Data Science for Mathematicians

Lesson 3: Probabilistic Foundations of Modeling

Department of Mathematics and Computer Science

Outline

From Geometry to Probability

Motivation: Moving beyond deterministic geometry

- Previous approach: Data as fixed vectors, models as subspaces
- Reality: Data contains **uncertainty**, **measurement error**, **randomness**
- Need: Models that can **quantify uncertainty** and make **predictions**

Key transition:

$$\boxed{\text{Deterministic Geometry} \longrightarrow \text{Probabilistic Framework}}$$

σ-Algebras

Definition: σ-Algebra

Let Ω be a non-empty set. A collection $\mathcal{F}$ of subsets of Ω is a σ-**algebra** if:

1. $\Omega \in \mathcal{F}$ (contains whole space)
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ (closed under complement)
3. $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ (closed under countable unions)

Consequences: $\emptyset \in \mathcal{F}$, closed under countable intersections, set differences

Example

For $\Omega = \{1, 2, 3, 4\}$ and $A = \{1, 2\}$:

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

Borel σ-Algebra

Definition: Borel σ-Algebra

The **Borel σ-algebra** on $\mathbb{R}$, denoted $\mathcal{B}(\mathbb{R})$, is the smallest σ-algebra containing all open intervals (a, b).

$\mathcal{B}(\mathbb{R})$ **contains:**

- All open sets, closed sets
- All intervals: (a, b), $[a, b]$, $[a, b)$, $(a, b]$
- Countable sets: singletons $\{x\}$, $\mathbb{Z}$, $\mathbb{Q}$

Example: Common Borel Sets

- $[0, 1]$ (closed interval)
- $\mathbb{Q} = \bigcup_{q \in \mathbb{Q}} \{q\}$ (countable union)
- Cantor set

Measures and Lebesgue Measure

Definition: Measure

A function $\mu : \mathcal{F} \to [0, \infty]$ is a **measure** if:

1. $\mu(A) \geq 0$ for all $A \in \mathcal{F}$
2. $\mu(\emptyset) = 0$
3. $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$ for pairwise disjoint A_i

Definition: Lebesgue Measure

The **Lebesgue measure** λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$:

- $\lambda([a, b]) = b - a$ (length of interval)
- Translation-invariant: $\lambda(A + t) = \lambda(A)$

Example

$\lambda([2, 5]) = 3, \quad \lambda(\{x\}) = 0, \quad \lambda(\mathbb{Q} \cap [0, 1]) = 0$

The Lebesgue Integral

Definition: Simple Function

$\phi : \Omega \to \mathbb{R}$ is **simple** if it takes finitely many values:

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

Definition: Lebesgue Integral

Let $\phi : \Omega \to \mathbb{R}$ be **simple**, denote

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

Example

For $\phi(x) = 2 \cdot \mathbf{1}_{[0,1)} + 5 \cdot \mathbf{1}_{[1,2)} + 1 \cdot \mathbf{1}_{[2,3]}$:

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

Riemann vs. Lebesgue Integration

Theorem: Riemann vs. Lebesgue

Let $f : [a, b] \to \mathbb{R}$ be bounded. Then:

1. If f is Riemann integrable, then it is Lebesgue integrable and the integrals agree
2. f is Riemann integrable $\Leftrightarrow$ f is continuous almost everywhere

Example: Dirichlet Function – Lebesgue but not Riemann

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

Not Riemann integrable (discontinuous everywhere)
Lebesgue integrable:

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

Probability Space

Definition: Probability Space

A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where:

- Ω: **sample space** (all possible outcomes)
- $\mathcal{F}$: σ-**algebra** (collection of events)
- $\mathbb{P}$: **probability measure** with $\mathbb{P}(\Omega) = 1$

Example: Fair Coin Toss

- $\Omega = \{H, T\}$
- $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$ (power set)
- $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = 0.5$

Example: Fair Die

$\Omega = \{1, 2, 3, 4, 5, 6\}, \quad \mathbb{P}(\{k\}) = 1/6$ for each k

Random Variables

Definition: Random Variable

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A **random variable** is a function $X : \Omega \to \mathbb{R}$ such that for every Borel set $B \subset \mathbb{R}$:

$$\{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{F}$$

This property is called **measurability**.

