Mathematics for Data Science

A Comprehensive Study Resource

Henry Baker

2023–2024

Hertie School

This document provides a self-contained treatment of the mathematical foundations essential for data science: probability theory, calculus, linear algebra, and optimisation.

Contents

Part I

Probability Theory

Chapter 1

Probability Theory

Learning Objectives

By the end of this chapter, you should be able to:

- Define sample spaces and events using set notation, and manipulate them using set operations
- State and apply the Kolmogorov axioms of probability
- Derive key properties of probability functions from the axioms
- Apply counting principles (permutations, combinations) to compute probabilities
- Use the inclusion-exclusion principle for unions of events
- Solve problems involving sampling with and without replacement

Prerequisites

This chapter assumes familiarity with:

- Basic set theory: unions ($\cup$), intersections ($\cap$), complements (A^c), and set containment ($\subseteq$)
- Summation notation and basic algebraic manipulation
- The factorial function: $n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1$

1.1 Sample Spaces and Events

Probability theory provides a mathematical framework for reasoning about uncertainty. Before we can assign probabilities, we must precisely describe what outcomes are possible and what collections of outcomes we care about.

Definition 1.1 (Sample Space)**.** A **sample space**, denoted S or Ω, is the set of all possible outcomes of a random experiment. Each element $\omega \in S$ is called an **outcome** or **sample point**.

Definition 1.2 (Event)**.** An **event** is any subset $A \subseteq S$ of the sample space. We say that event A **occurs** if the outcome of the experiment is an element of A.

Example 1.1 (Coin Flips)**.** Consider flipping a fair coin 10 times. The sample space is:

$$S = \{0,1\}^{10} = \{(s_1, s_2, \ldots, s_{10}) : s_j \in \{0,1\} \text{ for all } j\}$$

where we encode Tails as 0 and Heads as 1. This sample space has $|S| = 2^{10} = 1024$ outcomes.

Let A_n denote the event that the nth flip is Heads. Then:

- $A_1 = \{(1, s_2, \ldots, s_{10}) : s_j \in \{0,1\} \text{ for } 2 \leq j \leq 10\}$ is the event "first flip is Heads"
- $B = \bigcup_{n=1}^{10} A_n$ is the event "at least one flip is Heads"
- $C = \bigcap_{n=1}^{10} A_n$ is the event "all flips are Heads"
- $D = \bigcup_{n=1}^{9} (A_n \cap A_{n+1})$ is the event "at least two consecutive Heads"

Events as Questions

An event corresponds to a yes/no question about the outcome. The event A is the set of all outcomes for which the answer is "yes." For instance, "Did at least one coin land Heads?" corresponds to the event B above.

1.1.1 Set Operations on Events

Since events are sets, we use set operations to combine them:

Set Operations for Events

Let A and B be events in sample space S.

$$\begin{aligned} A \cup B &\quad (\text{"}A \text{ or } B \text{ occurs"}) \\ A \cap B &\quad (\text{"}A \text{ and } B \text{ both occur"}) \\ A^c = S \setminus A &\quad (\text{"}A \text{ does not occur"}) \\ A \setminus B = A \cap B^c &\quad (\text{"}A \text{ occurs but } B \text{ does not"}) \end{aligned}$$

Two events are **mutually exclusive** (or **disjoint**) if $A \cap B = \emptyset$.

Theorem 1.1 (De Morgan's Laws)**.** *For any events A and B:*

$$(A \cup B)^c = A^c \cap B^c \tag{1.1}$$

$$(A \cap B)^c = A^c \cup B^c \tag{1.2}$$

More generally, for any collection of events $\{A_i\}_{i \in I}$:

$$\left(\bigcup_{i \in I} A_i \right)^c = \bigcap_{i \in I} A_i^c \tag{1.3}$$

$$\left(\bigcap_{i \in I} A_i \right)^c = \bigcup_{i \in I} A_i^c \tag{1.4}$$

Proof. We prove Equation (1.1). An outcome $\omega \in (A \cup B)^c$ if and only if $\omega \notin A \cup B$, which holds if and only if $\omega \notin A$ and $\omega \notin B$. This is equivalent to $\omega \in A^c$ and $\omega \in B^c$, i.e., $\omega \in A^c \cap B^c$. The proof of Equation (1.2) is similar. ∎

De Morgan's Laws

De Morgan's laws formalise a natural duality: "not (A or B)" is the same as "(not A) and (not B)," while "not (A and B)" is the same as "(not A) or (not B)." These laws are essential for computing probabilities of complements.

1.2 The Axioms of Probability

We now define what it means to assign probabilities to events. The modern axiomatic foundation, due to Kolmogorov (1933), requires only three axioms from which all of probability theory follows.

Definition 1.3 (Probability Space)**.** A **probability space** is a triple $(S, \mathcal{F}, \mathbb{P})$ where:

1. S is a sample space
2. $\mathcal{F}$ is a collection of subsets of S (the events we can assign probabilities to)
3. $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a **probability function** satisfying the Kolmogorov axioms

Kolmogorov Axioms

A probability function $\mathbb{P}$ must satisfy:

1. **Non-negativity**: $\mathbb{P}(A) \geq 0$ for all events A
2. **Normalisation**: $\mathbb{P}(S) = 1$
3. **Countable additivity**: If $A_1, A_2, \ldots$ are pairwise disjoint events (i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$), then

$$\mathbb{P}\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mathbb{P}(A_j)$$

Remark. The third axiom (countable additivity) is stronger than finite additivity. It ensures that probability behaves well with limits, which is essential for working with continuous distributions and infinite sequences of events.

1.2.1 Properties Derived from the Axioms

All properties of probability follow logically from the three Kolmogorov axioms. We now derive the most important ones.

Theorem 1.2 (Probability of the Empty Set)**.** $\mathbb{P}(\emptyset) = 0$.

Proof. Consider the sequence $A_1 = S$, $A_2 = A_3 = \ldots = \emptyset$. These sets are pairwise disjoint, and $\bigcup_{j=1}^{\infty} A_j = S$. By countable additivity:

$$\mathbb{P}(S) = \mathbb{P}(S) + \sum_{j=2}^{\infty} \mathbb{P}(\emptyset)$$

Since $\mathbb{P}(S) = 1$ is finite, we must have $\sum_{j=2}^{\infty} \mathbb{P}(\emptyset) = 0$. Since $\mathbb{P}(\emptyset) \geq 0$ by non-negativity, this forces $\mathbb{P}(\emptyset) = 0$. ∎

Theorem 1.3 (Complement Rule)**.** *For any event A:*

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

Proof. Since A and A^c are disjoint and $A \cup A^c = S$, by countable additivity (applied to just two sets):

$$1 = \mathbb{P}(S) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$$

Rearranging gives $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$. ■

Theorem 1.4 (Monotonicity)**.** *If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.*

Proof. Write $B = A \cup (B \cap A^c)$. Since A and $B \cap A^c$ are disjoint:

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c) \geq \mathbb{P}(A)$$

where the inequality follows from non-negativity: $\mathbb{P}(B \cap A^c) \geq 0$. ■

Theorem 1.5 (Inclusion-Exclusion for Two Events)**.** *For any events A and B:*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Proof. We can write $A \cup B$ as a disjoint union:

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

By additivity: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$.

Similarly, $B = (A \cap B) \cup (B \cap A^c)$ is a disjoint union, so:

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

Thus $\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Substituting: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. ■

Summary of Basic Properties

For any events A, B in a probability space:

$$\mathbb{P}(\emptyset) = 0$$
$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$
$$A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$$
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

1.3 The Naive Definition of Probability

When all outcomes in a finite sample space are equally likely, probability calculations reduce to counting.

Definition 1.4 (Naive Probability)**.** If S is a finite sample space with equally likely outcomes, the **naive probability** of an event A is:

$$\mathbb{P}_{\text{naive}}(A) = \frac{|A|}{|S|} = \frac{\text{number of outcomes favourable to } A}{\text{total number of outcomes in } S}$$

When Naive Probability Fails

The naive definition only applies when:

1. The sample space is finite
2. All outcomes are equally likely

Many real-world situations violate these assumptions. For instance, if a coin is biased, the outcomes $\{H, T\}$ are not equally likely, and we cannot use $\mathbb{P}(\text{Heads}) = 1/2$.

Example 1.2 (Leibniz's Mistake)**.** Gottfried Wilhelm Leibniz, the co-inventor of calculus, incorrectly reasoned that when rolling two dice, the probability of getting a sum of 11 equals the probability of getting a sum of 12, because each can be achieved "in one way."

The error lies in conflating outcomes with events. With distinguishable dice (say, red and blue), the sample space has 36 equally likely outcomes: $\{(r, b) : r, b \in \{1, 2, 3, 4, 5, 6\}\}$.

- Sum of 11: $(5, 6)$ or $(6, 5)$ — two outcomes
- Sum of 12: $(6, 6)$ — one outcome

Thus $\mathbb{P}(\text{sum} = 11) = 2/36$ while $\mathbb{P}(\text{sum} = 12) = 1/36$.

Lesson: When order matters or items are distinguishable, label them explicitly to avoid undercounting.

1.4 Counting Principles

To apply the naive definition of probability, we need systematic methods for counting outcomes. This section develops the fundamental counting principles.

1.4.1 The Multiplication Rule

Theorem 1.6 (Multiplication Rule)**.** *If an experiment consists of k stages, where stage i has n_i possible outcomes (regardless of what happened in previous stages), then the total number of outcomes is:*

$$n_1 \times n_2 \times \cdots \times n_k$$

Tree Diagram Interpretation

The multiplication rule corresponds to counting paths through a tree diagram. At the root, we branch into n_1 choices for stage 1. From each of these, we branch into n_2 choices for stage 2, and so on. The total number of paths from root to leaves is the product $n_1 \cdot n_2 \cdots n_k$.

Remark. The multiplication rule holds even when stages are performed in different orders, because multiplication is commutative. This can be counterintuitive: the order in which we *think about* the stages need not match the temporal order in which they occur.

1.4.2 Permutations and Combinations

The distinction between permutations and combinations depends on whether the order of selection matters.

Definition 1.5 (Permutation)**.** A **permutation** is an ordered arrangement of objects. The number of ways to arrange k objects chosen from n distinct objects is:

$$P(n,k) = n \cdot (n-1) \cdot (n-2) \cdots (n-k+1) = \frac{n!}{(n-k)!}$$

For $k = n$, this gives $n!$, the number of ways to arrange all n objects.

Definition 1.6 (Combination)**.** A **combination** is an unordered selection of objects. The number of ways to choose k objects from n distinct objects (without regard to order) is the **binomial coefficient**:

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

This is read "n choose k." By convention, $\binom{n}{k} = 0$ if $k > n$ or $k < 0$.

Relationship Between Permutations and Combinations

To see why $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, consider a two-stage process:

1. Choose which k objects to select: $\binom{n}{k}$ ways
2. Arrange those k objects: $k!$ ways

By the multiplication rule, the total number of ordered arrangements is $\binom{n}{k} \cdot k!$. But this also equals $P(n,k) = \frac{n!}{(n-k)!}$. Solving:

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

Example 1.3 (Committee Selection)**.** From a group of 10 people, how many ways can we:

1. Form a committee of 3? Answer: $\binom{10}{3} = \frac{10!}{3!\cdot 7!} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120$
2. Elect a president, vice-president, and treasurer? Answer: $P(10,3) = 10 \cdot 9 \cdot 8 = 720$

The first counts unordered selections (order doesn't matter for a committee), while the second counts ordered selections (distinct roles).

1.4.3 Sampling With and Without Replacement

When sampling from a population, we must specify whether selected items are returned to the population before the next selection.

Counting Formulas for Sampling

When selecting k items from n distinct items:

	Order Matters	Order Doesn't Matter
With Replacement	n^k	$\binom{n+k-1}{k}$
Without Replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Example 1.4 (Four Scenarios)**.** Suppose we have 5 distinct balls (labelled 1–5) and select 3.

With replacement, order matters: Each selection has 5 choices, so $5^3 = 125$ outcomes. Example outcomes: $(1,1,1)$, $(1,2,3)$, $(3,2,1)$.

With replacement, order doesn't matter: This counts multisets. Using stars and bars: $\binom{5+3-1}{3} = \binom{7}{3} = 35$. Example outcomes: $\{1,1,1\}$, $\{1,2,3\}$.

Without replacement, order matters: This is $P(5,3) = 5 \cdot 4 \cdot 3 = 60$. Example outcomes: $(1,2,3)$, $(3,2,1)$ — these are distinct.

Without replacement, order doesn't matter: This is $\binom{5}{3} = 10$. Example outcomes: $\{1,2,3\}$, $\{1,4,5\}$ — here $(1,2,3)$ and $(3,2,1)$ represent the same outcome.

1.5 The Inclusion-Exclusion Principle

When computing the probability of a union of events, we must account for overlaps. The inclusion-exclusion principle generalises Theorem 1.5 to any finite number of events.

Theorem 1.7 (Inclusion-Exclusion Principle)**.** *For any events $A_1, A_2, \ldots, A_n$:*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n)$$

Equivalently:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k})$$

Proof. We prove this by induction on n.

Base case ($n = 2$): This is Theorem 1.5.

Inductive step: Assume the formula holds for $n-1$ events. Let $B = \bigcup_{i=1}^{n-1} A_i$. Then:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \mathbb{P}(B \cup A_n) = \mathbb{P}(B) + \mathbb{P}(A_n) - \mathbb{P}(B \cap A_n)$$

By the inductive hypothesis, $\mathbb{P}(B)$ expands correctly. Also:

$$B \cap A_n = \left(\bigcup_{i=1}^{n-1} A_i\right) \cap A_n = \bigcup_{i=1}^{n-1} (A_i \cap A_n)$$

Applying the inductive hypothesis to this union of $n-1$ events and combining terms yields the result. ■

Example 1.5 (Three Events)**.** For three events A, B, C:

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)$$

Why Inclusion-Exclusion Works

When we sum $\mathbb{P}(A_i)$, we overcount outcomes in multiple events. Subtracting pairwise intersections corrects for double-counting, but then we've subtracted too much for outcomes in three or more events. We add back triple intersections, subtract quadruple intersections, and so on. The alternating signs ensure each outcome is counted exactly once.

1.6 The Birthday Problem

The birthday problem is a classic application of counting principles that illustrates how quickly collision probabilities grow.

Example 1.6 (Birthday Problem)**.** In a room of k people, what is the probability that at least two share a birthday?

Assumptions: We assume 365 equally likely birthdays (ignoring leap years) and that birthdays are independent.

Strategy: It's easier to compute the complement — the probability that all k birthdays are distinct — and subtract from 1.

Solution: The sample space is all possible birthday assignments: $|S| = 365^k$.

For all birthdays to be distinct, we sample without replacement:

$$|\text{no matches}| = 365 \times 364 \times 363 \times \cdots \times (365 - k + 1) = \frac{365!}{(365-k)!}$$

Therefore:

$$\begin{aligned}
\mathbb{P}(\text{no match}) &= \frac{365 \times 364 \times \cdots \times (365-k+1)}{365^k} \\
&= \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365-k+1}{365} \\
&= \prod_{i=0}^{k-1} \frac{365-i}{365} = \prod_{i=0}^{k-1} \left(1 - \frac{i}{365}\right)
\end{aligned}$$

And finally:

$$\mathbb{P}(\text{at least one match}) = 1 - \prod_{i=0}^{k-1} \left(1 - \frac{i}{365}\right)$$

Birthday Problem Thresholds

- With $k = 23$ people, $\mathbb{P}(\text{match}) \approx 0.507$ (more likely than not!)
- With $k = 50$ people, $\mathbb{P}(\text{match}) \approx 0.970$
- With $k = 70$ people, $\mathbb{P}(\text{match}) \approx 0.999$

Why 23 is Surprisingly Small

The key insight is that we're not asking whether someone shares *your* birthday (which would require about 253 people for 50% probability), but whether *any pair* shares a birthday. With k people, there are $\binom{k}{2} = \frac{k(k-1)}{2}$ pairs to check. For $k = 23$, this gives 253 pairs — many opportunities for a match.

Assumptions Matter

The birthday problem assumes uniform distribution of birthdays. In reality, birthdays are not uniformly distributed (more births occur in certain months), which slightly increases the collision probability. The independence assumption also fails for twins and siblings in the same room.

1.7 Summary

This chapter established the foundations of probability theory:

Chapter Summary

Foundations:

- A probability space consists of a sample space S, a collection of events, and a probability function $\mathbb{P}$
- The Kolmogorov axioms (non-negativity, normalisation, countable additivity) define valid probability functions
- All probability properties derive from these three axioms

Key techniques:

- For finite, equally likely sample spaces: $\mathbb{P}(A) = |A|/|S|$
- Counting formulas depend on whether order matters and whether sampling is with replacement
- The complement rule: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ — often easier than direct calculation
- Inclusion-exclusion handles unions of overlapping events

Common pitfalls:

- Applying naive probability when outcomes aren't equally likely
- Confusing permutations (order matters) with combinations (order doesn't matter)
- Undercounting by treating distinguishable objects as indistinguishable

Chapter 2

Conditional Probability and Random Variables

Learning Objectives

By the end of this chapter, you should be able to:

- Define and compute conditional probabilities, understanding their geometric interpretation
- Apply Bayes' Rule to update probabilities given new evidence
- Use the Law of Total Probability to decompose complex probability calculations
- Distinguish between independence and conditional independence, and identify when each holds
- Define random variables and construct their probability mass functions (PMFs)
- Work with cumulative distribution functions (CDFs) and understand their relationship to PMFs
- Apply common discrete distributions (Bernoulli, Binomial, Discrete Uniform) to model real-world phenomena

Prerequisites

This chapter builds directly on Chapter 1 and assumes familiarity with:

- Sample spaces, events, and the Kolmogorov axioms
- Set operations: union ($\cup$), intersection ($\cap$), complement (A^c)
- Basic counting: permutations, combinations, and the binomial coefficient $\binom{n}{k}$
- The naive definition of probability for finite, equally likely sample spaces

2.1 Conditional Probability

In many situations, we have partial information about an experiment's outcome. We know that some event B has occurred, and we want to update our probability assessments accordingly. This leads to the fundamental concept of *conditional probability*.

