:

$$\frac{d\ell}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i.$$

Setting to zero: $\hat{\lambda}_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$.

(c) **Numerical computation:** With data $\{1.2, 0.8, 2.1, 1.5, 0.9\}$:

$$\bar{x} = \frac{1.2 + 0.8 + 2.1 + 1.5 + 0.9}{5} = \frac{6.5}{5} = 1.3.$$

Therefore, $\hat{\lambda}_{MLE} = \frac{1}{1.3} \approx 0.769$.

(d) **Bias analysis:** The MLE $\hat{\lambda} = 1/\bar{X}$ is **biased**.
For $X_i \sim \text{Exponential}(\lambda)$, we have $\bar{X} \sim \text{Gamma}(n, n\lambda)$.
Using properties of the inverse: $\mathbb{E}[1/\bar{X}] = \frac{n\lambda}{n-1}$ for $n > 1$.
Thus $\mathbb{E}[\hat{\lambda}_{MLE}] = \frac{n}{n-1}\lambda \neq \lambda$, so the estimator is biased.
An unbiased estimator would be $\tilde{\lambda} = \frac{n-1}{n\bar{x}} = \frac{n-1}{\sum_{i=1}^{n} x_i}$.

11. **(Central Limit Theorem Application)**

(a) By CLT, $\bar{X} \overset{\text{approx}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) = \mathcal{N}\left(50, \frac{16}{64}\right) = \mathcal{N}(50, 0.25)$.

The standard error is $\sigma_{\bar{X}} = \frac{4}{\sqrt{64}} = 0.5$ grams.

(b)

$$\begin{aligned}\mathbb{P}(49 < \bar{X} < 51) &= \mathbb{P}\left(\frac{49-50}{0.5} < Z < \frac{51-50}{0.5}\right) \\ &= \mathbb{P}(-2 < Z < 2) = 2\Phi(2) - 1 \approx 0.9545.\end{aligned}$$

(c) We need $\mathbb{P}(\bar{X} > c) = 0.05$, so $\mathbb{P}(\bar{X} \leq c) = 0.95$.

$$\mathbb{P}\left(Z \leq \frac{c-50}{0.5}\right) = 0.95 \implies \frac{c-50}{0.5} = 1.645.$$

Therefore, $c = 50 + 0.5 \times 1.645 = 50.82$ grams.

(d) We need $\mathbb{P}(|\bar{X} - 50| < 0.5) \geq 0.99$.

$$\mathbb{P}\left(-\frac{0.5}{4/\sqrt{n}} < Z < \frac{0.5}{4/\sqrt{n}}\right) \geq 0.99.$$

$$2\Phi\left(\frac{0.5\sqrt{n}}{4}\right) - 1 \geq 0.99 \implies \Phi\left(\frac{\sqrt{n}}{8}\right) \geq 0.995.$$

From standard normal tables, $\Phi^{-1}(0.995) \approx 2.576$.

$$\frac{\sqrt{n}}{8} \geq 2.576 \implies \sqrt{n} \geq 20.61 \implies n \geq 424.7.$$

Therefore, $n \geq 425$ bolts are needed.

12. **(Bayes' Theorem)**

(a) **Probability of defective item:**

$$\begin{aligned}\mathbb{P}(\text{Def}) &= \mathbb{P}(\text{Def}|A)\mathbb{P}(A) + \mathbb{P}(\text{Def}|B)\mathbb{P}(B) + \mathbb{P}(\text{Def}|C)\mathbb{P}(C) \\ &= 0.02 \times 0.50 + 0.03 \times 0.30 + 0.05 \times 0.20 \\ &= 0.010 + 0.009 + 0.010 = 0.029 = 2.9\%.\end{aligned}$$

(b) **Posterior probabilities:**

$$\begin{aligned}\mathbb{P}(A|\text{Def}) &= \frac{\mathbb{P}(\text{Def}|A)\mathbb{P}(A)}{\mathbb{P}(\text{Def})} = \frac{0.02 \times 0.50}{0.029} = \frac{0.010}{0.029} \approx 0.345, \\ \mathbb{P}(B|\text{Def}) &= \frac{0.03 \times 0.30}{0.029} = \frac{0.009}{0.029} \approx 0.310, \\ \mathbb{P}(C|\text{Def}) &= \frac{0.05 \times 0.20}{0.029} = \frac{0.010}{0.029} \approx 0.345.\end{aligned}$$

Note: These sum to 1, as expected.