Example: Coin Toss as a Random Variable

Example: Coin Toss as Random Variable

Recall our coin toss with $\Omega = \{H, T\}$. Define a random variable $X : \Omega \to \mathbb{R}$ by assigning numerical values:

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

This converts the abstract outcome "heads" into the number 1 and "tails" into 0.

Probabilistic computations:

- $\mathbb{P}(X = 1) = \mathbb{P}(\{H\}) = 0.5$
- $\mathbb{P}(X = 0) = \mathbb{P}(\{T\}) = 0.5$
- $\mathbb{P}(X \leq 0.5) = \mathbb{P}(\{T\}) = 0.5$

Key insight: This encoding is how we model binary outcomes (success/failure) in data science, leading directly to the Bernoulli distribution.

Example: Sum of Two Dice

Example: 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.
Define $S : \Omega \to \mathbb{R}$ as the sum: $S((i,j)) = i + j$

Key observation: S is not one-to-one—multiple outcomes map to the same value.

Example: $S = 7$ can occur via $(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)$

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

Other values:

- $\mathbb{P}(S = 2) = 1/36$ (only $(1,1)$)
- $\mathbb{P}(S = 12) = 1/36$ (only $(6,6)$)

Types of Random Variables

Discrete:

- Finite or countably infinite range
- Can ask $\mathbb{P}(X = x)$ for each value

Continuous:

- Uncountably infinite range
- $\mathbb{P}(X = x) = 0$ for any single value

Definition: PMF

Probability Mass Function:

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

1. $p_X(x) \geq 0$
2. $$\sum_x p_X(x) = 1$$

Definition: PDF

Probability Density Function:

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

1. $f_X(x) \geq 0$
2. $$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$$

Example: 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:

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

Verification: All values are non-negative, and $0.70 + 0.20 + 0.08 + 0.02 = 1$ ✓

Computing probabilities:

$$\mathbb{P}(X \geq 2) = p_X(2) + p_X(3) = 0.08 + 0.02 = 0.10$$

Example: Customer Arrivals

A coffee shop records the number of customers N arriving during each 5-minute interval. After analyzing many intervals, the shop determines the following PMF:

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

Verification: $0.10 + 0.25 + 0.30 + 0.20 + 0.10 + 0.05 = 1$ ✓

Computing probabilities:

$$\mathbb{P}(N \geq 3) = p_N(3) + p_N(4) + p_N(5) = 0.20 + 0.10 + 0.05 = 0.35$$

Example: Continuous Uniform Distribution

The waiting time T (in minutes) for a bus is uniformly distributed between 0 and 10 minutes. The PDF of T is:

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

Verification: $\int_{-\infty}^{\infty} f_T(t)\,\mathrm{d}t = \int_0^{10} 0.1\,\mathrm{d}t = 0.1 \times 10 = 1$ ✓

Probability of waiting 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: $\mathbb{P}(T = 3) = \int_3^3 f_T(t)\,\mathrm{d}t = 0$ (point probabilities are zero for continuous RVs)

Example: Exponential Distribution

The lifetime L (in years) of an electronic component is modeled by an exponential distribution with rate $\lambda = 0.5$. The PDF is:

$$f_L(\ell) = \begin{cases} 0.5e^{-0.5\ell} & \text{if } \ell \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Verification: $\displaystyle\int_0^\infty 0.5e^{-0.5\ell}\,\mathrm{d}\ell = \left[-e^{-0.5\ell}\right]_0^\infty = 0 - (-1) = 1$ ✓

Probability component lasts between 1 and 3 years:

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

Independent and Identically Distributed (i.i.d.)

Definition: i.i.d.

Random variables $X_1, X_2, \ldots, X_n$ are **i.i.d.** if:

1. **Independence:** For any sets $A_1, \ldots, A_n$:

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

2. **Identically Distributed:** All X_i have the same distribution

Example: i.i.d. Random Variables

Example: Rolling a 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
2. Each X_i follows the same discrete uniform distribution on $\{1, 2, 3, 4, 5, 6\}$

Non-example: Drawing Cards 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

Why i.i.d. matters:

- Simplifies analysis: joint probabilities become products
- Enables powerful limit theorems (LLN, CLT)

The Bernoulli Distribution

Definition: Bernoulli Distribution

$X \sim \text{Bernoulli}(p)$ if its PMF is:

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

- $\mathbb{P}(X = 1) = p$ (success)
- $\mathbb{P}(X = 0) = 1 - p$ (failure)

Example: 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 $X \sim \text{Bernoulli}(0.7)$.