Definition 2.1 (Conditional Probability)**.** Let A and B be events with $\mathbb{P}(B) > 0$. The **conditional probability of A given B** is:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{2.1}$$

This represents the probability that A occurs, given that we know B has occurred.

Conditional Probability Formula

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad \text{where } \mathbb{P}(B) > 0$$

Equivalently, by symmetry of intersection:

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \quad \text{where } \mathbb{P}(A) > 0$$

2.1.1 Geometric Interpretation

The definition of conditional probability has an elegant geometric interpretation. Consider a Venn diagram where areas represent probabilities.

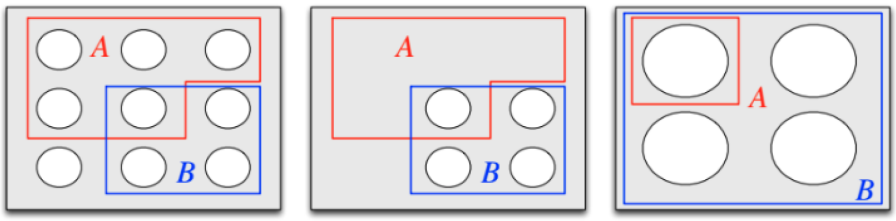


Figure 2.1: Geometric interpretation of conditional probability. When computing $\mathbb{P}(A \mid B)$, we "zoom in" on event B: we restrict our attention to outcomes where B occurred, then ask what fraction of these also lie in A. This requires re-normalising by dividing by $\mathbb{P}(B)$.

Zooming In and Re-normalising

When we condition on B, we are effectively:

1. **Restricting the sample space**: We discard all outcomes in B^c (the complement of B), since we know these didn't occur.
2. **Re-normalising**: The probabilities of the remaining outcomes no longer sum to 1 (they sum to $\mathbb{P}(B)$), so we divide by $\mathbb{P}(B)$ to restore a valid probability distribution.

In this new, restricted probability space, $\mathbb{P}(A \mid B)$ is simply the proportion of B that overlaps with A.

2.1.2 The Multiplication Rule

Rearranging the definition of conditional probability gives us a useful formula for joint probabilities:

Theorem 2.1 (Multiplication Rule for Probability). *For any events A and B with $\mathbb{P}(B) > 0$:*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B) \cdot \mathbb{P}(B) \tag{2.2}$$

By symmetry:

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

Proof. This follows immediately from Definition 2.1 by multiplying both sides by $\mathbb{P}(B)$. ■

The multiplication rule extends naturally to chains of events:

Corollary 2.2 (Chain Rule for Probability). *For events $A_1, A_2, \ldots, A_n$ with $\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) > 0$:*

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2 \mid A_1) \cdot \mathbb{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbb{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}) \tag{2.3}$$

2.1.3 Key Properties and Common Misconceptions

Conditional Probability is Not Symmetric

In general, $\mathbb{P}(A \mid B) \neq \mathbb{P}(B \mid A)$. These represent fundamentally different questions:

- $\mathbb{P}(A \mid B)$: "Given B occurred, how likely is A?"
- $\mathbb{P}(B \mid A)$: "Given A occurred, how likely is B?"

For example, $\mathbb{P}(\text{wet grass} \mid \text{rain})$ is high, but $\mathbb{P}(\text{rain} \mid \text{wet grass})$ might be moderate (sprinklers also wet grass).

Chronology Does Not Determine Conditioning Direction

A common misconception is that we can only condition on events that occurred *before* the event we're computing the probability of. This is false. Conditional probability is about *information*, not *causation* or *temporal order*.
Consider: "What is the probability that it rained yesterday, given that the grass is wet today?" This is a perfectly valid conditional probability $\mathbb{P}(\text{rain yesterday} \mid \text{wet grass today})$, even though we're conditioning on a later event.

Example 2.1 (Playing Cards). Two cards are drawn from a standard 52-card deck without replacement. Let:

- A = first card is a heart
- B = second card is red (hearts or diamonds)

What is $\mathbb{P}(B \mid A)$?

Solution: Using the conditional probability formula:

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

We compute each term:

$$\begin{aligned}\mathbb{P}(A) &= \frac{13}{52} = \frac{1}{4} \\ \mathbb{P}(A \cap B) &= \mathbb{P}(\text{1st is heart}) \times \mathbb{P}(\text{2nd is red} \mid \text{1st is heart}) \\ &= \frac{13}{52} \times \frac{25}{51} = \frac{13 \times 25}{52 \times 51} = \frac{325}{2652} = \frac{25}{204}\end{aligned}$$

Here, if the first card is a heart, there remain 51 cards, of which 25 are red (12 remaining hearts plus 13 diamonds).

Therefore:

$$\mathbb{P}(B \mid A) = \frac{25/204}{1/4} = \frac{25}{204} \times \frac{4}{1} = \frac{25}{51}$$

Verification: This makes sense! Given that the first card was a heart, there are 51 cards remaining, of which 25 are red.

2.2 Bayes' Rule

Bayes' Rule is one of the most important results in probability theory, providing a systematic way to update beliefs in light of new evidence.

2.2.1 Derivation and Statement

Theorem 2.3 (Bayes' Rule)**.** *For events A and B with $\mathbb{P}(A), \mathbb{P}(B) > 0$:*

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)} \tag{2.4}$$

Proof. From the definition of conditional probability:

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

By the multiplication rule (Theorem 2.1):

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

Substituting:

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

■

Bayes' Rule

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

In words: the probability of A given B equals the probability of B given A, times the prior probability of A, divided by the probability of B.

2.2.2 Interpretation: Updating Beliefs

Bayes' Rule has a natural interpretation in terms of *updating beliefs*:

The Bayesian Update

Let A be a hypothesis and B be observed evidence. Then:

- $\mathbb{P}(A)$ is the **prior probability**: our belief in A before seeing the evidence
- $\mathbb{P}(B \mid A)$ is the **likelihood**: how probable the evidence is if A is true
- $\mathbb{P}(B)$ is the **marginal likelihood** (or **evidence**): the overall probability of observing B
- $\mathbb{P}(A \mid B)$ is the **posterior probability**: our updated belief in A after seeing the evidence

Bayes' Rule tells us precisely how to update from prior to posterior:

$$\underbrace{\mathbb{P}(A \mid B)}_{\text{posterior}} = \frac{\overbrace{\mathbb{P}(B \mid A)}^{\text{likelihood}} \times \overbrace{\mathbb{P}(A)}^{\text{prior}}}{\underbrace{\mathbb{P}(B)}_{\text{evidence}}}$$

Prior vs Posterior Terminology

The terms "prior" and "posterior" refer specifically to beliefs *before* and *after* conditioning on evidence:

- **Prior**: $\mathbb{P}(A)$ — probability of A before considering B
- **Posterior**: $\mathbb{P}(A \mid B)$ — probability of A after learning B

Do not confuse these with the events A and B themselves. The terms describe our *state of knowledge*, not the events.

2.3 The Law of Total Probability

The Law of Total Probability (LOTP) allows us to compute the probability of an event by breaking it down into simpler cases.

Definition 2.2 (Partition)**.** A collection of events $\{A_1, A_2, \ldots, A_n\}$ forms a **partition** of the sample space S if:

1. The events are **mutually exclusive**: $A_i \cap A_j = \emptyset$ for $i \neq j$
2. The events are **collectively exhaustive**: $A_1 \cup A_2 \cup \cdots \cup A_n = S$

In other words, exactly one of the A_i occurs in any experiment.

Theorem 2.4 (Law of Total Probability)**.** *Let $\{A_1, A_2, \ldots, A_n\}$ be a partition of S with $\mathbb{P}(A_i) > 0$ for all i. Then for any event B:*

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i) \tag{2.5}$$

Proof. Since $\{A_i\}$ partitions S, we can write:

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

The events $B \cap A_i$ are pairwise disjoint (since the A_i are). By countable additivity:

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

Applying the multiplication rule to each term:

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i)$$

■

Law of Total Probability

If $\{A_1, \ldots, A_n\}$ partitions the sample space:

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i)$$

For the common case of two complementary events A and A^c:

$$\mathbb{P}(B) = \mathbb{P}(B \mid A) \cdot \mathbb{P}(A) + \mathbb{P}(B \mid A^c) \cdot \mathbb{P}(A^c)$$

Weighted Average Interpretation

LOTP says that $\mathbb{P}(B)$ is a weighted average of the conditional probabilities $\mathbb{P}(B \mid A_i)$, where the weights are $\mathbb{P}(A_i)$. Intuitively, we consider all the different "scenarios" (the A_i), compute how likely B is in each scenario, and average according to how likely each scenario is.

2.3.1 Tree Diagrams

Tree diagrams provide a visual method for organising LOTP calculations:

Example 2.2 (Tree Diagram Method)**.** Suppose a factory has two machines: Machine 1 produces 60% of items, Machine 2 produces 40%. Machine 1 has a 2% defect rate; Machine 2 has a 5% defect rate. What is the probability that a randomly selected item is defective?

Solution using LOTP:

Let D = "item is defective", M_1 = "produced by Machine 1", M_2 = "produced by Machine 2".

The partition is $\{M_1, M_2\}$ with:

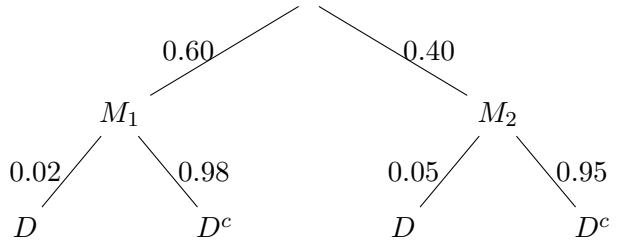
$$\begin{aligned} \mathbb{P}(M_1) = 0.60, \quad \mathbb{P}(M_2) = 0.40 \\ \mathbb{P}(D \mid M_1) = 0.02, \quad \mathbb{P}(D \mid M_2) = 0.05 \end{aligned}$$

By LOTP:

$$\begin{aligned} \mathbb{P}(D) &= \mathbb{P}(D \mid M_1) \cdot \mathbb{P}(M_1) + \mathbb{P}(D \mid M_2) \cdot \mathbb{P}(M_2) \\ &= 0.02 \times 0.60 + 0.05 \times 0.40 \\ &= 0.012 + 0.020 = 0.032 \end{aligned}$$

The overall defect rate is 3.2%.

Tree diagram representation:



To find $\mathbb{P}(D)$, multiply along each path to D and sum: $(0.60 \times 0.02) + (0.40 \times 0.05) = 0.032$.

2.3.2 Combining Bayes' Rule with LOTP

When applying Bayes' Rule, we often need to compute $\mathbb{P}(B)$ in the denominator. LOTP provides a systematic way to do this.

Theorem 2.5 (Bayes' Rule with LOTP)**.** *Let $\{A_1, \ldots, A_n\}$ be a partition of S with $\mathbb{P}(A_i) > 0$ for all i, and let B be an event with $\mathbb{P}(B) > 0$. Then:*

$$\mathbb{P}(A_j \mid B) = \frac{\mathbb{P}(B \mid A_j) \cdot \mathbb{P}(A_j)}{\sum_{i=1}^{n} \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i)} \tag{2.6}$$

Bayes' Rule with LOTP

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

$$\mathbb{P}(A_j \mid B) = \frac{\mathbb{P}(B \mid A_j) \cdot \mathbb{P}(A_j)}{\sum_{i=1}^{n} \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i)}$$

Example 2.3 (Medical Testing)**.** A disease affects 1% of the population. A test for the disease has:

- Sensitivity (true positive rate): $\mathbb{P}(\text{positive} \mid \text{disease}) = 0.95$
- Specificity (true negative rate): $\mathbb{P}(\text{negative} \mid \text{no disease}) = 0.90$

If a randomly selected person tests positive, what is the probability they have the disease?

Solution: Let D = has disease, T^+ = tests positive.

We want $\mathbb{P}(D \mid T^+)$. Using Bayes' Rule with LOTP:

$$\mathbb{P}(D \mid T^+) = \frac{\mathbb{P}(T^+ \mid D) \cdot \mathbb{P}(D)}{\mathbb{P}(T^+ \mid D) \cdot \mathbb{P}(D) + \mathbb{P}(T^+ \mid D^c) \cdot \mathbb{P}(D^c)}$$

We have:

$$\begin{aligned} \mathbb{P}(D) = 0.01, \quad \mathbb{P}(D^c) = 0.99 \\ \mathbb{P}(T^+ \mid D) = 0.95 \\ \mathbb{P}(T^+ \mid D^c) = 1 - 0.90 = 0.10 \quad \text{(false positive rate)} \end{aligned}$$

Substituting:

$$\begin{aligned} \mathbb{P}(D \mid T^+) &= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.10 \times 0.99} \\ &= \frac{0.0095}{0.0095 + 0.099} = \frac{0.0095}{0.1085} \approx 0.088 \end{aligned}$$

Despite testing positive, there is only about an 8.8% chance of actually having the disease!

Base Rate Fallacy

The medical testing example illustrates the **base rate fallacy**. The low prior probability ($\mathbb{P}(D) = 0.01$) dominates: even with a positive test, the posterior probability remains modest. This is because the false positives from the large healthy population (99% of people) outnumber the true positives from the small diseased population (1% of people). **Key insight**: When a condition is rare, even accurate tests produce many false positives relative to true positives.

2.3.3 Bayes' Rule with Extra Conditioning

Sometimes we want to apply Bayes' Rule within a restricted context, conditioning on additional evidence E.

Theorem 2.6 (Bayes' Rule with Extra Conditioning). *For events A, B, and E with $\mathbb{P}(A \cap E) > 0$ and $\mathbb{P}(B \cap E) > 0$:*

$$\mathbb{P}(A \mid B, E) = \frac{\mathbb{P}(B \mid A, E) \cdot \mathbb{P}(A \mid E)}{\mathbb{P}(B \mid E)} \tag{2.7}$$

The LOTP also has a version with extra conditioning:

Theorem 2.7 (LOTP with Extra Conditioning). *Let $\{A_1, \ldots, A_n\}$ be a partition with $\mathbb{P}(A_i \cap E) > 0$ for all i. Then:*

$$\mathbb{P}(B \mid E) = \sum_{i=1}^{n} \mathbb{P}(B \mid A_i, E) \cdot \mathbb{P}(A_i \mid E) \tag{2.8}$$

Combining these:

Bayes' Rule with LOTP and Extra Conditioning

$$\mathbb{P}(A_j \mid B, E) = \frac{\mathbb{P}(B \mid A_j, E) \cdot \mathbb{P}(A_j \mid E)}{\sum_{i=1}^{n} \mathbb{P}(B \mid A_i, E) \cdot \mathbb{P}(A_i \mid E)}$$

2.4 The Monty Hall Problem

The Monty Hall problem is a classic probability puzzle that illustrates the subtleties of conditional probability. It is named after the host of the American television game show "Let's Make a Deal."

Example 2.4 (The Monty Hall Problem)**.** You are on a game show with three doors. Behind one door is a car; behind the other two are goats. You choose a door (say, Door 1). The host, Monty Hall, who knows what's behind each door, opens one of the other two doors to reveal a goat (say, Door 3). He then asks: "Would you like to switch to Door 2?"

Question: Should you switch? What is the probability of winning if you switch versus if you stay?

2.4.1 Setup and Assumptions

The problem requires careful specification of the assumptions:

1. The car is equally likely to be behind any of the three doors
2. The contestant initially chooses Door 1
3. Monty always opens a door that:
   - The contestant did not choose
   - Does not contain the car
   - If two doors satisfy these conditions, Monty chooses uniformly at random between them
4. Monty then offers the contestant the option to switch

2.4.2 Solution Using Bayes' Rule

Let C_i denote the event that the car is behind Door i, and let M_j denote the event that Monty opens Door j.