(c) **Both defective items from Machine C:**

Using independence, if two items are selected and both are defective:

$$\mathbb{P}(\text{both from } C|\text{both defective}) = [\mathbb{P}(C|\text{Def})]^2 \approx (0.345)^2 \approx 0.119.$$

13. **(Bayesian Inference)**

   (a) **Prior moments:** For $\theta \sim \text{Beta}(2,2)$:

$$\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta} = \frac{2}{4} = 0.5.$$

$$\text{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{2\cdot 2}{16\cdot 5} = \frac{4}{80} = 0.05.$$

   (b) **Posterior distribution:** With $h = 8$ heads and $t = 2$ tails from $n = 10$ flips:

$$\theta|D \sim \text{Beta}(\alpha+h, \beta+t) = \text{Beta}(2+8, 2+2) = \text{Beta}(10,4).$$

   (c) **Posterior mean:**

$$\mathbb{E}[\theta|D] = \frac{10}{10+4} = \frac{10}{14} = \frac{5}{7} \approx 0.714.$$

   MLE: $\hat{\theta}_{MLE} = \frac{h}{n} = \frac{8}{10} = 0.8$.
   The posterior mean (0.714) is pulled toward the prior mean (0.5) relative to the MLE (0.8).

   (d) **MAP estimate:** For $\text{Beta}(\alpha,\beta)$, the mode (MAP) is:

$$\hat{\theta}_{MAP} = \frac{\alpha-1}{\alpha+\beta-2} = \frac{10-1}{10+4-2} = \frac{9}{12} = 0.75.$$

   Summary: MLE = 0.8, MAP = 0.75, Posterior Mean = 0.714. The prior information pulls estimates toward 0.5.

   (e) **95% Credible interval:** The interval is $(q_{0.025}, q_{0.975})$ where q_p denotes the p-th quantile of Beta(10, 4).
   Using Beta distribution tables or software:

$$q_{0.025} \approx 0.491, \quad q_{0.975} \approx 0.897.$$

   The 95% credible interval is approximately $(0.491, 0.897)$.

14. **(Naive Bayes Classification)**

   (a) **Computing posteriors for $\mathbf{x} = (3,4)$:**
   For class C_1: $\mu_1 = 2, \sigma_1 = 1$ for x_1 and $\mu_2 = 5, \sigma_2 = 2$ for x_2.

$$\mathbb{P}(x_1 = 3|C_1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(3-2)^2}{2}\right) = \frac{1}{\sqrt{2\pi}} e^{-0.5} \approx 0.242,$$
$$\mathbb{P}(x_2 = 4|C_1) = \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{(4-5)^2}{8}\right) = \frac{1}{2\sqrt{2\pi}} e^{-0.125} \approx 0.176.$$

   Unnormalized posterior: $\mathbb{P}(C_1) \cdot \mathbb{P}(x_1|C_1) \cdot \mathbb{P}(x_2|C_1) = 0.4 \times 0.242 \times 0.176 \approx 0.0170$.
   For class C_2: $\mu_1 = 4, \sigma_1 = 1.5$ and $\mu_2 = 3, \sigma_2 = 1$.

$$\mathbb{P}(x_1 = 3|C_2) = \frac{1}{1.5\sqrt{2\pi}} \exp\left(-\frac{(3-4)^2}{4.5}\right) \approx 0.213,$$
$$\mathbb{P}(x_2 = 4|C_2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(4-3)^2}{2}\right) \approx 0.242.$$

   Unnormalized posterior: $0.6 \times 0.213 \times 0.242 \approx 0.0309$.