Computing probabilities using the general formula:

$$\mathbb{P}(X = x) = p^x(1-p)^{1-x}$$

- $\mathbb{P}(X = 1) = 0.7^1 \cdot 0.3^0 = 0.7$
- $\mathbb{P}(X = 0) = 0.7^0 \cdot 0.3^1 = 0.3$

Quick check: $\mathbb{E}[X] = 0 \cdot 0.3 + 1 \cdot 0.7 = 0.7 = p$
The expectation of a Bernoulli random variable equals its success probability.

Example: Titanic Survival

Consider the Titanic dataset. Let X_i be a random variable representing the survival status of passenger i:

$$X_i = \begin{cases} 1 & \text{if passenger } i \text{ survived} \\ 0 & \text{otherwise} \end{cases}$$

We model this as $X_i \sim \text{Bernoulli}(p)$, where p is the survival probability.

Estimation approaches:

- **Naive model:** Estimate p as the overall survival rate in the dataset
- **Logistic regression:** Model p_i as a function of passenger features (class, sex, age)

Key insight: Each passenger's survival is a draw from a Bernoulli distribution with a potentially unique success parameter p_i.

The Binomial Distribution

Definition: Binomial Distribution

If $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p) and $Y = \sum_{i=1}^{n} X_i$, then $Y \sim \text{Binomial}(n, p)$:

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

Example: 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 $p = 0.5$:

$$Y \sim \text{Binomial}(5, 0.5)$$

Computing specific probabilities:

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

$$\mathbb{P}(Y = 0) = \binom{5}{0}(0.5)^5 = 1 \cdot 0.03125 = 0.03125$$

$$\mathbb{P}(Y = 5) = \binom{5}{5}(0.5)^5 = 1 \cdot 0.03125 = 0.03125$$

Example: Multiple Free Throws

A basketball player has an 80% free throw success rate. If she attempts $n = 10$ free throws, let Y be the total number of successful shots.
Then $Y \sim \text{Binomial}(10, 0.8)$.

Probability of making exactly 8 shots:

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

Probability of making at least 9 shots:

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

The Gaussian (Normal) Distribution

Definition: Gaussian Distribution

$X \sim \mathcal{N}(\mu, \sigma^2)$ if its PDF is:

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

- μ: mean (location parameter)
- σ^2: variance (spread parameter)
- $\mathcal{N}(0,1)$: **standard normal distribution**

Why is Gaussian so important?

$\downarrow$

Central Limit Theorem

Properties of the Gaussian Distribution

Theorem: Symmetry

$f_X(\mu + h) = f_X(\mu - h)$ for all h

Theorem: Unimodality

Unique global maximum at $x = \mu$

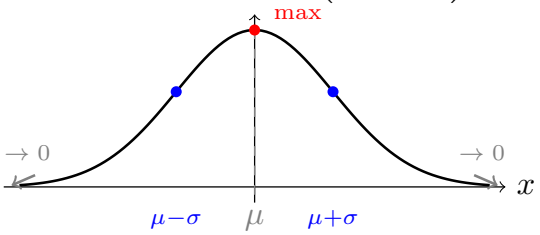
Theorem: Asymptotic Behavior

$$\lim_{x \to \pm\infty} f_X(x) = 0$$

Theorem: Inflection Points

Located at $x = \mu \pm \sigma$

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



Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$

Example: Standard Normal Distribution

Consider the standard normal distribution $Z \sim \mathcal{N}(0, 1)$ with $\mu = 0$ and $\sigma = 1$.