Prior probabilities (before any doors are opened):

$$\mathbb{P}(C_1) = \mathbb{P}(C_2) = \mathbb{P}(C_3) = \frac{1}{3}$$

Likelihoods (probability Monty opens Door 3, given where the car is):

$$\mathbb{P}(M_3 \mid C_1) = \frac{1}{2} \quad \text{(car at Door 1: Monty chooses randomly between Doors 2 and 3)}$$
$$\mathbb{P}(M_3 \mid C_2) = 1 \quad \text{(car at Door 2: Monty must open Door 3)}$$
$$\mathbb{P}(M_3 \mid C_3) = 0 \quad \text{(car at Door 3: Monty cannot open Door 3)}$$

Applying LOTP to find $\mathbb{P}(M_3)$:

$$\begin{aligned}\mathbb{P}(M_3) &= \mathbb{P}(M_3 \mid C_1)\,\mathbb{P}(C_1) + \mathbb{P}(M_3 \mid C_2)\,\mathbb{P}(C_2) + \mathbb{P}(M_3 \mid C_3)\,\mathbb{P}(C_3) \\ &= \frac{1}{2}\cdot\frac{1}{3} + 1\cdot\frac{1}{3} + 0\cdot\frac{1}{3} = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}\end{aligned}$$

Applying Bayes' Rule to find posterior probabilities:

$$\mathbb{P}(C_1 \mid M_3) = \frac{\mathbb{P}(M_3 \mid C_1)\,\mathbb{P}(C_1)}{\mathbb{P}(M_3)} = \frac{\frac{1}{2}\cdot\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

$$\mathbb{P}(C_2 \mid M_3) = \frac{\mathbb{P}(M_3 \mid C_2)\,\mathbb{P}(C_2)}{\mathbb{P}(M_3)} = \frac{1\cdot\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

$$\mathbb{P}(C_3 \mid M_3) = \frac{\mathbb{P}(M_3 \mid C_3)\,\mathbb{P}(C_3)}{\mathbb{P}(M_3)} = \frac{0\cdot\frac{1}{3}}{\frac{1}{2}} = 0$$

Monty Hall Solution

After Monty opens Door 3 to reveal a goat:

- Probability of winning by **staying** with Door 1: $\frac{1}{3}$
- Probability of winning by **switching** to Door 2: $\frac{2}{3}$

You should switch! Switching doubles your probability of winning.

2.4.3 Intuitive Explanation

Why Switching Wins More Often

Here is an intuitive way to understand the result:
Initial choice: When you first pick Door 1, you have a $\frac{1}{3}$ chance of being right and a $\frac{2}{3}$ chance of being wrong.
Monty's action: Monty's reveal doesn't change where the car is—it only provides information. Crucially, Monty *always* reveals a goat, so his action is constrained by the car's location.
Two scenarios:

1. If you initially chose correctly ($\frac{1}{3}$ chance): Switching loses.
2. If you initially chose incorrectly ($\frac{2}{3}$ chance): Monty reveals the other goat, so switching wins.

Since you initially chose incorrectly $\frac{2}{3}$ of the time, switching wins $\frac{2}{3}$ of the time.
Alternative perspective: Imagine there are 100 doors with 1 car and 99 goats. You pick one door. Monty then opens 98 doors, all revealing goats. Would you switch to the one remaining door? Almost certainly yes—your initial chance of being right was only 1%.

The Role of Monty's Knowledge

The solution critically depends on Monty *knowing* where the car is and *always* opening a door with a goat. If Monty opened a random door (that happened to have a goat), the probabilities would be different. The information Monty's action provides depends on the *process* by which he chooses which door to open.

2.5 Independence of Events

Two events are independent if learning that one occurred gives no information about whether the other occurred.

Definition 2.3 (Independence of Two Events)**.** Events A and B are **independent** if:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B) \tag{2.9}$$

We write $A \perp B$ to denote that A and B are independent.

Theorem 2.8 (Equivalent Characterisations of Independence)**.** *The following are equivalent (assuming* $\mathbb{P}(A), \mathbb{P}(B) > 0$*):*

1. $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$
2. $\mathbb{P}(A \mid B) = \mathbb{P}(A)$
3. $\mathbb{P}(B \mid A) = \mathbb{P}(B)$

Proof. (1) $\Rightarrow$ (2): If $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$, then:

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

(2) $\Rightarrow$ (1): If $\mathbb{P}(A \mid B) = \mathbb{P}(A)$, then:

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

The equivalence of (1) and (3) follows by symmetry. ■

Independence

Events A and B are independent if and only if knowing whether B occurred does not change the probability of A:

$$A \perp B \quad \Longleftrightarrow \quad \mathbb{P}(A \mid B) = \mathbb{P}(A) \quad \Longleftrightarrow \quad \mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$$

2.5.1 Properties of Independence

Theorem 2.9 (Independence is Symmetric)**.** *If* $A \perp B$*, then* $B \perp A$*.*

Proof. Independence is defined by $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$, which is symmetric in A and B. ■

Theorem 2.10 (Independence with Complements)**.** *If A and B are independent, then so are:*

1. *A and B^c*
2. *A^c and B*
3. *A^c and B^c*

Proof. We prove (1); the others follow similarly.

$$\begin{aligned}\mathbb{P}(A \cap B^c) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \quad (\text{since } A = (A \cap B) \cup (A \cap B^c)) \\ &= \mathbb{P}(A) - \mathbb{P}(A)\,\mathbb{P}(B) \quad (\text{by independence of } A, B) \\ &= \mathbb{P}(A)(1 - \mathbb{P}(B)) \\ &= \mathbb{P}(A)\,\mathbb{P}(B^c)\end{aligned}$$

■

2.5.2 Independence vs Disjointness

Independence ≠ Disjointness

Independence and disjointness are fundamentally different concepts:

- **Disjoint** (mutually exclusive): $A \cap B = \emptyset$, so $\mathbb{P}(A \cap B) = 0$
- **Independent**: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$

If A and B are disjoint with $\mathbb{P}(A), \mathbb{P}(B) > 0$, then they are *not* independent. In fact, they are strongly dependent: knowing A occurred tells you B definitely did not occur!
Disjoint events with positive probability can only be independent if at least one has probability zero.

Example 2.5 (Disjoint Implies Dependent)**.** Let $A =$ "roll a 1 on a fair die" and $B =$ "roll a 6 on a fair die." These events are disjoint ($A \cap B = \emptyset$).

Check independence:

$$\begin{aligned}\mathbb{P}(A) \cdot \mathbb{P}(B) &= \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} \\ \mathbb{P}(A \cap B) &= 0\end{aligned}$$

Since $0 \neq \frac{1}{36}$, the events are not independent. Indeed, knowing you rolled a 1 tells you with certainty that you did not roll a 6.

2.5.3 Independence of Multiple Events

For more than two events, independence requires more than just pairwise independence.

Definition 2.4 (Mutual Independence)**.** Events $A_1, A_2, \ldots, A_n$ are **mutually independent** if for every subset $\{i_1, i_2, \ldots, i_k\} \subseteq \{1, 2, \ldots, n\}$:

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}) \tag{2.10}$$

Example 2.6 (Three Events: Pairwise vs Mutual Independence)**.** For three events A, B, C to be mutually independent, we need *all four* conditions:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) \tag{2.11}$$
$$\mathbb{P}(A \cap C) = \mathbb{P}(A)\,\mathbb{P}(C) \tag{2.12}$$
$$\mathbb{P}(B \cap C) = \mathbb{P}(B)\,\mathbb{P}(C) \tag{2.13}$$
$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C) \tag{2.14}$$

Conditions (2.11)–(2.13) constitute **pairwise independence**. Condition (2.14) is additional.

Pairwise independence does not imply mutual independence.

Example 2.7 (Pairwise but Not Mutually Independent)**.** Consider two fair coin flips. Let:

$$A = \text{“first flip is Heads”}$$
$$B = \text{“second flip is Heads”}$$
$$C = \text{“both flips give the same result”}$$

We have $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$.

Checking pairwise independence:

$$\mathbb{P}(A \cap B) = \mathbb{P}(HH) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A)\,\mathbb{P}(B) \quad \checkmark$$
$$\mathbb{P}(A \cap C) = \mathbb{P}(HH) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A)\,\mathbb{P}(C) \quad \checkmark$$
$$\mathbb{P}(B \cap C) = \mathbb{P}(HH) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(B)\,\mathbb{P}(C) \quad \checkmark$$

Checking triple independence:

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(HH) = \frac{1}{4} \neq \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C) \quad \times$$

The events are pairwise independent but *not* mutually independent.

2.5.4 Conditional Independence

Events can be independent given some information, even if they are not unconditionally independent (and vice versa).

Definition 2.5 (Conditional Independence)**.** Events A and B are **conditionally independent given** E if:

$$\mathbb{P}(A \cap B \mid E) = \mathbb{P}(A \mid E) \cdot \mathbb{P}(B \mid E) \tag{2.15}$$

We write $A \perp B \mid E$.

Independence and Conditional Independence Are Different

Three important facts:

1. **Conditional independence given E does not imply conditional independence given E^c**

2. **Conditional independence does not imply (unconditional) independence**
3. **Independence does not imply conditional independence**

Example 2.8 (Phone Call Example)**.** Consider whether two friends, Alice and Bob, call you on a given day. Suppose their decisions to call are made independently.

Let:

- A = "Alice calls"
- B = "Bob calls"
- E = "exactly one friend calls"

Independence: A and B are independent (by assumption).

Conditional dependence given E: Given that exactly one friend called, knowing Alice called tells you Bob did not call. So A and B are *dependent* given E:

$$\mathbb{P}(B \mid A, E) = 0 \neq \mathbb{P}(B \mid E)$$

Conditional independence given E^c: Given that it's not the case that exactly one friend called (i.e., zero or two called), knowing Alice called tells you Bob must have called too. So A and B are also dependent given E^c.

This example shows that independence does not imply conditional independence, and conditional independence given E says nothing about conditional independence given E^c.

Why Conditioning Can Create Dependence

Conditioning on an event that involves both A and B can create dependence even if A and B are initially independent. The event E acts as a "constraint" linking A and B—once we know E occurred, the outcomes of A and B must jointly satisfy this constraint, inducing dependence.
This phenomenon is sometimes called **explaining away** or **Berkson's paradox** in causal inference.

2.6 Random Variables

A random variable provides a numerical summary of the outcome of a random experiment.

Definition 2.6 (Random Variable)**.** A **random variable** is a function $X : S \to \mathbb{R}$ that assigns a real number $X(s)$ to each outcome s in the sample space S.

Random Variables as Numerical Summaries

A random variable extracts a number from each experimental outcome. The outcome s contains all the details of what happened; the random variable $X(s)$ is a numerical summary of interest.
For example, if we flip a coin 10 times, the outcome s is a sequence like $HHTTHTHHTH$.
Possible random variables include:

- $X(s)$ = number of heads in s
- $Y(s)$ = length of longest run of heads
- $Z(s)$ = 1 if the first flip is heads, 0 otherwise

Example 2.9 (Coin Flips)**.** Consider flipping a fair coin twice. The sample space is $S = \{HH, HT, TH, TT\}$.

Define random variables:

- X = number of heads: $X(HH) = 2$, $X(HT) = 1$, $X(TH) = 1$, $X(TT) = 0$
- Y = number of tails: $Y(HH) = 0$, $Y(HT) = 1$, $Y(TH) = 1$, $Y(TT) = 2$
- I = indicator for first flip being heads: $I(HH) = 1$, $I(HT) = 1$, $I(TH) = 0$, $I(TT) = 0$

Note that $Y(s) = 2 - X(s)$ for all s—random variables can be related through functions.

Definition 2.7 (Support of a Random Variable)**.** The **support** of a random variable X is the set of all values x such that $\mathbb{P}(X = x) > 0$:

$$\text{supp}(X) = \{x \in \mathbb{R} : \mathbb{P}(X = x) > 0\}$$

2.7 Probability Mass Functions

For discrete random variables (those taking countably many values), the probability mass function completely describes the distribution.

Definition 2.8 (Probability Mass Function)**.** The **probability mass function (PMF)** of a discrete random variable X is the function $p_X : \mathbb{R} \to [0, 1]$ defined by:

$$p_X(x) = \mathbb{P}(X = x) \tag{2.16}$$

The PMF gives the probability that X takes each possible value. It is positive on the support of X and zero elsewhere.

Theorem 2.11 (Properties of PMFs)**.** *A function $p : \mathbb{R} \to \mathbb{R}$ is a valid PMF if and only if:*

1. ***Non-negativity**: $p(x) \geq 0$ for all x*
2. ***Normalisation**: $\sum_x p(x) = 1$ (summing over all x in the support)*

Proof. **Necessity**: If $p = p_X$ for some random variable X:

1. $p_X(x) = \mathbb{P}(X = x) \geq 0$ by non-negativity of probability.
2. The events $\{X = x\}$ for x in the support form a partition of S, so:

$$\sum_x p_X(x) = \sum_x \mathbb{P}(X = x) = \mathbb{P}(S) = 1$$

Sufficiency: Any function satisfying these properties can be used to define a probability distribution on the support, which in turn defines a random variable. ∎

PMF Properties

For any PMF p_X:

1. $p_X(x) \geq 0$ for all x
2. $\sum_x p_X(x) = 1$
3. $\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x)$ for any set A

Example 2.10 (Constructing a PMF)**.** Let X be the number of heads in two fair coin flips.

Step 1: Identify the sample space and map outcomes to values:

Outcome s	$X(s)$
TT	0
TH	1
HT	1
HH	2

Step 2: Calculate probabilities for each value of X:

$$p_X(0) = \mathbb{P}(X = 0) = \mathbb{P}(\{TT\}) = \frac{1}{4}$$
$$p_X(1) = \mathbb{P}(X = 1) = \mathbb{P}(\{TH, HT\}) = \frac{2}{4} = \frac{1}{2}$$
$$p_X(2) = \mathbb{P}(X = 2) = \mathbb{P}(\{HH\}) = \frac{1}{4}$$

Verification: $\frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$ ✓

The PMF is:

$$p_X(x) = \begin{cases} 1/4 & \text{if } x = 0 \\ 1/2 & \text{if } x = 1 \\ 1/4 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

2.8 Cumulative Distribution Functions

The cumulative distribution function provides an alternative characterisation of a random variable's distribution.

Definition 2.9 (Cumulative Distribution Function)**.** The **cumulative distribution function (CDF)** of a random variable X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by:

$$F_X(x) = \mathbb{P}(X \leq x) \tag{2.17}$$

Theorem 2.12 (Properties of CDFs)**.** *Any CDF F satisfies:*

1. ***Monotonicity**: F is non-decreasing: if $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$*
2. ***Limits at infinity**: $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$*

3. ***Right-continuity***: *F is right-continuous:* $\lim_{h \to 0^+} F(x+h) = F(x)$

Proof. **Monotonicity**: If $x_1 \leq x_2$, then $\{X \leq x_1\} \subseteq \{X \leq x_2\}$, so by monotonicity of probability:

$$F(x_1) = \mathbb{P}(X \leq x_1) \leq \mathbb{P}(X \leq x_2) = F(x_2)$$

Limits: As $x \to -\infty$, the events $\{X \leq x\}$ decrease to the empty set, so $F(x) \to 0$. As $x \to \infty$, the events $\{X \leq x\}$ increase to S, so $F(x) \to 1$.

Right-continuity: This follows from the continuity of probability for decreasing sequences of events. ■

2.8.1 Relationship Between PMF and CDF

For discrete random variables, the PMF and CDF are related by:

Theorem 2.13 (PMF-CDF Relationship)**.** *For a discrete random variable X with PMF p_X and CDF F_X:*

1. ***CDF from PMF***:

$$F_X(x) = \sum_{t \leq x} p_X(t) \tag{2.18}$$

2. ***PMF from CDF***: *At any point x in the support,*

$$p_X(x) = F_X(x) - \lim_{t \to x^-} F_X(t) \tag{2.19}$$

The PMF equals the size of the jump in the CDF at x.

CDF for Discrete Random Variables

- The CDF of a discrete random variable is a step function
- Jumps occur at points in the support; the jump size equals the PMF value
- Between jumps, the CDF is constant (flat)

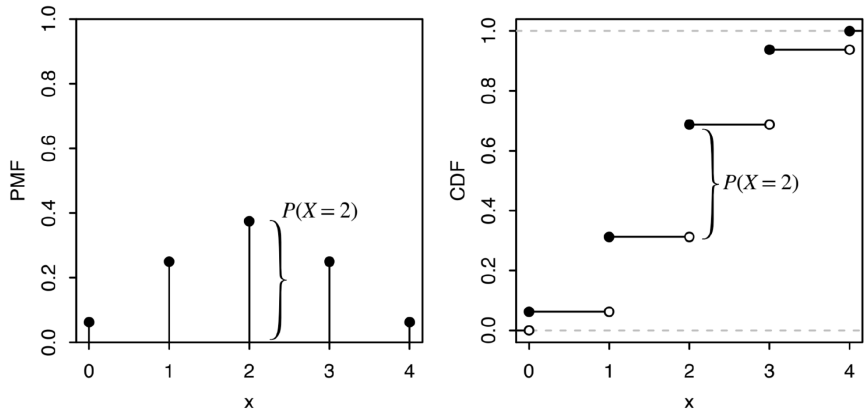


Figure 2.2: PMF and CDF for $X \sim \text{Binomial}(4, 1/2)$. The PMF shows point masses at $x = 0, 1, 2, 3, 4$. The CDF is a step function that jumps at each point in the support, with jump sizes equal to the PMF values.

2.9 Common Discrete Distributions

Several discrete distributions arise repeatedly in applications. Understanding these "named" distributions provides building blocks for modelling real-world phenomena.

2.9.1 Bernoulli Distribution

The Bernoulli distribution models a single trial with two outcomes (success/failure).

Definition 2.10 (Bernoulli Distribution)**.** A random variable X has a **Bernoulli distribution** with parameter $p \in [0,1]$, written $X \sim \text{Bern}(p)$, if:

$$p_X(x) = \mathbb{P}(X = x) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.20}$$

Equivalently: $p_X(x) = p^x(1-p)^{1-x}$ for $x \in \{0,1\}$.