(b) **Classification:** Since $0.0309 > 0.0170$, classify as C_2.
Normalized posterior:

$$\mathbb{P}(C_2|\mathbf{x}) = \frac{0.0309}{0.0170 + 0.0309} = \frac{0.0309}{0.0479} \approx \boxed{0.645}.$$

(c) **Decision boundary:** The boundary occurs where:

$$\mathbb{P}(C_1)\mathbb{P}(x_1|C_1)\mathbb{P}(x_2|C_1) = \mathbb{P}(C_2)\mathbb{P}(x_1|C_2)\mathbb{P}(x_2|C_2).$$

Taking logarithms and simplifying (after substantial algebra):

$$\log(0.4) - \frac{(x_1-2)^2}{2} - \frac{(x_2-5)^2}{8} - \log(2) = \log(0.6) - \frac{(x_1-4)^2}{4.5} - \frac{(x_2-3)^2}{2} - \log(1.5).$$

This simplifies to a quadratic equation in x_1 and x_2:

$$0.278x_1^2 - 0.125x_2^2 - 0.778x_1 + 1.75x_2 + C = 0,$$

where C is a constant determined by the remaining terms. This is a hyperbola in the (x_1, x_2) plane.

15. **(Comprehensive Problem)**

(a) **Joint PMF:**

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i} = p^T(1-p)^{n-T},$$

where $T = \sum_{i=1}^{n} x_i$. The joint PMF depends only on T, not on the individual values. □

(b) **Distribution of T:** By definition, $T \sim \text{Binomial}(n, p)$.

$$\mathbb{E}[T] = np, \qquad \text{Var}(T) = np(1-p). \quad \square$$

(c) **MLE and unbiasedness:** The log-likelihood is:

$$\ell(p) = T \log p + (n-T)\log(1-p).$$

Taking derivative: $\frac{d\ell}{dp} = \frac{T}{p} - \frac{n-T}{1-p} = 0$. Solving: $\hat{p} = T/n$.
Unbiasedness: $\mathbb{E}[\hat{p}] = \mathbb{E}[T/n] = \frac{1}{n}\mathbb{E}[T] = \frac{np}{n} = p$. □

(d) **Chebyshev bound:**

$$\text{Var}(\hat{p}) = \text{Var}(T/n) = \frac{1}{n^2}\text{Var}(T) = \frac{p(1-p)}{n}.$$

By Chebyshev: $\mathbb{P}(|\hat{p} - p| \geq \epsilon) \leq \frac{\text{Var}(\hat{p})}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2}$.
Since $p(1-p) \leq \frac{1}{4}$ for all $p \in [0,1]$ (maximum at $p = 0.5$):

$$\mathbb{P}(|\hat{p} - p| \geq \epsilon) \leq \frac{1}{4n\epsilon^2}. \quad \square$$

(e) **CLT application:** By CLT, for large n:

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{d} \mathcal{N}(0,1).$$

Therefore, $\hat{p} \overset{\text{approx}}{\sim} \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$. □

(f) **95% Confidence interval:** With $T = 45$ and $n = 500$:

$$\hat{p} = \frac{45}{500} = 0.09.$$

Using the approximate variance $\hat{p}(1-\hat{p})/n = \frac{0.09 \times 0.91}{500} = 0.0001638$:

$$\text{SE} = \sqrt{0.0001638} \approx 0.0128.$$

The 95% CI is:

$$\hat{p} \pm 1.96 \times \text{SE} = 0.09 \pm 1.96 \times 0.0128 = 0.09 \pm 0.025.$$

The 95% confidence interval is approximately $(0.065, 0.115)$ or $(6.5\%, 11.5\%)$.