Applying the theorems:

- **Symmetry:** The PDF is symmetric about 0: $f_Z(-2) = f_Z(2)$
- **Maximum:** Occurs at $z = 0$:

$$f_Z(0) = \frac{1}{\sqrt{2\pi}} \approx 0.3989$$

- **Inflection points:** At $z = -1$ and $z = 1$:

$$f_Z(\pm 1) = \frac{1}{\sqrt{2\pi}} e^{-1/2} \approx 0.2420$$

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

Mathematical Expectation

Definition: Expectation – Measure-Theoretic

For random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$:

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

Computational Formulas:

- **Discrete:** $\mathbb{E}[X] = \sum_{x \in S} x \cdot p_X(x)$
- **Continuous:** $\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\, \mathrm{d}x$

Example: 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:

$$\mathbb{E}[X] = \sum_{k=1}^{6} k \cdot p_X(k) = \sum_{k=1}^{6} k \cdot \frac{1}{6}$$

$$= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3.5$$

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

Example: Expectation of a Bernoulli Random Variable

Let $X \sim \text{Bernoulli}(p)$, so $X = 1$ with probability p and $X = 0$ with probability $1 - p$.

Computing the expectation:

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

Elegant result: The expectation of an indicator random variable equals the probability of the event it indicates.

Example: If X represents whether a coin lands heads with $p = 0.7$:

$$\mathbb{E}[X] = 0.7$$

Example: 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:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = \int_a^b x \cdot \frac{1}{b-a}\,\mathrm{d}x$$

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

Intuition: The mean of a uniform distribution is the midpoint of its support.
Example: If $X \sim \text{Uniform}(0, 10)$, then $\mathbb{E}[X] = 5$.

Example: Expectation of an Exponential Distribution

Let $X \sim \text{Exponential}(\lambda)$ with PDF $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.

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

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

Example: If a light bulb's lifetime follows Exponential(0.1) (measured in years), its expected lifetime is $\mathbb{E}[X] = 10$ years.

Linearity of Expectation

Theorem: Linearity of Expectation

For any random variables X, Y (not necessarily independent) and constants $a, b \in \mathbb{R}$:

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

Proof (Continuous Case).

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

□

Example: 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.

Direct approach (tedious): Use the binomial distribution

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

Using linearity (elegant): Define indicator variables

$$X_i = \mathbf{1}_{\{\text{flip } i \text{ is heads}\}} \quad \Rightarrow \quad X = X_1 + X_2 + \cdots + X_n$$

By linearity:

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

Key insight: No need to sum over all $n+1$ binomial terms!

Example: Expected Sum of Two Dice Without Joint Distribution

Let $S = X_1 + X_2$ be the sum of two fair six-sided dice.

Direct approach: Enumerate all 36 outcomes

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

Using linearity:

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

Crucial observation: This works *regardless of whether X_1 and X_2 are independent*. Even if the dice were dependent, linearity still holds.
This universality—linearity requires no assumptions about independence—is what makes it such a powerful tool.

Example: 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:

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

Finding $\mathbb{E}[\bar{X}]$ directly would require knowing the distribution of $\bar{X}$, which can be complicated.

Using linearity:

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

Elegant result: The expected value of the sample mean equals the population mean—this holds for *any* distribution, derived without computing a single integral.

Variance

Definition: Variance

For random variable X with mean $\mu = \mathbb{E}[X]$, denote the variance of X as:

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

Standard deviation $\sigma_X = \sqrt{\mathrm{Var}(X)}$.

Theorem: Computational Formula

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

Proof.

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 - 2X\mu + \mu^2] \\ &= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - 2\mu^2 + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}$$

Example: 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 $1/6$.

Step 1: Compute $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$

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

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

Step 2: Apply the computational formula

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

Standard deviation: $\sigma_X = \sqrt{35/12} \approx 1.71$

Example: Variance of a Uniform Distribution

Let $X \sim \text{Uniform}(a, b)$ with PDF $f(x) = \frac{1}{b-a}$ for $x \in [a, b]$.

Computing the moments:

$$\mathbb{E}[X] = \frac{a+b}{2}, \qquad \mathbb{E}[X^2] = \frac{a^2+ab+b^2}{3}$$

Applying the computational formula:

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

Example: If $X \sim \text{Uniform}(0, 1)$, then $\text{Var}(X) = 1/12 \approx 0.083$ and $\sigma_X \approx 0.289$.

Example: 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:

$$\mathbb{E}[X] = \frac{1}{\lambda}, \qquad \mathbb{E}[X^2] = \frac{2}{\lambda^2}$$

Applying the computational formula:

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

Key property: $\sigma_X = 1/\lambda = \mathbb{E}[X]$
The standard deviation equals the mean—this is characteristic of the exponential distribution.