Bernoulli Distribution

$X \sim \text{Bern}(p)$:

- Support: $\{0,1\}$
- PMF: $\mathbb{P}(X=1) = p$, $\mathbb{P}(X=0) = 1-p$
- Interpretation: Models a single trial with success probability p

Example 2.11 (Bernoulli Applications)**.**

- Coin flip: $X = 1$ if heads, $X = 0$ if tails; $X \sim \text{Bern}(0.5)$ for a fair coin
- Customer purchase: $X = 1$ if purchase made, $X = 0$ otherwise
- Medical test result: $X = 1$ if positive, $X = 0$ if negative

2.9.2 Binomial Distribution

The Binomial distribution models the number of successes in multiple independent trials.

Definition 2.11 (Binomial Distribution)**.** A random variable X has a **Binomial distribution** with parameters $n \in \{0,1,2,\ldots\}$ and $p \in [0,1]$, written $X \sim \text{Bin}(n,p)$, if:

$$p_X(k) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0,1,2,\ldots,n \tag{2.21}$$

Deriving the Binomial PMF

Consider n independent trials, each with success probability p. Let X count the number of successes.
To have exactly k successes:

- We need to choose which k of the n trials are successes: $\binom{n}{k}$ ways

- Those k trials must all succeed: probability p^k
- The remaining $n-k$ trials must all fail: probability $(1-p)^{n-k}$

By independence, the probability of any specific arrangement is $p^k(1-p)^{n-k}$. Multiplying by the number of arrangements gives the PMF.

Binomial Distribution

$X \sim \text{Bin}(n, p)$:

- Support: $\{0, 1, 2, \ldots, n\}$
- PMF: $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- Interpretation: Number of successes in n independent Bernoulli(p) trials
- Special case: $\text{Bin}(1, p) = \text{Bern}(p)$

Theorem 2.14 (Sum of Bernoullis). *If $X_1, X_2, \ldots, X_n$ are independent Bern(p) random variables, then:*

$$X_1 + X_2 + \cdots + X_n \sim Bin(n, p)$$

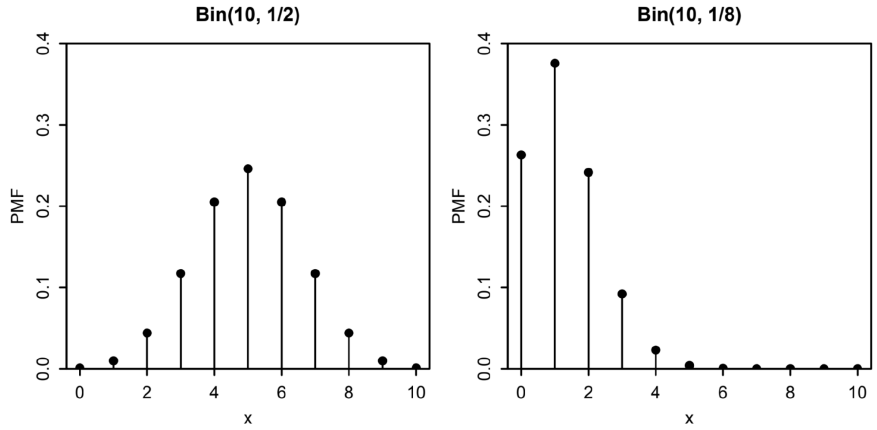


Figure 2.3: Binomial PMFs for various values of n and p. As n increases, the distribution becomes more spread out. The parameter p controls the location of the peak.

Example 2.12 (Quality Control). A factory produces items with a 5% defect rate. In a batch of 20 items, what is the probability of finding exactly 2 defective items?

Let X = number of defective items. Then $X \sim \text{Bin}(20, 0.05)$.

$$\begin{aligned}\mathbb{P}(X = 2) &= \binom{20}{2}(0.05)^2(0.95)^{18} \\ &= 190 \times 0.0025 \times 0.3972 \\ &\approx 0.189\end{aligned}$$

2.9.3 Discrete Uniform Distribution

The Discrete Uniform distribution assigns equal probability to each value in a finite set.

Definition 2.12 (Discrete Uniform Distribution)**.** A random variable X has a **Discrete Uniform distribution** on a finite set C, written $X \sim \text{DUnif}(C)$, if:

$$p_X(x) = \mathbb{P}(X = x) = \begin{cases} \frac{1}{|C|} & \text{if } x \in C \\ 0 & \text{otherwise} \end{cases} \tag{2.22}$$

Discrete Uniform Distribution

$X \sim \text{DUnif}(C)$:

- Support: The finite set C
- PMF: $\mathbb{P}(X = x) = 1/|C|$ for all $x \in C$
- For any subset $A \subseteq C$: $\mathbb{P}(X \in A) = |A|/|C|$

Example 2.13 (Dice Roll)**.** Let X be the result of rolling a fair six-sided die. Then $X \sim \text{DUnif}(\{1, 2, 3, 4, 5, 6\})$:

$$\mathbb{P}(X = k) = \frac{1}{6} \quad \text{for } k = 1, 2, 3, 4, 5, 6$$

The probability of rolling an even number:

$$\mathbb{P}(X \in \{2, 4, 6\}) = \frac{|\{2, 4, 6\}|}{|\{1, 2, 3, 4, 5, 6\}|} = \frac{3}{6} = \frac{1}{2}$$

Connection to Naive Probability

The Discrete Uniform distribution is the probabilistic analogue of the naive definition of probability. When $X \sim \text{DUnif}(C)$, computing $\mathbb{P}(X \in A)$ reduces to counting: we just need to find $|A|$ and $|C|$.

2.10 Summary

This chapter developed the core concepts of conditional probability and introduced random variables:

Chapter Summary

Conditional Probability:

- $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ — probability of A given B occurred
- Represents "zooming in" on outcomes where B occurred and re-normalising
- $\mathbb{P}(A \mid B) \neq \mathbb{P}(B \mid A)$ in general; chronology does not determine conditioning direction

Bayes' Rule and LOTP:

- Bayes' Rule: $\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A)\,\mathbb{P}(A)/\mathbb{P}(B)$
- Updates prior belief $\mathbb{P}(A)$ to posterior $\mathbb{P}(A \mid B)$ using likelihood $\mathbb{P}(B \mid A)$

- LOTP: $\mathbb{P}(B) = \sum_i \mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)$ for any partition $\{A_i\}$
- Combined: Use LOTP to compute the denominator in Bayes' Rule

Independence:

- $A \perp B$ iff $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$ iff $\mathbb{P}(A \mid B) = \mathbb{P}(A)$
- Independence $\neq$ disjointness (disjoint events with positive probability are dependent)
- Mutual independence requires all subset products, not just pairwise
- Independence and conditional independence are distinct concepts

Random Variables and Distributions:

- Random variable: function $X : S \to \mathbb{R}$ assigning numbers to outcomes
- PMF: $p_X(x) = \mathbb{P}(X = x)$; must be non-negative and sum to 1
- CDF: $F_X(x) = \mathbb{P}(X \leq x)$; non-decreasing, right-continuous, limits 0 and 1
- Common distributions: Bernoulli (single trial), Binomial (n trials), Discrete Uniform (equally likely)

Looking ahead: Chapter 3 extends these ideas to joint distributions of multiple random variables, introducing expectation and variance as summary measures.

Chapter 3

Joint Random Variables

Learning Objectives

After completing this chapter, you should be able to:

- Define joint random variables and understand how functions of multiple random variables work
- Distinguish between joint, marginal, and conditional probability mass functions
- Apply marginalisation to extract single-variable distributions from joint distributions
- Verify independence of random variables using the factorisation criterion
- Explain the distinction between independence and conditional independence
- Compute expectations and variances for common distributions (Binomial, Multinomial, Poisson)
- Apply the linearity of expectation and understand when variance is additive
- Introduce covariance as a measure of how random variables vary together

Prerequisites: This chapter assumes familiarity with basic probability theory (**??**) and conditional probability (**??**), including Bayes' theorem and the law of total probability. You should be comfortable with the definitions of random variables, probability mass functions, and cumulative distribution functions.

3.1 Joint Random Variables and Their Distributions

When modelling real-world phenomena, we often need to consider multiple random quantities simultaneously. For instance, we might want to study both the height and weight of individuals in a population, or the number of sunny days and average temperature in a month. Joint random variables provide the framework for analysing such relationships.

3.1.1 Functions of Multiple Random Variables

Definition 3.1 (Joint Random Variable)**.** Given an experiment with sample space S, suppose X and Y are random variables that map each outcome $s \in S$ to values $X(s)$ and $Y(s)$ respectively.

For any function $g : \mathbb{R}^2 \to \mathbb{R}$, the composition $g(X, Y)$ defines a new random variable that maps each outcome s to $g(X(s), Y(s))$.

This definition tells us that when we have two random variables X and Y defined on the same sample space, any function of these variables is itself a random variable. Common examples include:

- The sum $X + Y$
- The product XY
- The maximum $\max(X, Y)$
- The indicator $\mathbf{1}_{X>Y}$ (equals 1 if $X > Y$, 0 otherwise)

Example 3.1 (Sum of Dice Rolls)**.** Let X be the outcome of rolling a fair six-sided die, and let Y be the outcome of rolling a second independent fair die. The sum $S = X + Y$ is a new random variable taking values in $\{2, 3, \ldots, 12\}$. The probability that $S = 7$ is determined by counting the pairs (x, y) with $x + y = 7$:

$$P(S = 7) = P(\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}) = \frac{6}{36} = \frac{1}{6}.$$

3.1.2 The Random Walk (Motivating Example)

A *random walk* is a mathematical model that describes a path consisting of successive random steps. It serves as an excellent illustration of how joint random variables arise naturally.

Example 3.2 (Simple Random Walk)**.** Consider a particle starting at position 0 on the integer line. At each time step, it moves either one step to the right (with probability p) or one step to the left (with probability $1 - p$). Let X_i denote the i-th step:

$$X_i = \begin{cases} +1 & \text{with probability } p \\ -1 & \text{with probability } 1 - p \end{cases}$$

After n steps, the particle's position is $S_n = X_1 + X_2 + \cdots + X_n$. The joint distribution of $(X_1, X_2, \ldots, X_n)$ determines the distribution of S_n.

For a symmetric random walk ($p = 1/2$), the position after n steps has:

- Expected position: $\mathbb{E}[S_n] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n \cdot 0 = 0$
- Variance: $\mathrm{Var}(S_n) = n \,\mathrm{Var}(X_1) = n \cdot 1 = n$ (since the X_i are independent)

The number of right steps R follows a Binomial(n, p) distribution, and the final position is $S_n = R - (n - R) = 2R - n$. Thus:

$$P(S_n = k) = P(R = \tfrac{n+k}{2}) = \binom{n}{\frac{n+k}{2}} p^{\frac{n+k}{2}} (1 - p)^{\frac{n-k}{2}}$$

for $k \in \{-n, -n + 2, \ldots, n - 2, n\}$ (i.e., k has the same parity as n).

3.2 Joint Distributions: How Two Random Variables Interact

The joint distribution captures all probabilistic information about how two (or more) random variables behave together. We now formalise the key concepts.

3.2.1 Joint Probability Mass Function

Definition 3.2 (Joint PMF)**.** For discrete random variables X and Y, the **joint probability mass function** is defined as:

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

for all possible values x and y.

Properties of Joint PMF

The joint PMF must satisfy two conditions:

1. **Non-negativity:** $p_{X,Y}(x, y) \geq 0$ for all x, y
2. **Normalisation:** $\sum_x \sum_y p_{X,Y}(x, y) = 1$

The joint PMF can be visualised as a three-dimensional bar chart, where the height of each bar at position (x, y) represents the probability $P(X = x, Y = y)$.

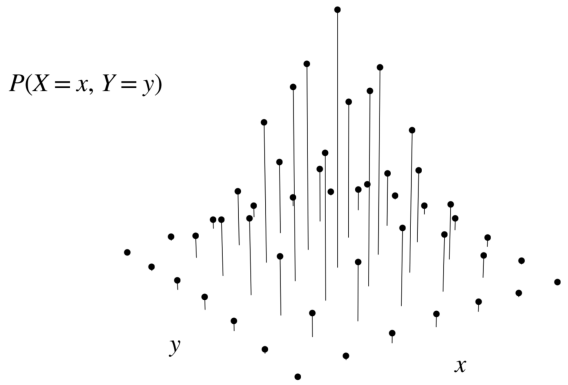


Figure 3.1: Visualisation of a joint PMF as a 3D bar chart. Each bar's height represents the probability of the corresponding (x, y) pair.

3.2.2 Joint Cumulative Distribution Function

Definition 3.3 (Joint CDF)**.** For random variables X and Y (discrete or continuous), the **joint cumulative distribution function** is:

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

The joint CDF gives the probability that X is at most x *and* Y is at most y simultaneously. For discrete random variables, it can be computed from the joint PMF:

$$F_{X,Y}(x,y) = \sum_{x' \le x} \sum_{y' \le y} p_{X,Y}(x', y')$$

3.3 Marginal Distributions

Given the joint distribution of (X, Y), we can recover the distribution of X alone (or Y alone) by *marginalising* over the other variable.

3.3.1 Marginal PMF

Definition 3.4 (Marginal PMF)**.** The **marginal PMF of** X is obtained by summing the joint PMF over all values of Y:

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

Similarly, the **marginal PMF of** Y is:

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

Geometric Interpretation of Marginalisation

Marginalisation "flattens" the joint distribution along one axis. If the joint PMF is visualised as a 3D bar chart, the marginal PMF of X is obtained by projecting all bars onto the X-axis, summing the heights of bars with the same x-coordinate.

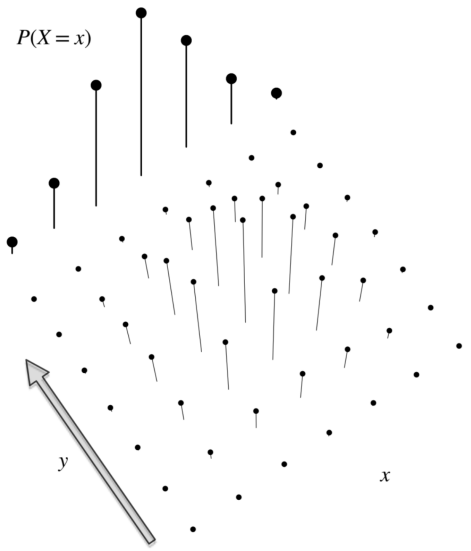


Figure 3.2: Marginal PMF of X obtained by summing over all values of Y. The 2D distribution along the X-axis represents $p_X(x)$.

The term "marginal" comes from the practice of writing row and column totals in the margins of contingency tables, as we shall see in the examples below.

Theorem 3.1 (Law of Total Probability for Random Variables)**.** *The marginal PMF $p_X(x)$ is a valid probability mass function. That is:*

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

Proof. Non-negativity follows from the non-negativity of the joint PMF. For normalisation:

$$\sum_x p_X(x) = \sum_x \sum_y p_{X,Y}(x,y) = \sum_x \sum_y P(X = x, Y = y) = 1$$

where the last equality uses the normalisation property of the joint PMF. ■

3.4 Conditional Distributions

Conditional distributions describe the behaviour of one random variable given knowledge about another.

3.4.1 Conditional PMF

Definition 3.5 (Conditional PMF)**.** The **conditional PMF of Y given $X = x$** is:

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

provided that $P(X = x) > 0$.

Key Relationship: Joint

The three types of distributions are related by:

$$\underbrace{p_{X,Y}(x,y)}_{\text{Joint}} = \underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} \cdot \underbrace{p_X(x)}_{\text{Marginal}}$$

This is the *chain rule of probability* for random variables.

Geometric Interpretation

If the joint PMF is a 3D bar chart, the conditional PMF $p_{Y|X}(y|x)$ is obtained by taking a "slice" at a fixed value of $X = x$, then renormalising so that the slice sums to 1. You are conditioning on knowing the value of X, and asking: given this information, what is the distribution of Y?

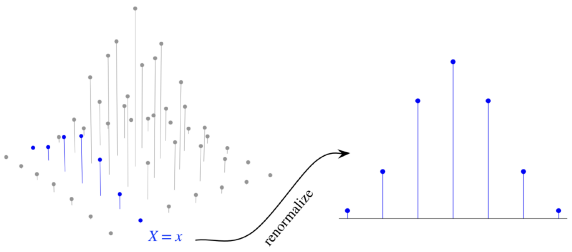


Figure 3.3: Conditional PMF of Y given $X = x$: a cross-section of the joint distribution, renormalised.

3.5 Independence of Random Variables

Independence is a fundamental concept that captures when knowledge of one variable provides no information about another.

Definition 3.6 (Independence of Random Variables)**.** Random variables X and Y are **independent** if and only if their joint distribution factorises as the product of their marginals:

- **Discrete case:** $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ for all x, y
- **Continuous case:** $P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y)$ for all x, y

Testing for Independence

To verify that X and Y are independent, you must check that $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ holds for **all** pairs (x, y).
To show that X and Y are **not** independent, it suffices to find **one** pair (x, y) where the equality fails.

3.5.1 Conditional Independence

Independence can also be defined relative to a third variable.

Definition 3.7 (Conditional Independence)**.** Random variables X and Y are **conditionally independent given** Z if:

$$P(X \leq x, Y \leq y \mid Z = z) = P(X \leq x \mid Z = z) \cdot P(Y \leq y \mid Z = z)$$

for all x, y, and z where $P(Z = z) > 0$.

Independence vs Conditional Independence

Independence does not imply conditional independence, and **conditional independence does not imply independence**. These are distinct concepts.

Example 3.3 (Independence Does Not Imply Conditional Independence)**.** Let X indicate whether your friend Bob calls you next Friday, and Y indicate whether your friend Alice calls. Suppose X and Y are independent (Bob's and Alice's decisions are unrelated).

Let Z indicate that *exactly one* friend calls. Conditional on $Z = 1$, if $X = 1$ (Bob calls), then necessarily $Y = 0$ (Alice does not call), and vice versa. Thus X and Y are perfectly (negatively) dependent given Z, even though they are marginally independent.

Example 3.4 (Conditional Independence Does Not Imply Independence)**.** Consider a medical scenario where X is an indicator for smoking, Y is an indicator for lung cancer, and Z is an indicator for a genetic marker that predisposes to both smoking behaviour and lung cancer.

It may be that X and Y are conditionally independent given Z (once we control for genetics, smoking and cancer are unrelated in this hypothetical), yet marginally dependent (smokers have higher cancer rates in the population).

Remark. The distinction between independence and conditional independence is central to causal inference. Much of applied causal analysis involves finding conditions under which variables become conditionally independent, enabling identification of causal effects. This idea underlies techniques such as regression adjustment and instrumental variables.

3.6 Worked Example: Gene and Disease

Let us work through a complete example involving joint, marginal, and conditional distributions.

Example 3.5 (Gene and Disease Association)**.** A random sample from a population yields the following data:

- X: indicator for the presence of a certain gene (1 = present, 0 = absent)
- Y: indicator for developing a certain disease (1 = disease, 0 = no disease)

The joint distribution is given by the contingency table:

	$Y = 1$	$Y = 0$	Marginal of X
$X = 1$	5/100	20/100	25/100
$X = 0$	3/100	72/100	75/100
Marginal of Y	8/100	92/100	1

Table 3.1: Contingency table for gene (X) and disease (Y) with marginal distributions shown in the margins.

Joint PMF: The entries in the interior of the table give the joint PMF directly. For instance, $P(X = 1, Y = 1) = 5/100 = 0.05$.

Marginal PMFs:

- $P(X = 1) = 5/100 + 20/100 = 25/100 = 0.25$
- $P(X = 0) = 3/100 + 72/100 = 75/100 = 0.75$
- $P(Y = 1) = 5/100 + 3/100 = 8/100 = 0.08$
- $P(Y = 0) = 20/100 + 72/100 = 92/100 = 0.92$

Conditional Distribution of Y given $X = 1$:

$$P(Y = 1 \mid X = 1) = \frac{P(X = 1, Y = 1)}{P(X = 1)} = \frac{5/100}{25/100} = \frac{5}{25} = 0.2$$

$$P(Y = 0 \mid X = 1) = \frac{P(X = 1, Y = 0)}{P(X = 1)} = \frac{20/100}{25/100} = \frac{20}{25} = 0.8$$

Thus, the conditional distribution of Y given $X = 1$ is Bernoulli(0.2): among those with the gene, 20% develop the disease.

Testing for Independence:

$$P(X = 1) \cdot P(Y = 1) = 0.25 \times 0.08 = 0.02$$
$$P(X = 1, Y = 1) = 0.05$$

Since $0.05 \neq 0.02$, we have found a pair where the factorisation fails, so X and Y are **not independent**. The gene and disease are associated—having the gene increases the probability of disease from the baseline rate of 8% to 20%.

3.7 Extended Worked Example: Joint to Marginal to Conditional

The following example demonstrates the systematic computation of marginal and conditional distributions from a joint PMF.

Example 3.6 (Complete Analysis of a Joint Distribution)**.** Consider two discrete random variables X and Y with the following joint distribution:

$p_{X,Y}(x,y)$	$Y=1$	$Y=2$	$Y=3$
$X=1$	0.10	0.20	0.10
$X=2$	0.05	0.25	0.10
$X=3$	0.05	0.10	0.05

Step 1: Verify Normalisation

$$\sum_x \sum_y p_{X,Y}(x,y) = 0.10 + 0.20 + 0.10 + 0.05 + 0.25 + 0.10 + 0.05 + 0.10 + 0.05 = 1.00 \checkmark$$

Step 2: Compute Marginal PMF of X

$$p_X(1) = p_{X,Y}(1,1) + p_{X,Y}(1,2) + p_{X,Y}(1,3) = 0.10 + 0.20 + 0.10 = 0.40$$
$$p_X(2) = p_{X,Y}(2,1) + p_{X,Y}(2,2) + p_{X,Y}(2,3) = 0.05 + 0.25 + 0.10 = 0.40$$
$$p_X(3) = p_{X,Y}(3,1) + p_{X,Y}(3,2) + p_{X,Y}(3,3) = 0.05 + 0.10 + 0.05 = 0.20$$

Step 3: Compute Marginal PMF of Y

$$p_Y(1) = p_{X,Y}(1,1) + p_{X,Y}(2,1) + p_{X,Y}(3,1) = 0.10 + 0.05 + 0.05 = 0.20$$
$$p_Y(2) = p_{X,Y}(1,2) + p_{X,Y}(2,2) + p_{X,Y}(3,2) = 0.20 + 0.25 + 0.10 = 0.55$$
$$p_Y(3) = p_{X,Y}(1,3) + p_{X,Y}(2,3) + p_{X,Y}(3,3) = 0.10 + 0.10 + 0.05 = 0.25$$

Step 4: Compute Conditional PMF of X given Y

For $Y = 1$ (where $p_Y(1) = 0.20$):

$$p_{X|Y}(1|1) = \frac{0.10}{0.20} = 0.50, \quad p_{X|Y}(2|1) = \frac{0.05}{0.20} = 0.25, \quad p_{X|Y}(3|1) = \frac{0.05}{0.20} = 0.25$$

For $Y = 2$ (where $p_Y(2) = 0.55$):

$$p_{X|Y}(1|2) = \frac{0.20}{0.55} \approx 0.364, \quad p_{X|Y}(2|2) = \frac{0.25}{0.55} \approx 0.455, \quad p_{X|Y}(3|2) = \frac{0.10}{0.55} \approx 0.182$$

For $Y = 3$ (where $p_Y(3) = 0.25$):

$$p_{X|Y}(1|3) = \frac{0.10}{0.25} = 0.40, \quad p_{X|Y}(2|3) = \frac{0.10}{0.25} = 0.40, \quad p_{X|Y}(3|3) = \frac{0.05}{0.25} = 0.20$$

Step 5: Compute Conditional PMF of Y given X

For $X = 1$ (where $p_X(1) = 0.40$):

$$p_{Y|X}(1|1) = \frac{0.10}{0.40} = 0.25, \quad p_{Y|X}(2|1) = \frac{0.20}{0.40} = 0.50, \quad p_{Y|X}(3|1) = \frac{0.10}{0.40} = 0.25$$

For $X = 2$ (where $p_X(2) = 0.40$):

$$p_{Y|X}(1|2) = \frac{0.05}{0.40} = 0.125, \quad p_{Y|X}(2|2) = \frac{0.25}{0.40} = 0.625, \quad p_{Y|X}(3|2) = \frac{0.10}{0.40} = 0.25$$

For $X = 3$ (where $p_X(3) = 0.20$):

$$p_{Y|X}(1|3) = \frac{0.05}{0.20} = 0.25, \quad p_{Y|X}(2|3) = \frac{0.10}{0.20} = 0.50, \quad p_{Y|X}(3|3) = \frac{0.05}{0.20} = 0.25$$

3.8 Expectation

The expectation (or expected value, or mean) of a random variable is a measure of its "centre"—a weighted average of the possible values, where the weights are the probabilities.

Definition 3.8 (Expected Value)**.** For a discrete random variable X, the **expected value** (or **mean**) is:

$$\mathbb{E}[X] = \sum_x x \cdot P(X = x) = \sum_x x \cdot p_X(x)$$

Interpretation of Expectation

The expected value $\mathbb{E}[X]$ represents the long-run average of X if the experiment were repeated infinitely many times. It need not be a value that X can actually take.

Example 3.7 (Fair Die)**.** Let X be the result of rolling a fair six-sided die. Each outcome has probability $1/6$, so:

$$\mathbb{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5$$

Note that X never actually equals 3.5—the expected value is a theoretical average, not a possible outcome.

Example 3.8 (Coin Flips)**.** Let X be the number of heads when flipping a fair coin twice. The PMF is:

$$P(X = 0) = \frac{1}{4}, \quad P(X = 1) = \frac{1}{2}, \quad P(X = 2) = \frac{1}{4}$$

Thus:

$$\mathbb{E}[X] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 0 + \frac{1}{2} + \frac{1}{2} = 1$$

Example 3.9 (Bernoulli Distribution)**.** For $X \sim \text{Bernoulli}(p)$:

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

The expected value of a Bernoulli random variable is simply its success probability.

3.8.1 Linearity of Expectation

One of the most powerful properties of expectation is its linearity.

Theorem 3.2 (Linearity of Expectation)**.** *For any random variables X and Y (not necessarily independent) and any constants a, b, c:*

1. $\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b$
2. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
3. *More generally:* $\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i + c\right] = \sum_{i=1}^{n} a_i\,\mathbb{E}[X_i] + c$

Linearity Only Works for Linear Functions

Linearity of expectation does **not** extend to non-linear functions. In general:

$$\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$$

For example, $\mathbb{E}[X^2] \neq (\mathbb{E}[X])^2$ in general. The difference $\mathbb{E}[X^2] - (\mathbb{E}[X])^2$ is precisely the variance, as we shall see.

Example 3.10 (Linearity in Action)**.** Let X be the number of heads in two coin flips ($\mathbb{E}[X] = 1$), and let Z be the result of rolling a fair die ($\mathbb{E}[Z] = 3.5$). For $W = X + Z$:

$$\mathbb{E}[W] = \mathbb{E}[X + Z] = \mathbb{E}[X] + \mathbb{E}[Z] = 1 + 3.5 = 4.5$$

This holds regardless of whether X and Z are independent.

3.9 Variance

While expectation measures the centre of a distribution, variance measures its spread—how far values typically deviate from the mean.

Definition 3.9 (Variance)**.** The **variance** of a random variable X with mean $\mu = \mathbb{E}[X]$ is:

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

The **standard deviation** is $\sigma = \sqrt{\mathrm{Var}(X)}$.

Computational Formula for Variance

Expanding the definition using linearity of expectation yields a computationally convenient formula:

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

This says: "the variance is the expected value of the square minus the square of the expected value."

Derivation of computational formula.

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2X\,\mathbb{E}[X] + (\mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - 2\,\mathbb{E}[X]\,\mathbb{E}[X] + (\mathbb{E}[X])^2 \quad \text{(linearity, and } \mathbb{E}[X] \text{ is a constant)} \\ &= \mathbb{E}[X^2] - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}$$

■

Example 3.11 (Variance of a Fair Die)**.** Let X be the result of rolling a fair six-sided die. We computed $\mathbb{E}[X] = 3.5$.

For $\mathbb{E}[X^2]$:

$$\mathbb{E}[X^2] = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{1 + 4 + 9 + 16 + 25 + 36}{6} = \frac{91}{6} \approx 15.17$$

Thus:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{91}{6} - (3.5)^2 = \frac{91}{6} - \frac{49}{4} = \frac{182 - 147}{12} = \frac{35}{12} \approx 2.92$$

3.9.1 Properties of Variance

Theorem 3.3 (Variance Properties)**.** *For random variables X and Y and constants a, c:*

1. $\mathrm{Var}(c) = 0$ *for any constant c*
2. $\mathrm{Var}(X + c) = \mathrm{Var}(X)$ *(shifting by a constant does not change spread)*
3. $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$ *(scaling by a scales variance by a^2)*
4. *If X and Y are* ***independent****:* $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Variance is NOT Linear

Unlike expectation, variance does not obey simple linearity:

- $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X) \neq a\,\mathrm{Var}(X)$ (unless $a \in \{0, 1\}$)
- $\mathrm{Var}(X + Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$ in general (only holds for independent X, Y)

The general formula is $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$, where covariance measures the dependence between X and Y. See Section 3.11 and **??** for details.

3.10 Mean and Variance of Common Distributions

We now derive the mean and variance for several important discrete distributions. For comprehensive reference material on these and other distributions, see **??**.

3.10.1 Bernoulli Distribution

Definition 3.10 (Bernoulli Distribution)**.** A random variable X follows a **Bernoulli distribution** with parameter $p \in [0,1]$, written $X \sim \text{Bernoulli}(p)$, if:

$$P(X = 1) = p, \quad P(X = 0) = 1 - p$$

Theorem 3.4 (Bernoulli Mean and Variance)**.** *For $X \sim$ Bernoulli(p):*

$$\mathbb{E}[X] = p, \qquad \text{Var}(X) = p(1-p)$$

Proof. **Mean:**

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

Variance: First compute $\mathbb{E}[X^2]$. Since $X \in \{0,1\}$, we have $X^2 = X$, so $\mathbb{E}[X^2] = \mathbb{E}[X] = p$. Thus:

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

■

Remark. The variance $p(1-p)$ is maximised when $p = 1/2$, giving $\text{Var}(X) = 1/4$. This makes intuitive sense: a fair coin has maximum uncertainty, while a coin that always lands heads ($p = 1$) or always tails ($p = 0$) has zero variance.

3.10.2 Binomial Distribution

Definition 3.11 (Binomial Distribution)**.** A random variable X follows a **Binomial distribution** with parameters $n \in \mathbb{N}$ and $p \in [0,1]$, written $X \sim \text{Binomial}(n,p)$, if it represents the number of successes in n independent Bernoulli(p) trials:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

Theorem 3.5 (Binomial Mean and Variance)**.** *For $X \sim$ Binomial(n,p):*

$$\mathbb{E}[X] = np, \qquad \text{Var}(X) = np(1-p)$$

Proof. The key insight is that a Binomial(n,p) random variable can be written as a sum of independent Bernoulli(p) random variables. Let $X_1, X_2, \ldots, X_n$ be independent with $X_i \sim$ Bernoulli(p). Then $X = X_1 + X_2 + \cdots + X_n$.

Mean: By linearity of expectation:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p = np$$

Variance: Since the X_i are independent:

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

■

Intuition

The mean np is exactly what you would expect: if each trial succeeds with probability p, then on average np out of n trials will succeed. The variance $np(1-p)$ grows linearly with n, but the standard deviation $\sqrt{np(1-p)}$ grows only as $\sqrt{n}$, so the *relative* spread (coefficient of variation) decreases as n increases.

Example 3.12 (Coin Flipping)**.** Flip a fair coin 100 times. Let X be the number of heads. Then $X \sim \text{Binomial}(100, 0.5)$, so:

$$\mathbb{E}[X] = 100 \times 0.5 = 50, \qquad \text{Var}(X) = 100 \times 0.5 \times 0.5 = 25$$

The standard deviation is $\sigma = 5$, so we expect roughly 50 ± 10 heads (within two standard deviations) about 95% of the time.

3.10.3 Multinomial Distribution

The multinomial distribution generalises the binomial to experiments with more than two possible outcomes.

Definition 3.12 (Multinomial Distribution)**.** Consider n independent trials, each resulting in one of k categories with probabilities $p_1, p_2, \ldots, p_k$ (where $\sum_{j=1}^{k} p_j = 1$). Let X_j be the count of outcomes in category j. The vector $(X_1, \ldots, X_k)$ follows a **Multinomial distribution**:

$$P(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

where $n_1 + n_2 + \cdots + n_k = n$.

Theorem 3.6 (Multinomial Marginal Moments)**.** *For a Multinomial$(n; p_1, \ldots, p_k)$ distribution, each marginal X_j satisfies:*

$$\mathbb{E}[X_j] = np_j, \qquad \text{Var}(X_j) = np_j(1-p_j)$$

Proof. Each category j can be viewed as a "success" with probability p_j, while all other outcomes are "failures." Thus X_j, the count in category j, follows a Binomial(n, p_j) distribution marginally. The result follows from Theorem 3.5. ■

Remark. The multinomial distribution also has a covariance structure. For $i \neq j$:

$$\text{Cov}(X_i, X_j) = -np_i p_j$$

The negative covariance reflects the constraint $\sum_j X_j = n$: if more outcomes fall in category i, fewer can fall in category j.

3.10.4 Poisson Distribution

The Poisson distribution models the number of events occurring in a fixed interval when events happen at a constant average rate.

Definition 3.13 (Poisson Distribution)**.** A random variable X follows a **Poisson distribution** with parameter $\lambda > 0$, written $X \sim \text{Poisson}(\lambda)$, if:

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

Theorem 3.7 (Poisson Mean and Variance)**.** *For $X \sim Poisson(\lambda)$:*

$$\mathbb{E}[X] = \lambda, \qquad \mathrm{Var}(X) = \lambda$$

Proof. **Mean:**

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} k \cdot \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{(the } k=0 \text{ term vanishes)} \\
&= \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda} \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} \quad \text{(substituting } m = k-1) \\
&= \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda
\end{aligned}$$

Variance: We compute $\mathbb{E}[X(X-1)]$ first (this is a useful trick):

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{k=0}^{\infty} k(k-1) \cdot \frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=2}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-2)!} \\
&= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = \lambda^2 e^{-\lambda} \cdot e^{\lambda} = \lambda^2
\end{aligned}$$

Now:

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

Therefore:

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

■

Poisson: Mean Equals Variance

A distinctive property of the Poisson distribution is that $\mathbb{E}[X] = \mathrm{Var}(X) = \lambda$. This property is sometimes used as a diagnostic: if empirical data show the sample mean approximately equal to the sample variance, a Poisson model may be appropriate.