Properties of Variance

Theorem: Scaling Property

For constants $a, b \in \mathbb{R}$:

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

Proof.

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

□

Example: 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: $F = \frac{9}{5}C + 32$

Applying the scaling property:

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

Thus $\sigma_F = 9$ degrees Fahrenheit.

Key observations:

- The additive constant 32 does *not* contribute to the variance
- 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.

Covariance

Definition: Covariance

For random variables X, Y with means μ_X, μ_Y:

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

Interpretation:

- $\text{Cov}(X, Y) > 0$: X and Y tend to move together
- $\text{Cov}(X, Y) < 0$: X and Y tend to move oppositely
- $\text{Cov}(X, Y) = 0$: X and Y are **uncorrelated**

Important: Independence $\Rightarrow$ Uncorrelated, but not vice versa!

Note: $\text{Cov}(X, X) = \text{Var}(X)$

Covariance: The Geometric Connection

Fundamental insight: Covariance $\approx$ Dot Product

For data vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, define centered vectors:

$$\tilde{\mathbf{x}} = \mathbf{x} - \bar{x}\mathbf{1}, \quad \tilde{\mathbf{y}} = \mathbf{y} - \bar{y}\mathbf{1}$$

Sample covariance:

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

Sample covariance = scaled **dot product** of centered vectors

Statistical correlation $\equiv$ Geometric alignment

The Covariance Matrix

Definition: Covariance Matrix

For random vector $\mathbf{X} = [X_1, \ldots, X_p]^T$ with mean $\boldsymbol{\mu}$:

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

where $\Sigma_{ij} = \text{Cov}(X_i, X_j)$.

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

Key Properties:

1. **Symmetric:** $\Sigma = \Sigma^T$
2. **Positive semi-definite:** $\mathbf{a}^T \Sigma \mathbf{a} \geq 0$ for all $\mathbf{a}$
3. **Diagonal entries:** variances

Chebyshev's Inequality

Lemma: Chebyshev's Inequality

For random variable X with mean μ and variance σ^2:

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

Proof.

Define indicator $\mathbf{1}_{|X-\mu|\geq\epsilon}$. When $|X - \mu| \geq \epsilon$:

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

Taking expectations:

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

Dividing by ϵ^2 gives the result. □

Significance: Universal bound using only mean and variance – works for *any* distribution!

The Law of Large Numbers

Theorem: Weak Law of Large Numbers

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for any $\epsilon > 0$:

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

Proof.

First: $\mathbb{E}[\bar{X}_n] = \mu$ and $\text{Var}(\bar{X}_n) = \sigma^2/n$.
By Chebyshev: $\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0$ as $n \to \infty$. □

Interpretation: Sample mean converges to true mean as $n \to \infty$
Justification: Why sample statistics estimate population parameters

The Central Limit Theorem

Theorem: Central Limit Theorem

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance $\sigma^2 > 0$. Then, as $n \to \infty$,

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

Equivalently, for large n: $\bar{X}_n \overset{\text{approx}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

Profound implication:

Distribution of sample means approaches Gaussian **regardless** of the underlying distribution

Why Gaussian is everywhere: Many phenomena arise from aggregating small independent effects

CLT: Proof Sketch

Proof Outline using Characteristic Functions.

Step 1: Let $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$ (standardized, assuming $\mu = 0$, $\sigma = 1$)

Step 2: Characteristic function: $\varphi_{Z_n}(t) = \left[\varphi_X(t/\sqrt{n})\right]^n$

Step 3: Taylor expand: $\varphi_X(s) = 1 - \frac{s^2}{2} + O(s^3)$

Step 4: Substitute $s = t/\sqrt{n}$:

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

Step 5: Take limit using $(1 + a/n)^n \to e^a$:

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

This is the characteristic function of $\mathcal{N}(0,1)$. By Levy's theorem: $Z_n \xrightarrow{d} \mathcal{N}(0,1)$. □

Example: Averaging Dice Rolls

Consider rolling a fair six-sided die. Each roll X_i has:

- Mean: $\mu = 3.5$
- Variance: $\sigma^2 = 35/12 \approx 2.917$

The distribution is discrete uniform—far from Gaussian!

Rolling $n = 100$ dice: By CLT, the sample mean is approximately normal:

$$\bar{X}_{100} \overset{\text{approx}}{\sim} \mathcal{N}\left(3.5, \frac{2.917}{100}\right) = \mathcal{N}(3.5, 0.0292)$$

Standard deviation of the sample mean: $\sigma/\sqrt{n} \approx 0.171$

Constructing a 95% confidence interval:
For the standard normal, $\mathbb{P}(-1.96 < Z < 1.96) \approx 0.95$.

Substituting $Z_n = \dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ and rearranging:

$$\mathbb{P}\left(\mu - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \bar{X}_n < \mu + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

For $n = 100$ dice rolls:

$$\begin{aligned}\bar{X}_{100} &\in [3.5 - 1.96(0.171), 3.5 + 1.96(0.171)] \\ &\approx [3.16, 3.84]\end{aligned}$$

Interpretation: Approximately 95% of the time, the average of 100 dice rolls will fall in this interval.

The Estimation Problem

Setting:

- Observed data: $D = \{x_1, \ldots, x_n\}$
- Model: Parametric distribution $f(x|\theta)$
- Goal: Estimate θ from data

Examples:

- Coin flips $\rightarrow$ estimate bias p
- Heights $\rightarrow$ estimate mean μ, variance σ^2
- Linear regression $\rightarrow$ estimate coefficients β

Approach: Maximum Likelihood Estimation (MLE)

The Likelihood Function

Definition: Likelihood Function

Given data $D = \{x_1, \ldots, x_n\}$ from i.i.d. distribution $f(x|\theta)$:

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

Key distinction:

- $f(x|\theta)$: probability of data given fixed θ
- $L(\theta|D)$: same expression, but data is fixed, θ varies

Example: Bernoulli Likelihood

Consider flipping a coin 3 times and observing $D = \{1, 0, 1\}$ (Heads, Tails, Heads).
The likelihood function for the unknown bias p is:

$$L(p|D) = p \cdot (1-p) \cdot p = p^2(1-p)$$

Evaluating at different parameter values:

- $L(0.5|D) = (0.5)^2(0.5) = 0.125$
- $L(0.7|D) = (0.7)^2(0.3) = 0.147$
- $L(0.6|D) = (0.6)^2(0.4) = 0.144$
- $L(2/3|D) = (2/3)^2(1/3) = 4/27 \approx 0.148$

Observation: $p = 2/3$ gives the highest likelihood—this is the MLE (sample proportion = 2 heads / 3 flips).

Maximum Likelihood Estimation

Definition: MLE

The **maximum likelihood estimate** is:

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

Practical approach: Maximize **log-likelihood** instead

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

Why?

- Products $\rightarrow$ sums (easier calculus)
- Avoids numerical underflow
- log is monotonic, so same maximizer

Method: Set $\frac{\partial \ell}{\partial \theta} = 0$ and solve for θ

Example: MLE for Bernoulli(p)

Data: n trials with n_1 successes, $n_0 = n - n_1$ failures

Step 1: Likelihood

$$L(p|D) = p^{n_1}(1-p)^{n_0}$$

Step 2: Log-likelihood

$$\ell(p|D) = n_1 \log p + n_0 \log(1-p)$$

Step 3: Differentiate

$$\frac{d\ell}{dp} = \frac{n_1}{p} - \frac{n_0}{1-p} = 0$$

Step 4: Solve

$$n_1(1-\hat{p}) = n_0\hat{p} \quad \Rightarrow \quad \boxed{\hat{p}_{\text{MLE}} = \frac{n_1}{n}}$$

Result: MLE = sample proportion (intuitive!)

Connecting OLS and MLE

Linear model: $y_i = \mathbf{x}_i^T \beta + \epsilon_i \quad$ with $\quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$

This implies: $y_i | \mathbf{x}_i \sim \mathcal{N}(\mathbf{x}_i^T \beta, \sigma^2)$

Log-likelihood:

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

To maximize w.r.t. β**:**

$$\hat{\beta}_{\text{MLE}} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 = \arg\min_{\beta} \|y - X\beta\|^2$$

OLS = MLE Under Gaussian Errors

Key Result:

Under assumption $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$:

$$\hat{\beta}_{\text{MLE}} = \hat{\beta}_{\text{OLS}} = (X^T X)^{-1} X^T y$$

Significance:

- **Geometric view** (Lesson 2): OLS = orthogonal projection
- **Probabilistic view** (Lesson 3): OLS = MLE
- Same answer, different perspectives!