Example 3.13 (Poisson Approximation to Binomial)**.** When n is large and p is small such that $\lambda = np$ is moderate, the Binomial(n, p) distribution is well-approximated by Poisson(λ). This can be seen by comparing moments:

- Binomial mean: $np = \lambda$ ✓
- Binomial variance: $np(1-p) \approx np = \lambda$ when p is small ✓

3.11 Introduction to Covariance

We have seen that $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ only holds when X and Y are independent. To understand what happens in the general case, we need the concept of covariance.

Definition 3.14 (Covariance)**.** The **covariance** of random variables X and Y is:

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

Covariance Computational Formula

Expanding the definition gives:

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

Covariance is "the expectation of the product minus the product of the expectations."

Intuition

Covariance measures how X and Y "vary together":

- If X tends to be above its mean when Y is above its mean (and below when below), then $\mathrm{Cov}(X,Y) > 0$ (positive covariance)
- If X tends to be above its mean when Y is below its mean, then $\mathrm{Cov}(X,Y) < 0$ (negative covariance)
- If there is no systematic relationship, $\mathrm{Cov}(X,Y) \approx 0$

Theorem 3.8 (Variance of a Sum). *For any random variables X and Y:*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$$

Corollary 3.9. *If X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$, so $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.*

Zero Covariance Does Not Imply Independence

If $\mathrm{Cov}(X,Y) = 0$, we say X and Y are **uncorrelated**. However, uncorrelated does not imply independent. Covariance only captures *linear* relationships; X and Y can be perfectly dependent (e.g., $Y = X^2$) yet have zero covariance.
A detailed treatment of covariance, correlation, and the distinction between independence and uncorrelatedness appears in **??**.

3.12 Summary

Chapter Summary

Joint Distributions:

- Joint PMF: $p_{X,Y}(x,y) = P(X = x, Y = y)$
- Joint CDF: $F_{X,Y}(x,y) = P(X \le x, Y \le y)$

Marginal and Conditional:

- Marginal: $p_X(x) = \sum_y p_{X,Y}(x,y)$
- Conditional: $p_{Y|X}(y|x) = \dfrac{p_{X,Y}(x,y)}{p_X(x)}$
- Chain rule: $p_{X,Y}(x,y) = p_{Y|X}(y|x) \cdot p_X(x)$

Independence:

- $X \perp Y$ iff $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all x, y
- Independence $\not\Rightarrow$ Conditional independence

Expectation and Variance:

- $\mathbb{E}[X] = \sum_x x \cdot P(X = x)$
- Linearity: $\mathbb{E}[aX + bY + c] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y] + c$ (always)
- $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
- $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ only if $X \perp Y$

Key Distribution Moments:

Distribution	Mean	Variance
Bernoulli(p)	p	$p(1-p)$
Binomial(n, p)	np	$np(1-p)$
Poisson(λ)	λ	λ

Part II

Calculus

Chapter 4

Calculus I: Differentiation and Maximum Likelihood Estimation

Learning Objectives

By the end of this chapter, you should be able to:

- State and apply the fundamental rules of differentiation: constant, power, sum, product, quotient, and chain rules
- Derive and apply formulas for exponential and logarithmic functions
- Understand the relationship between differentiation and finding optima
- Construct likelihood functions from probability models
- Apply maximum likelihood estimation to derive estimators for common distributions
- Use the log-likelihood to simplify optimisation problems
- Verify maxima using the second derivative test

Prerequisites

This chapter assumes familiarity with:

- Basic algebraic manipulation and function notation
- The concept of a limit (intuitive understanding suffices)
- Probability distributions: Bernoulli, Binomial, Poisson (from Chapter 2)
- The i.i.d. assumption for random samples

4.1 Differentiation: Foundations

Differentiation is the mathematical tool for measuring *instantaneous rates of change*. Given a function $f(x)$, its derivative $f'(x)$ tells us how rapidly f changes as x varies. In data science, differentiation is indispensable for:

- **Optimisation**: Finding maxima and minima of functions (e.g., maximising likelihood, minimising loss)
- **Sensitivity analysis**: Understanding how outputs depend on inputs
- **Gradient-based learning**: Training neural networks via backpropagation

Definition 4.1 (Derivative)**.** The **derivative** of a function f at a point x is defined as the limit:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{4.1}$$

provided this limit exists. Alternative notations include $\frac{df}{dx}$, $\frac{d}{dx}f(x)$, and $Df(x)$.

The Derivative as a Slope

The derivative $f'(x)$ is the slope of the tangent line to the curve $y = f(x)$ at the point $(x, f(x))$. The limit definition captures this by considering the slope of secant lines through $(x, f(x))$ and $(x+h, f(x+h))$, then taking $h \to 0$ to get the tangent.

4.2 Differentiation Rules

Rather than computing derivatives from the limit definition each time, we use a collection of rules that cover most practical cases. We present each rule with a formal statement and proof.

4.2.1 Rule 1: Constant Rule

Theorem 4.1 (Constant Rule)**.** *If $f(x) = c$ where c is a constant, then:*

$$\frac{d}{dx}[c] = 0$$

Proof. Using the limit definition:

$$\frac{d}{dx}[c] = \lim_{h \to 0} \frac{c - c}{h} = \lim_{h \to 0} \frac{0}{h} = \lim_{h \to 0} 0 = 0$$

■

Intuition

A constant function is a horizontal line. Horizontal lines have zero slope everywhere, so the derivative is zero.

4.2.2 Rule 2: Power Rule

Theorem 4.2 (Power Rule). *For any real number n:*

$$\frac{d}{dx}[x^n] = nx^{n-1} \tag{4.2}$$

Power Rule

$$\frac{d}{dx}[x^n] = nx^{n-1}$$

Examples: $\frac{d}{dx}[x^3] = 3x^2$, $\frac{d}{dx}[x^{-1}] = -x^{-2}$, $\frac{d}{dx}[\sqrt{x}] = \frac{d}{dx}[x^{1/2}] = \frac{1}{2}x^{-1/2}$

Proof. We prove this for positive integer n using the binomial theorem. For $f(x) = x^n$:

$$f'(x) = \lim_{h\to 0} \frac{(x+h)^n - x^n}{h}$$

Expanding $(x+h)^n$ using the binomial theorem:

$$(x+h)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k}h^k = x^n + nx^{n-1}h + \binom{n}{2}x^{n-2}h^2 + \cdots + h^n$$

Therefore:

$$\begin{aligned} f'(x) &= \lim_{h\to 0} \frac{x^n + nx^{n-1}h + \binom{n}{2}x^{n-2}h^2 + \cdots + h^n - x^n}{h} \\ &= \lim_{h\to 0} \frac{nx^{n-1}h + \binom{n}{2}x^{n-2}h^2 + \cdots + h^n}{h} \\ &= \lim_{h\to 0} \left(nx^{n-1} + \binom{n}{2}x^{n-2}h + \cdots + h^{n-1} \right) \\ &= nx^{n-1} \end{aligned}$$

The result extends to all real n via logarithmic differentiation (see Section 4.3.4). ∎

4.2.3 Rule 3: Constant Multiple Rule

Theorem 4.3 (Constant Multiple Rule). *If c is a constant and f is differentiable, then:*

$$\frac{d}{dx}[c \cdot f(x)] = c \cdot \frac{d}{dx}[f(x)] = c \cdot f'(x) \tag{4.3}$$

Proof. Using the limit definition:

$$\begin{aligned} \frac{d}{dx}[c \cdot f(x)] &= \lim_{h\to 0} \frac{c \cdot f(x+h) - c \cdot f(x)}{h} \\ &= \lim_{h\to 0} c \cdot \frac{f(x+h) - f(x)}{h} \\ &= c \cdot \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \\ &= c \cdot f'(x) \end{aligned}$$

∎

Intuition

Constants "factor out" of derivatives. This is because differentiation is a *linear operator*: scaling a function scales its rate of change by the same factor.

4.2.4 Rule 4: Sum and Difference Rule

Theorem 4.4 (Sum and Difference Rule)**.** *If f and g are differentiable, then:*

$$\frac{d}{dx}[f(x) \pm g(x)] = f'(x) \pm g'(x) \tag{4.4}$$

Proof. For the sum (the difference follows analogously):

$$\begin{aligned}\frac{d}{dx}[f(x) + g(x)] &= \lim_{h \to 0} \frac{[f(x+h) + g(x+h)] - [f(x) + g(x)]}{h} \\ &= \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} + \lim_{h \to 0} \frac{g(x+h) - g(x)}{h} \\ &= f'(x) + g'(x)\end{aligned}$$

■

Linearity of Differentiation

Combining the constant multiple and sum rules, differentiation is a **linear operator**:

$$\frac{d}{dx}[af(x) + bg(x)] = a \cdot f'(x) + b \cdot g'(x)$$

for any constants a and b.

4.2.5 Rule 5: Product Rule

Theorem 4.5 (Product Rule)**.** *If f and g are differentiable, then:*

$$\frac{d}{dx}[f(x) \cdot g(x)] = f'(x) \cdot g(x) + f(x) \cdot g'(x) \tag{4.5}$$

Product Rule

$$(fg)' = f'g + fg'$$

Mnemonic: "derivative of first times second, plus first times derivative of second."

Proof. Using a clever algebraic manipulation:

$$\frac{d}{dx}[f(x)g(x)] = \lim_{h \to 0} \frac{f(x+h)g(x+h) - f(x)g(x)}{h}$$

We add and subtract $f(x+h)g(x)$ in the numerator:

$$\begin{aligned}
&= \lim_{h\to 0} \frac{f(x+h)g(x+h) - f(x+h)g(x) + f(x+h)g(x) - f(x)g(x)}{h} \\
&= \lim_{h\to 0} \frac{f(x+h)[g(x+h) - g(x)] + g(x)[f(x+h) - f(x)]}{h} \\
&= \lim_{h\to 0} f(x+h) \cdot \frac{g(x+h) - g(x)}{h} + \lim_{h\to 0} g(x) \cdot \frac{f(x+h) - f(x)}{h} \\
&= f(x) \cdot g'(x) + g(x) \cdot f'(x)
\end{aligned}$$

where we used that $\lim_{h\to 0} f(x+h) = f(x)$ by continuity (differentiability implies continuity). ■

Example 4.1 (Product Rule Application)**.** Find $\frac{d}{dx}[x^2 \sin(x)]$.

Solution: Let $f(x) = x^2$ and $g(x) = \sin(x)$. Then $f'(x) = 2x$ and $g'(x) = \cos(x)$.

$$\frac{d}{dx}[x^2 \sin(x)] = 2x \cdot \sin(x) + x^2 \cdot \cos(x) = 2x \sin(x) + x^2 \cos(x)$$

The Derivative of a Product is NOT the Product of Derivatives

A common error is to assume $(fg)' = f' \cdot g'$. This is **false**. For example:

$$\frac{d}{dx}[x \cdot x] = \frac{d}{dx}[x^2] = 2x \neq 1 \cdot 1 = 1$$

Always use the product rule.

4.2.6 Rule 6: Quotient Rule

Theorem 4.6 (Quotient Rule)**.** *If f and g are differentiable and $g(x) \neq 0$, then:*

$$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{[g(x)]^2} \tag{4.6}$$

Quotient Rule

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$$

Mnemonic: "lo d-hi minus hi d-lo, over lo-lo" (where "hi" = numerator, "lo" = denominator).

Proof. We can derive this from the product rule by writing $\frac{f}{g} = f \cdot g^{-1}$:

$$\begin{aligned}
\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] &= \frac{d}{dx}[f(x) \cdot (g(x))^{-1}] \\
&= f'(x) \cdot (g(x))^{-1} + f(x) \cdot \frac{d}{dx}[(g(x))^{-1}]
\end{aligned}$$

Using the chain rule (proved below), $\frac{d}{dx}[(g(x))^{-1}] = -\frac{g'(x)}{(g(x))^2}$. Therefore:

$$\begin{aligned}
\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] &= \frac{f'(x)}{g(x)} - \frac{f(x) \cdot g'(x)}{(g(x))^2} \\
&= \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{(g(x))^2}
\end{aligned}$$

■

Example 4.2 (Quotient Rule Application)**.** Find $\frac{d}{dx}\left[\frac{x^2+1}{x-3}\right]$.

Solution: Let $f(x) = x^2 + 1$ and $g(x) = x - 3$. Then $f'(x) = 2x$ and $g'(x) = 1$.

$$\begin{aligned}\frac{d}{dx}\left[\frac{x^2+1}{x-3}\right] &= \frac{2x(x-3) - (x^2+1)(1)}{(x-3)^2} \\ &= \frac{2x^2 - 6x - x^2 - 1}{(x-3)^2} \\ &= \frac{x^2 - 6x - 1}{(x-3)^2}\end{aligned}$$

4.2.7 Rule 7: Chain Rule

The chain rule is perhaps the most important differentiation rule, as it allows us to differentiate *compositions* of functions.

Theorem 4.7 (Chain Rule)**.** *If g is differentiable at x and f is differentiable at $g(x)$, then the composite function $f \circ g$ is differentiable at x, and:*

$$\frac{d}{dx}[f(g(x))] = f'(g(x)) \cdot g'(x) \tag{4.7}$$

Chain Rule

Using Leibniz notation, if $y = f(u)$ and $u = g(x)$, then:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

In words: "the derivative of the outer function (evaluated at the inner) times the derivative of the inner function."

Proof Sketch. The intuitive argument proceeds as follows. If $u = g(x)$, then a small change Δx in x produces a change $\Delta u \approx g'(x)\Delta x$ in u. This change in u then produces a change $\Delta y \approx f'(u)\Delta u$ in y. Combining:

$$\frac{\Delta y}{\Delta x} \approx \frac{f'(u)\Delta u}{\Delta x} = f'(u) \cdot \frac{\Delta u}{\Delta x} \approx f'(g(x)) \cdot g'(x)$$

Taking limits as $\Delta x \to 0$ yields the result. A rigorous proof requires careful handling of the case when $g(x + h) = g(x)$ for arbitrarily small h. ■

Example 4.3 (Chain Rule Application)**.** Find $\frac{d}{dx}\left[(x^2+1)^{10}\right]$.

Solution: Let $u = g(x) = x^2 + 1$ (inner function) and $y = f(u) = u^{10}$ (outer function).

Step 1: Differentiate the outer function: $\frac{dy}{du} = 10u^9$

Step 2: Differentiate the inner function: $\frac{du}{dx} = 2x$

Step 3: Apply the chain rule:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx} = 10u^9 \cdot 2x = 10(x^2+1)^9 \cdot 2x = 20x(x^2+1)^9$$

Example 4.4 (Chain Rule: Normal Distribution PDF)**.** The standard normal probability density function is:

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

Find $\phi'(x)$.

Solution: First, factor out the constant $\frac{1}{\sqrt{2\pi}}$:

$$\phi'(x) = \frac{1}{\sqrt{2\pi}} \cdot \frac{d}{dx}\left[e^{-\frac{1}{2}x^2}\right]$$

Let $u = -\frac{1}{2}x^2$ (inner) and $y = e^u$ (outer). Then:

$$\frac{dy}{du} = e^u$$
$$\frac{du}{dx} = -x$$

By the chain rule:

$$\frac{d}{dx}\left[e^{-\frac{1}{2}x^2}\right] = e^{-\frac{1}{2}x^2} \cdot (-x) = -xe^{-\frac{1}{2}x^2}$$

Therefore:

$$\phi'(x) = \frac{1}{\sqrt{2\pi}} \cdot \left(-xe^{-\frac{1}{2}x^2}\right) = -\frac{x}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} = -x \cdot \phi(x)$$

This shows that the derivative of the standard normal PDF is $-x$ times the PDF itself.

4.3 Exponential and Logarithmic Derivatives

Exponential and logarithmic functions are ubiquitous in statistics and data science: they appear in probability distributions, likelihood functions, entropy, and information theory. Understanding their derivatives is essential.

4.3.1 The Natural Exponential Function

Theorem 4.8 (Derivative of the Natural Exponential)**.** *The function $f(x) = e^x$ is its own derivative:*

$$\frac{d}{dx}[e^x] = e^x \tag{4.8}$$

Exponential Derivative

$$\frac{d}{dx}[e^x] = e^x$$

This is the *defining property* of the number $e \approx 2.71828$. The exponential function is the unique function (up to scaling) that equals its own derivative.

Proof. This can be proved from the definition $e = \lim_{n\to\infty}(1+\frac{1}{n})^n$, but a cleaner approach uses the series definition:

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

Differentiating term by term:

$$\frac{d}{dx}[e^x] = 0 + 1 + \frac{2x}{2!} + \frac{3x^2}{3!} + \cdots = 1 + x + \frac{x^2}{2!} + \cdots = e^x$$

■

Corollary 4.9 (Chain Rule with Exponential). *For any differentiable function $u(x)$:*

$$\frac{d}{dx}[e^{u(x)}] = e^{u(x)} \cdot u'(x)$$

4.3.2 General Exponential Functions

Theorem 4.10 (Derivative of General Exponential). *For any positive constant $a > 0$, $a \neq 1$:*

$$\frac{d}{dx}[a^x] = a^x \ln(a) \tag{4.9}$$

Proof. We use the identity $a^x = e^{x \ln a}$. By the chain rule:

$$\frac{d}{dx}[a^x] = \frac{d}{dx}[e^{x \ln a}] = e^{x \ln a} \cdot \ln(a) = a^x \ln(a)$$

■

Power Rule vs Exponential Rule

The **power rule** applies when the *base* is the variable: $\frac{d}{dx}[x^n] = nx^{n-1}$
The **exponential rule** applies when the *exponent* is the variable: $\frac{d}{dx}[a^x] = a^x \ln(a)$
These are fundamentally different operations. A common error is to apply the power rule to exponential functions.