Implication: Statistical inference (confidence intervals, hypothesis tests) now possible via probability theory

Conditional Probability

Definition: Conditional Probability

For events A, B with $\mathbb{P}(B) > 0$:

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

Intuition: Restrict sample space to B, rescale probabilities

Product Rule: $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$

Example: Rolling a Die

Suppose we roll a fair six-sided die. Let A be the event "the outcome is even" and B be the event "the outcome is greater than 3."

- $A = \{2, 4, 6\}$, so $\mathbb{P}(A) = 3/6 = 1/2$
- $B = \{4, 5, 6\}$, so $\mathbb{P}(B) = 3/6 = 1/2$
- $A \cap B = \{4, 6\}$, so $\mathbb{P}(A \cap B) = 2/6 = 1/3$

Computing the conditional probability:

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

Intuition: Among the outcomes greater than 3 (namely 4, 5, 6), two of them (4 and 6) are even. So the conditional probability is 2/3.

Example: Drawing Two Aces in a Row

What is the probability of drawing two Aces in a row from a standard 52-card deck (without replacement)?
Let $A_1 =$ "first card is an Ace" and $A_2 =$ "second card is an Ace."

Using the product rule:

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

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

Key insight: Conditioning on A_1 changes the probability of A_2—this is why drawing without replacement creates dependence.

Example: Medical Testing

A medical test for a disease has the following characteristics:

- Disease prevalence: $\mathbb{P}(D) = 0.01$ (1% of population)
- Sensitivity: $\mathbb{P}(T^+|D) = 0.95$ (95% true positive rate)
- Specificity: $\mathbb{P}(T^-|\neg D) = 0.90$ (90% true negative rate)

Therefore, $\mathbb{P}(T^+|\neg D) = 0.10$ (10% false positive rate).

Computing joint probability using the product rule:

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

The key question: If a patient tests positive, what is the probability they actually have the disease?
This requires **inverting** the conditional probability $\rightarrow$ Bayes' Theorem

Bayes' Theorem

Theorem: Bayes' Theorem

For events A, B with $\mathbb{P}(A), \mathbb{P}(B) > 0$:

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

Proof.

From the product rule:

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

Equating and dividing by $\mathbb{P}(B)$:

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

Example: Medical Test Revisited

Given: $\mathbb{P}(D) = 0.01$, $\mathbb{P}(T^+|D) = 0.95$, $\mathbb{P}(T^+|\neg D) = 0.10$
Question: If a patient tests positive, what is $\mathbb{P}(D|T^+)$?

Step 1: Compute total probability of positive test using law of total probability

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

Breakdown:

- True positives: 0.0095 (diseased people who test positive)
- False positives: 0.099 (healthy people who test positive)

Step 2: Apply Bayes' Theorem

$$\mathbb{P}(D|T^+) = \frac{\mathbb{P}(T^+|D)\mathbb{P}(D)}{\mathbb{P}(T^+)} = \frac{0.95 \times 0.01}{0.1085} = \frac{0.0095}{0.1085} \approx \boxed{0.088}$$

Surprising result: Even with a positive test, there is only about an 8.8% chance the patient has the disease!

Why is this counterintuitive?

- The disease is rare (1% prevalence)
- False positives from healthy population (10% of 99%) dominate true positives
- Most positive tests come from the large healthy population, not the small diseased population

Lesson: Base rates (priors) matter enormously in probabilistic reasoning.

Bayesian Inference

Recast Bayes for parameters and data:

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

- **Prior** $\mathbb{P}(\theta)$: Belief about θ *before* data
- **Likelihood** $\mathbb{P}(D|\theta)$: Same as MLE's $L(\theta|D)$
- **Posterior** $\mathbb{P}(\theta|D)$: Updated belief *after* data
- **Evidence** $\mathbb{P}(D)$: Normalizing constant

$$\boxed{\text{Prior Belief} + \text{Data} \longrightarrow \text{Updated Posterior Belief}}$$

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

Example: 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**:

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

Why Beta?