Example 4.5 (General Exponential Derivatives).

1. $\frac{d}{dx}[10^x] = 10^x \ln(10)$
2. $\frac{d}{dx}[2^{4x}]$: Here $u = 4x$, so by the chain rule:

$$\frac{d}{dx}[2^{4x}] = 2^{4x} \ln(2) \cdot 4 = 4 \ln(2) \cdot 2^{4x}$$

3. $\frac{d}{dx}[2^{4x} + 4x^2] = 4 \ln(2) \cdot 2^{4x} + 8x$

More generally, when differentiating $a^{u(x)}$ where u is a function of x:

$$\frac{d}{dx}[a^{u(x)}] = a^{u(x)} \cdot \ln(a) \cdot u'(x) \tag{4.10}$$

4.3.3 The Natural Logarithm

Theorem 4.11 (Derivative of the Natural Logarithm). *For $x > 0$:*

$$\frac{d}{dx}[\ln(x)] = \frac{1}{x} \tag{4.11}$$

Logarithmic Derivative

$$\frac{d}{dx}[\ln(x)] = \frac{1}{x}$$

This is the inverse relationship to $\frac{d}{dx}[e^x] = e^x$.

Proof. Since $\ln(x)$ is the inverse of e^x, we have $e^{\ln x} = x$. Differentiating both sides:

$$\frac{d}{dx}[e^{\ln x}] = \frac{d}{dx}[x]$$

By the chain rule on the left:

$$e^{\ln x} \cdot \frac{d}{dx}[\ln x] = 1$$

Since $e^{\ln x} = x$:

$$x \cdot \frac{d}{dx}[\ln x] = 1 \implies \frac{d}{dx}[\ln x] = \frac{1}{x}$$

■

Corollary 4.12 (Chain Rule with Logarithm). *For any positive differentiable function $u(x) > 0$:*

$$\frac{d}{dx}[\ln(u(x))] = \frac{u'(x)}{u(x)}$$

Example 4.6 (Logarithmic Derivatives).

1. $\frac{d}{dx}[\ln(x^2+1)]$: Let $u = x^2+1$, so $u' = 2x$:

$$\frac{d}{dx}[\ln(x^2+1)] = \frac{2x}{x^2+1}$$

2. $\frac{d}{dx}[\ln(1-\theta)]$: Let $u = 1-\theta$, so $\frac{du}{d\theta} = -1$:

$$\frac{d}{d\theta}[\ln(1-\theta)] = \frac{-1}{1-\theta} = -\frac{1}{1-\theta}$$

This derivative appears frequently in MLE derivations.

4.3.4 Logarithmic Differentiation

Logarithmic differentiation is a powerful technique for differentiating complicated products, quotients, and functions with variable bases and exponents.

Theorem 4.13 (General Power Rule via Logarithmic Differentiation). *For $x > 0$ and any real n:*

$$\frac{d}{dx}[x^n] = nx^{n-1}$$

Proof. Write $y = x^n$. Taking logarithms: $\ln y = n \ln x$.

Differentiating both sides with respect to x:

$$\frac{1}{y} \cdot \frac{dy}{dx} = \frac{n}{x}$$

Solving for $\frac{dy}{dx}$:

$$\frac{dy}{dx} = y \cdot \frac{n}{x} = x^n \cdot \frac{n}{x} = nx^{n-1}$$

■

4.4 Power Series: A Brief Introduction

Power series provide a way to represent functions as "infinite polynomials." This is particularly useful because polynomials are easy to differentiate, integrate, and manipulate.

Definition 4.2 (Power Series)**.** A **power series centred at** a is an infinite sum of the form:

$$\sum_{k=0}^{\infty} c_k(x-a)^k = c_0 + c_1(x-a) + c_2(x-a)^2 + c_3(x-a)^3 + \cdots \tag{4.12}$$

where $c_0, c_1, c_2, \ldots$ are constants called the **coefficients**. When $a = 0$, this simplifies to:

$$\sum_{k=0}^{\infty} c_k x^k = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \cdots$$

Why Power Series Matter

Many "ugly" functions (exponentials, logarithms, trigonometric functions) have power series representations. When we express a function as a power series:

- We can **differentiate** term by term using the power rule
- We can **integrate** term by term
- We can **approximate** the function by truncating the series

Convergence

Not every power series converges for all x. The **radius of convergence** R determines where the series is valid:

- If $|x-a| < R$: the series converges (the partial sums approach a finite limit)
- If $|x-a| > R$: the series diverges (the partial sums grow without bound)
- At $|x-a| = R$: convergence must be checked individually

For example, the geometric series $\sum_{k=0}^{\infty} x^k$ converges to $\frac{1}{1-x}$ only when $|x| < 1$.

Example 4.7 (Geometric Series)**.** For $|x| < 1$:

$$\sum_{k=0}^{\infty} x^k = 1 + x + x^2 + x^3 + \cdots = \frac{1}{1-x}$$

To verify: if $S = 1 + x + x^2 + \cdots$, then $xS = x + x^2 + x^3 + \cdots = S - 1$, so $S - xS = 1$, giving $S = \frac{1}{1-x}$.

The Taylor series (covered in detail in **??**) provides a systematic way to find the coefficients c_k for a given function. For now, the key insight is that power series connect to differentiation: the coefficients are determined by the function's derivatives at the centre point a.

4.5 Optimisation: Finding Maxima and Minima

One of the most important applications of differentiation is finding the maximum or minimum values of a function. This is directly relevant to Maximum Likelihood Estimation.

4.5.1 Critical Points

Definition 4.3 (Critical Point)**.** A **critical point** of a function f is a point $x = c$ where either:

1. $f'(c) = 0$ (the derivative is zero), or
2. $f'(c)$ does not exist

Theorem 4.14 (Fermat's Theorem)**.** *If f has a local maximum or minimum at c, and $f'(c)$ exists, then $f'(c) = 0$.*

Converse is False

The converse of Fermat's theorem is *not* true: $f'(c) = 0$ does not guarantee that c is a local extremum. For example, $f(x) = x^3$ has $f'(0) = 0$, but $x = 0$ is neither a maximum nor a minimum (it's an inflection point).

4.5.2 Second Derivative Test

The second derivative test helps classify critical points.

Theorem 4.15 (Second Derivative Test)**.** *Suppose f'' is continuous near c and $f'(c) = 0$. Then:*

1. *If $f''(c) > 0$, then f has a **local minimum** at c*
2. *If $f''(c) < 0$, then f has a **local maximum** at c*
3. *If $f''(c) = 0$, the test is **inconclusive***

Second Derivative Test for MLE

In Maximum Likelihood Estimation, we want to **maximise** the likelihood. After finding a critical point $\hat{\theta}$ where $\frac{d\ell}{d\theta}\big|_{\theta=\hat{\theta}} = 0$, we verify it's a maximum by checking:

$$\left.\frac{d^2\ell}{d\theta^2}\right|_{\theta=\hat{\theta}} < 0$$

A negative second derivative confirms we have a maximum.

Concavity and the Second Derivative

The second derivative measures **concavity**:

- $f''(x) > 0$: the function is **concave up** (shaped like a cup $\cup$) – holds water
- $f''(x) < 0$: the function is **concave down** (shaped like a cap $\cap$) – sheds water

At a critical point with $f'(c) = 0$:

- Concave up $\implies$ local minimum (the cup's bottom)
- Concave down $\implies$ local maximum (the cap's peak)

4.6 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a fundamental method for estimating the parameters of a statistical model. Given observed data, MLE finds the parameter values that make the observed data *most probable*.

4.6.1 The Core Idea

The MLE Philosophy

Suppose we observe data $x_1, x_2, \ldots, x_n$. We believe this data comes from some probability distribution with unknown parameter θ. Different values of θ assign different probabilities to our observed data:

- Some values of θ make our data look very likely
- Some values of θ make our data look very unlikely

MLE chooses the value of θ that makes our observed data *as likely as possible*.

4.6.2 The Likelihood Function

Definition 4.4 (Likelihood Function)**.** Given a statistical model with parameter θ and observed data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, the **likelihood function** is:

$$L(\theta; \boldsymbol{x}) = f(\boldsymbol{x}; \theta) \tag{4.13}$$

where $f(\boldsymbol{x}; \theta)$ is the joint PMF (for discrete data) or joint PDF (for continuous data), viewed as a function of θ with the data $\boldsymbol{x}$ held fixed.

Likelihood for i.i.d. Data

If $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) with common PMF/PDF $f(x; \theta)$, then:

$$L(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i; \theta) \tag{4.14}$$

The likelihood is the **product** of individual probabilities.

Likelihood is NOT Probability

The likelihood $L(\theta; \boldsymbol{x})$ is **not** a probability distribution over θ:

- It need not integrate to 1 over θ
- It should not be interpreted as $\mathbb{P}(\theta \mid \boldsymbol{x})$

Rather, likelihood measures the *plausibility* of different parameter values given the observed data. For Bayesian inference that does provide $\mathbb{P}(\theta \mid \boldsymbol{x})$, see later chapters.

4.6.3 The Log-Likelihood

In practice, we almost always work with the *logarithm* of the likelihood.

Definition 4.5 (Log-Likelihood)**.** The **log-likelihood function** is:

$$\ell(\theta; \boldsymbol{x}) = \ln L(\theta; \boldsymbol{x}) \tag{4.15}$$

Why Use Log-Likelihood?

The log-likelihood has several computational and numerical advantages:

1. **Products become sums**: For i.i.d. data, $\ell(\theta) = \sum_{i=1}^{n} \ln f(x_i; \theta)$
2. **Powers become products**: $\ln(\theta^k) = k \ln \theta$, simplifying derivatives
3. **Numerical stability**: Products of many small probabilities can underflow; sums of logs are more stable
4. **Same optimum**: Since $\ln$ is monotonically increasing, $\arg\max_\theta L(\theta) = \arg\max_\theta \ell(\theta)$

Key Logarithm Rules for MLE

$$\begin{aligned} \ln(ab) &= \ln(a) + \ln(b) \\ \ln\left(\frac{a}{b}\right) &= \ln(a) - \ln(b) \\ \ln\left(\prod_{j=1}^{n} a_j\right) &= \sum_{j=1}^{n} \ln(a_j) \\ \ln(a^b) &= b \ln(a) \end{aligned}$$

4.6.4 The MLE Procedure

Steps for Maximum Likelihood Estimation

1. **Specify the model**: Identify the distribution and its parameter(s) θ
2. **Write the likelihood**: $L(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i; \theta)$
3. **Take the log**: $\ell(\theta) = \sum_{i=1}^{n} \ln f(x_i; \theta)$
4. **Differentiate**: Compute $\frac{d\ell}{d\theta}$ (the *score function*)
5. **Set to zero and solve**: Solve $\frac{d\ell}{d\theta} = 0$ for $\hat{\theta}_{\text{MLE}}$
6. **Verify maximum**: Check $\frac{d^2\ell}{d\theta^2}\big|_{\theta=\hat{\theta}} < 0$

4.7 MLE Examples

We now work through MLE derivations for several important distributions. These examples illustrate the general procedure and build familiarity with the calculus involved.

4.7.1 Example 1: Bernoulli Distribution

Example 4.8 (MLE for Bernoulli Distribution)**.** Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli(θ), where $\theta \in (0,1)$ is the unknown success probability. Each X_i takes value 1 (success) with probability θ and 0 (failure) with probability $1-\theta$.

Step 1: PMF

$$f(x;\theta) = \theta^x(1-\theta)^{1-x}, \quad x \in \{0,1\}$$

Step 2: Likelihood

$$\begin{aligned} L(\theta;\boldsymbol{x}) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \end{aligned}$$

Let $s = \sum_{i=1}^n x_i$ denote the total number of successes. Then:

$$L(\theta) = \theta^s(1-\theta)^{n-s}$$

Step 3: Log-likelihood

$$\ell(\theta) = s\ln\theta + (n-s)\ln(1-\theta)$$

Step 4: Differentiate

$$\frac{d\ell}{d\theta} = \frac{s}{\theta} + (n-s)\cdot\frac{-1}{1-\theta} = \frac{s}{\theta} - \frac{n-s}{1-\theta}$$

Step 5: Set to zero and solve

$$\begin{aligned} 0 &= \frac{s}{\hat\theta} - \frac{n-s}{1-\hat\theta} \\ \frac{s}{\hat\theta} &= \frac{n-s}{1-\hat\theta} \\ s(1-\hat\theta) &= (n-s)\hat\theta \\ s - s\hat\theta &= n\hat\theta - s\hat\theta \\ s &= n\hat\theta \end{aligned}$$

Therefore:

$$\boxed{\hat\theta_{\text{MLE}} = \frac{s}{n} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}}$$

The MLE is the **sample proportion**: the number of successes divided by the number of trials.

Step 6: Verify maximum

$$\frac{d^2\ell}{d\theta^2} = -\frac{s}{\theta^2} - \frac{n-s}{(1-\theta)^2}$$

For $\theta \in (0,1)$ and $0 < s < n$, both terms are negative, so $\frac{d^2\ell}{d\theta^2} < 0$. This confirms we have a maximum.

4.7.2 Example 2: Binomial Distribution (Goalie Example)

Example 4.9 (MLE for Binomial: Goalie Saves)**.** A goalkeeper faces $m = 5$ penalty kicks in each of $n = 4$ games. The number of saves in each game is recorded:

$$\text{Game 1: 1 save}, \quad \text{Game 2: 3 saves}, \quad \text{Game 3: 2 saves}, \quad \text{Game 4: 2 saves}$$

Assuming each save is an independent Bernoulli trial with unknown success probability θ, we model $X_i \sim \text{Binomial}(5, \theta)$.

Step 1: PMF

The Binomial PMF for $m = 5$ trials is:

$$f(k;\theta) = \binom{5}{k}\theta^k(1-\theta)^{5-k}, \quad k \in \{0,1,2,3,4,5\}$$

Step 2: Likelihood

The observations are $x_1 = 1$, $x_2 = 3$, $x_3 = 2$, $x_4 = 2$. The likelihood is:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{4}\binom{5}{x_i}\theta^{x_i}(1-\theta)^{5-x_i} \\ &= \binom{5}{1}\binom{5}{3}\binom{5}{2}\binom{5}{2}\cdot\theta^{1+3+2+2}(1-\theta)^{(5-1)+(5-3)+(5-2)+(5-2)} \\ &= (5)(10)(10)(10)\cdot\theta^8(1-\theta)^{12} \\ &= 5000\cdot\theta^8(1-\theta)^{12} \end{aligned}$$

Step 3: Log-likelihood

$$\ell(\theta) = \ln(5000) + 8\ln\theta + 12\ln(1-\theta)$$

Step 4: Differentiate

$$\frac{d\ell}{d\theta} = \frac{8}{\theta} - \frac{12}{1-\theta}$$

Step 5: Set to zero and solve

$$\begin{aligned} 0 &= \frac{8}{\hat\theta} - \frac{12}{1-\hat\theta} \\ \frac{8}{\hat\theta} &= \frac{12}{1-\hat\theta} \\ 8(1-\hat\theta) &= 12\hat\theta \\ 8 - 8\hat\theta &= 12\hat\theta \\ 8 &= 20\hat\theta \end{aligned}$$

Therefore:

$$\boxed{\hat\theta_{\text{MLE}} = \frac{8}{20} = \frac{2}{5} = 0.4}$$

Interpretation: The MLE equals $\frac{\text{total saves}}{\text{total kicks}} = \frac{1+3+2+2}{4\times5} = \frac{8}{20} = 0.4$.

Step 6: Verify maximum

$$\frac{d^2\ell}{d\theta^2} = -\frac{8}{\theta^2} - \frac{12}{(1-\theta)^2} < 0$$

for all $\theta \in (0,1)$, confirming a maximum.

General Binomial MLE

For i.i.d. observations $X_1, \ldots, X_n$ from Binomial(m, θ):

$$\hat{\theta}_{\text{MLE}} = \frac{\sum_{i=1}^{n} x_i}{mn} = \frac{\text{total successes}}{\text{total trials}}$$

This is the sample proportion of successes across all trials.

4.7.3 Example 3: Poisson Distribution

Example 4.10 (MLE for Poisson Distribution)**.** Let $X_1, X_2, \ldots, X_n$ be i.i.d. Poisson(λ), where $\lambda > 0$ is the unknown rate parameter.

Step 1: PMF

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

Step 2: Likelihood

$$\begin{aligned} L(\lambda; \boldsymbol{x}) &= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\ &= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \end{aligned}$$

Step 3: Log-likelihood

Let $s = \sum_{i=1}^{n} x_i$. Then:

$$\ell(\lambda) = s \ln \lambda - n\lambda - \sum_{i=1}^{n} \ln(x_i!)$$

The last term is a constant (does not depend on λ).

Step 4: Differentiate

$$\frac{d\ell}{d\lambda} = \frac{s}{\lambda} - n$$

Step 5: Set to zero and solve

$$\begin{aligned} 0 &= \frac{s}{\hat{\lambda}} - n \\ n &= \frac{s}{\hat{\lambda}} \\ \hat{\lambda} &= \frac{s}{n} \end{aligned}$$

Therefore:

$$\boxed{\hat{\lambda}_{\text{MLE}} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}}$$

The MLE for the Poisson rate is the **sample mean**.