- Support is $[0,1]$—appropriate for a probability
- Flexible: $\alpha = \beta = 1$ gives uniform (no preference)
- **Conjugate prior:** Beta prior + Binomial likelihood = Beta posterior

Data: We flip the coin n times and observe h heads and $t = n - h$ tails.
Likelihood: $\mathbb{P}(D|\theta) = \binom{n}{h}\theta^h(1-\theta)^t \propto \theta^h(1-\theta)^t$

Applying Bayes' Theorem:

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

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

Numerical example: Start with uniform prior $\text{Beta}(1,1)$. Observe 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.

The Naive Bayes Classifier

Goal: Classify observation $\mathbf{x} = (x_1, \ldots, x_p)$ into class C_k

Bayesian approach

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

Decision rule

$$\hat{C} = \arg\max_k \mathbb{P}(\mathbf{x}|C_k)\mathbb{P}(C_k)$$

Definition: Naive Bayes Assumption

Features are **conditionally independent** given class:

$$\mathbb{P}(\mathbf{x}|C_k) = \prod_{j=1}^{p} \mathbb{P}(x_j|C_k)$$

Final classifier:

$$\hat{C} = \arg\max_k \left[\log \mathbb{P}(C_k) + \sum_{j=1}^{p} \log \mathbb{P}(x_j|C_k) \right]$$

Example: Email Spam Detection

Consider a spam filter with two classes: $C_1 = \text{Spam}$ and $C_2 = \text{Not Spam}$.
Each email is represented by binary features indicating word presence:

- $x_1 = 1$ if "free" appears, 0 otherwise
- $x_2 = 1$ if "meeting" appears, 0 otherwise
- $x_3 = 1$ if "winner" appears, 0 otherwise

Training data estimates:

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

New email: Contains "free" and "winner" but not "meeting"
So $\mathbf{x} = (x_1 = 1, x_2 = 0, x_3 = 1)$

Computing unnormalized posteriors:

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

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

Normalized posterior:

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

Decision: Classify as **Spam** with 99.6% confidence.

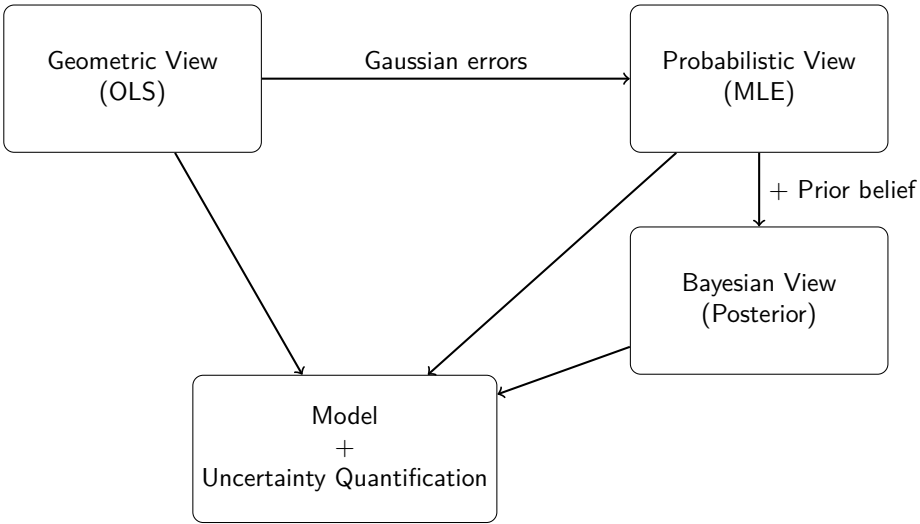
Summary: Three Pillars

1. **Distributions as Models**
   - Data = realizations of random variables
   - Bernoulli/Binomial for classification
   - Gaussian for continuous variables (CLT justification)

2. **Parameter Estimation via MLE**
   - Likelihood function: $L(\theta|D) = \prod_i f(x_i|\theta)$
   - MLE: maximize likelihood (or log-likelihood)
   - **OLS = MLE** under Gaussian errors

3. **Moments and Relationships**
   - Expectation (center), Variance (spread)
   - Covariance $\approx$ Dot product of centered vectors
   - LLN and CLT: foundations of statistical inference

The Big Picture



Questions?

"Probability theory is nothing but common sense reduced to calculation."

— Pierre-Simon Laplace