Step 6: Verify maximum

$$\frac{d^2\ell}{d\lambda^2} = -\frac{s}{\lambda^2} < 0$$

for $\lambda > 0$ and $s > 0$, confirming a maximum.

4.7.4 Example 4: Exponential Distribution

Example 4.11 (MLE for Exponential Distribution)**.** Let $X_1, X_2, \ldots, X_n$ be i.i.d. Exponential(λ), where $\lambda > 0$ is the rate parameter.

Step 1: PDF

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x \geq 0$$

Step 2: Likelihood

$$\begin{aligned} L(\lambda; \boldsymbol{x}) &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \\ &= \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \end{aligned}$$

Step 3: Log-likelihood

Let $s = \sum_{i=1}^{n} x_i$. Then:

$$\ell(\lambda) = n \ln \lambda - \lambda s$$

Step 4: Differentiate

$$\frac{d\ell}{d\lambda} = \frac{n}{\lambda} - s$$

Step 5: Set to zero and solve

$$\begin{aligned} 0 &= \frac{n}{\hat{\lambda}} - s \\ s &= \frac{n}{\hat{\lambda}} \\ \hat{\lambda} &= \frac{n}{s} \end{aligned}$$

Therefore:

$$\boxed{\hat{\lambda}_{\text{MLE}} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}}$$

The MLE for the exponential rate is the **reciprocal of the sample mean**.

Step 6: Verify maximum

$$\frac{d^2\ell}{d\lambda^2} = -\frac{n}{\lambda^2} < 0$$

for $\lambda > 0$, confirming a maximum.

Exponential MLE

The Exponential(λ) distribution has mean $\frac{1}{\lambda}$. The MLE sets the population mean equal to the sample mean:

$$\frac{1}{\hat{\lambda}} = \bar{x} \implies \hat{\lambda} = \frac{1}{\bar{x}}$$

This is an example of the *method of moments* coinciding with MLE.

4.7.5 Example 5: Normal Distribution (Mean)

Example 4.12 (MLE for Normal Mean (Known Variance))**.** Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$, where μ is unknown and σ^2 is known.

Step 1: PDF

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

Step 2: Likelihood

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

Step 3: Log-likelihood

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

Step 4: Differentiate

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

Step 5: Set to zero and solve

$$\begin{aligned} 0 &= \sum_{i=1}^{n}(x_i - \hat{\mu}) \\ 0 &= \sum_{i=1}^{n} x_i - n\hat{\mu} \\ n\hat{\mu} &= \sum_{i=1}^{n} x_i \end{aligned}$$

Therefore:

$$\boxed{\hat{\mu}_{\text{MLE}} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}}$$

The MLE for the normal mean is the **sample mean**.

Step 6: Verify maximum

$$\frac{d^2\ell}{d\mu^2} = -\frac{n}{\sigma^2} < 0$$

confirming a maximum.

4.8 Summary: MLE Pattern Recognition

Common MLE Results

Distribution	Parameter	MLE
Bernoulli(θ)	θ (success prob.)	$\bar{x}$ (sample proportion)
Binomial(m, θ)	θ (success prob.)	$\frac{\sum x_i}{mn}$ (total proportion)
Poisson(λ)	λ (rate)	$\bar{x}$ (sample mean)
Exponential(λ)	λ (rate)	$1/\bar{x}$ (reciprocal of sample mean)
Normal(μ, σ^2)	μ (mean, σ^2 known)	$\bar{x}$ (sample mean)

Why Sample Mean Appears So Often

For many distributions in the exponential family, the MLE takes a simple form involving the sample mean or sample proportion. This is not coincidental: for exponential family distributions, there exist *sufficient statistics* that capture all information about the parameter, and the MLE is typically a function of these sufficient statistics.

4.9 Properties of Maximum Likelihood Estimators

Maximum Likelihood Estimators have several desirable statistical properties that make them the default choice in many applications.

Properties of MLEs (Overview)

Under suitable regularity conditions, MLEs are:

1. **Consistent**: As $n \to \infty$, $\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta$ (converges in probability to the true value)
2. **Asymptotically efficient**: Among all consistent estimators, MLEs achieve the smallest asymptotic variance (the Cramér-Rao lower bound)
3. **Asymptotically normal**: $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})$ where $I(\theta)$ is the Fisher information
4. **Invariant**: If $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$ for any function g

These properties are studied in detail in advanced statistics courses. For now, the key takeaway is that MLE is a principled, well-behaved estimation method with strong theoretical foundations.

4.10 Practice Exercises

Differentiation

1. Find $\frac{d}{dx}[x^4 - 3x^2 + 5x - 7]$.
2. Find $\frac{d}{dx}[(2x + 1)^5]$ using the chain rule.
3. Find $\frac{d}{dx}\left[\frac{e^x}{x^2+1}\right]$ using the quotient rule.
4. Find $\frac{d}{dx}[\ln(x^2 + e^x)]$.
5. Find $\frac{d}{dx}[x^x]$ for $x > 0$ using logarithmic differentiation.

Maximum Likelihood Estimation

1. Given a sample $x_1, x_2, \ldots, x_n$ from a Geometric(θ) distribution with PMF $f(k;\theta) = \theta(1-\theta)^{k-1}$ for $k \in \{1, 2, 3, \ldots\}$, find the MLE of θ.

2. Given a sample from a Uniform$[0, \theta]$ distribution with PDF $f(x;\theta) = \frac{1}{\theta}$ for $x \in [0, \theta]$, find the MLE of θ. (Hint: The calculus approach fails here; think about which values of θ make the likelihood positive.)

3. For the Normal(μ, σ^2) distribution with both parameters unknown, derive the MLEs $\hat{\mu}$ and $\hat{\sigma}^2$.

4. Verify using the second derivative test that your answer to Exercise 1 is indeed a maximum.

Solutions

Solutions to these exercises are available in the appendix or from your instructor.

Chapter 5

Power Series and Integration

Learning Objectives

By the end of this chapter, you should be able to:

- Derive Taylor and Maclaurin series for common functions using the coefficient-matching approach
- Write out the first several terms of Taylor expansions for e^x, $\sin(x)$, $\cos(x)$, $\ln(1+x)$, and $(1-x)^{-1}$
- Determine the radius of convergence for a power series using the ratio test
- Compute indefinite and definite integrals using fundamental techniques
- Apply integration by substitution and integration by parts
- State and apply the Fundamental Theorem of Calculus
- Understand how power series connect to moment generating functions in probability theory

Prerequisites

This chapter assumes familiarity with:

- Differentiation rules: power, chain, product, and quotient rules (from Chapter 4)
- Exponential and logarithmic functions and their derivatives
- The concept of a limit and convergence (intuitive understanding suffices)
- Summation notation and factorial notation ($n!$)

5.1 Taylor Series: Motivation and Derivation

Many functions that arise in statistics and data science—exponentials, logarithms, trigonometric functions—are difficult to compute directly. Taylor series provide a powerful technique: represent these "ugly" functions as *infinite polynomials*, which are far easier to differentiate, integrate, and approximate.

Why Approximate with Polynomials?

Polynomials are the "nicest" functions we have:

- They can be evaluated using only addition and multiplication
- Their derivatives and integrals are other polynomials
- They are continuous and smooth everywhere

If we can approximate a complicated function by a polynomial, we inherit all these nice properties.

5.1.1 The Central Question

Suppose we have a function $f(x)$ that is difficult to work with directly. We want to find a polynomial $p(x)$ that closely approximates $f(x)$ near some point $x = a$. The question is: **what should the coefficients of this polynomial be?**

Assume $f(x)$ can be represented by a power series centred at a:

$$f(x) = \sum_{k=0}^{\infty} c_k(x-a)^k = c_0 + c_1(x-a) + c_2(x-a)^2 + c_3(x-a)^3 + \cdots \tag{5.1}$$

Our task is to determine the coefficients $c_0, c_1, c_2, \ldots$ in terms of the function f and its derivatives.

5.1.2 Deriving the Coefficients

The key insight is that we want the polynomial to *match* the function and all its derivatives at the expansion point $x = a$:

$$\begin{aligned} p(a) &= f(a) \\ p'(a) &= f'(a) \\ p''(a) &= f''(a) \\ p'''(a) &= f'''(a) \\ &\vdots \end{aligned} \tag{5.2}$$

Let us work through these conditions systematically to find each coefficient.

Finding c_0: The Zeroth Coefficient

Starting with the series:

$$p(x) = c_0 + c_1(x-a) + c_2(x-a)^2 + c_3(x-a)^3 + \cdots$$

Evaluating at $x = a$: all terms with $(x-a)$ vanish, leaving only c_0:

$$p(a) = c_0 + 0 + 0 + \cdots = c_0$$

Applying the matching condition $p(a) = f(a)$:

$$\boxed{c_0 = f(a)} \tag{5.3}$$

Finding c_1: The First Coefficient

Differentiating the series:

$$p'(x) = c_1 + 2c_2(x-a) + 3c_3(x-a)^2 + 4c_4(x-a)^3 + \cdots$$

Evaluating at $x = a$:

$$p'(a) = c_1 + 0 + 0 + \cdots = c_1$$

Applying $p'(a) = f'(a)$:

$$\boxed{c_1 = f'(a)} \tag{5.4}$$

Finding c_2: The Second Coefficient

Differentiating again:

$$p''(x) = 2c_2 + 6c_3(x-a) + 12c_4(x-a)^2 + \cdots$$

At $x = a$:

$$p''(a) = 2c_2$$

Applying $p''(a) = f''(a)$:

$$2c_2 = f''(a) \implies \boxed{c_2 = \frac{f''(a)}{2!}}$$

Finding c_3: The Third Coefficient

Differentiating once more:

$$p'''(x) = 6c_3 + 24c_4(x-a) + \cdots$$

At $x = a$:

$$p'''(a) = 6c_3$$

Applying $p'''(a) = f'''(a)$:

$$6c_3 = f'''(a) \implies \boxed{c_3 = \frac{f'''(a)}{3!}}$$

The General Pattern

A clear pattern emerges. The numerical coefficients $1, 2, 6, 24, \ldots$ are factorials: $0! = 1$, $1! = 1$, $2! = 2$, $3! = 6$, $4! = 24$. In general:

$$c_k = \frac{f^{(k)}(a)}{k!} \tag{5.5}$$

where $f^{(k)}(a)$ denotes the k-th derivative of f evaluated at a.

5.1.3 The Taylor Series Formula

Definition 5.1 (Taylor Series)**.** The **Taylor series** of a function $f(x)$ centred at $x = a$ is:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k \tag{5.6}$$

provided f has derivatives of all orders at a and the series converges to $f(x)$.

Taylor Series Expanded

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots$$

Definition 5.2 (Maclaurin Series)**.** A **Maclaurin series** is a Taylor series centred at $a = 0$:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!}x^k = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots \tag{5.7}$$

Definition 5.3 (Taylor Polynomial)**.** The **n-th Taylor polynomial** (or n-th partial sum) is the finite truncation:

$$p_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k \tag{5.8}$$

This provides an approximation to $f(x)$ near $x = a$, with accuracy improving as n increases.

5.2 Common Maclaurin Series

Several Maclaurin series appear repeatedly in mathematics, statistics, and data science. You should commit these to memory.

5.2.1 The Exponential Function

Example 5.1 (Maclaurin Series for e^x)**.** Find the Maclaurin series for $f(x) = e^x$.

Solution: The exponential function has a remarkable property: all its derivatives equal itself.

$$\begin{aligned} f(x) &= e^x & f(0) &= 1 \\ f'(x) &= e^x & f'(0) &= 1 \\ f''(x) &= e^x & f''(0) &= 1 \\ f^{(k)}(x) &= e^x & f^{(k)}(0) &= 1 \quad \text{for all } k \end{aligned}$$

Substituting into the Maclaurin formula:

$$e^x = \sum_{k=0}^{\infty} \frac{1}{k!}x^k = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

Exponential Series

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots \tag{5.9}$$

This series converges for all real x (radius of convergence $R = \infty$).

The Exponential Series in Statistics

The exponential series appears throughout probability theory:

- The Poisson PMF: $\frac{\lambda^k e^{-\lambda}}{k!}$ uses the exponential series
- Moment generating functions often involve e^{tx}
- The normal distribution's PDF contains $e^{-x^2/2}$

5.2.2 Trigonometric Functions

Example 5.2 (Maclaurin Series for $\sin(x)$)**.** Find the Maclaurin series for $f(x) = \sin(x)$.

Solution: The derivatives of sine cycle with period 4:

$$\begin{aligned} f(x) &= \sin(x) & f(0) &= 0 \\ f'(x) &= \cos(x) & f'(0) &= 1 \\ f''(x) &= -\sin(x) & f''(0) &= 0 \\ f'''(x) &= -\cos(x) & f'''(0) &= -1 \\ f^{(4)}(x) &= \sin(x) & f^{(4)}(0) &= 0 \end{aligned}$$

The pattern $0, 1, 0, -1, 0, 1, 0, -1, \ldots$ repeats. Only odd-powered terms survive:

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1}$$

Example 5.3 (Maclaurin Series for $\cos(x)$)**.** By similar reasoning (or by differentiating the sine series):

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k}$$

Trigonometric Series

$$\sin(x) = x - \frac{x^3}{6} + \frac{x^5}{120} - \cdots = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!} \tag{5.10}$$

$$\cos(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \cdots = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!} \tag{5.11}$$

Both series converge for all real x (radius of convergence $R = \infty$).

Remark. Notice that $\sin(x)$ has only odd powers (it's an odd function: $\sin(-x) = -\sin(x)$), while $\cos(x)$ has only even powers (it's an even function: $\cos(-x) = \cos(x)$).

5.2.3 The Natural Logarithm

Example 5.4 (Maclaurin Series for $\ln(1+x)$)**.** Find the Maclaurin series for $f(x) = \ln(1+x)$.

Solution: We compute successive derivatives:

$$\begin{aligned} f(x) &= \ln(1+x) & f(0) &= 0 \\ f'(x) &= \frac{1}{1+x} = (1+x)^{-1} & f'(0) &= 1 \\ f''(x) &= -(1+x)^{-2} & f''(0) &= -1 \\ f'''(x) &= 2(1+x)^{-3} & f'''(0) &= 2 \\ f^{(4)}(x) &= -6(1+x)^{-4} & f^{(4)}(0) &= -6 \end{aligned}$$

The pattern: $f^{(k)}(0) = (-1)^{k+1}(k-1)!$ for $k \geq 1$. Therefore:

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k$$

Logarithmic Series

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots = \sum_{k=1}^{\infty} \frac{(-1)^{k+1} x^k}{k} \tag{5.12}$$

This series converges for $-1 < x \leq 1$ (radius of convergence $R = 1$).

Limited Convergence

Unlike the exponential and trigonometric series, the logarithmic series only converges for $|x| \leq 1$ (and conditionally at $x = 1$). For $|x| > 1$, the series diverges. This is because $\ln(1+x)$ has a singularity at $x = -1$ (where it becomes $\ln(0) = -\infty$).

5.2.4 The Geometric Series

Theorem 5.1 (Geometric Series)**.** *For $|x| < 1$:*

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k = 1 + x + x^2 + x^3 + \cdots \tag{5.13}$$

Proof. Let $S_n = 1 + x + x^2 + \cdots + x^n$. Multiplying by x:

$$xS_n = x + x^2 + \cdots + x^{n+1}$$

Subtracting: $S_n - xS_n = 1 - x^{n+1}$, so $S_n = \frac{1-x^{n+1}}{1-x}$.

For $|x| < 1$, we have $x^{n+1} \to 0$ as $n \to \infty$, giving $S = \lim_{n\to\infty} S_n = \frac{1}{1-x}$. ■

Geometric Series Variants

By substitution and manipulation:

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \cdots = \sum_{k=0}^{\infty} (-1)^k x^k \quad (|x| < 1) \tag{5.14}$$

$$\frac{1}{(1-x)^2} = 1 + 2x + 3x^2 + 4x^3 + \cdots = \sum_{k=0}^{\infty} (k+1) x^k \quad (|x| < 1) \tag{5.15}$$

The second follows from differentiating the first with respect to x.

5.2.5 Taylor Series for $\ln(x)$ at $x = 1$

Example 5.5 (Taylor Series for $\ln(x)$ at $x = 1$)**.** Find the Taylor polynomials p_0, p_1, p_2, and p_3 for $f(x) = \ln(x)$ centred at $a = 1$.

Solution: We need the derivatives of $f(x)$ evaluated at $x = 1$:

$$\begin{aligned} f(x) &= \ln(x) & f(1) &= \ln(1) = 0 \\ f'(x) &= \frac{1}{x} & f'(1) &= 1 \\ f''(x) &= -\frac{1}{x^2} & f''(1) &= -1 \\ f'''(x) &= \frac{2}{x^3} & f'''(1) &= 2 \end{aligned}$$

Using the Taylor polynomial formula $p_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(1)}{k!}(x-1)^k$:

$$\begin{aligned} p_0(x) &= f(1) = 0 \\ p_1(x) &= f(1) + f'(1)(x-1) = 0 + 1 \cdot (x-1) = x - 1 \\ p_2(x) &= p_1(x) + \frac{f''(1)}{2!}(x-1)^2 = (x-1) + \frac{-1}{2}(x-1)^2 \\ &= (x-1) - \frac{1}{2}(x-1)^2 \\ p_3(x) &= p_2(x) + \frac{f'''(1)}{3!}(x-1)^3 = (x-1) - \frac{1}{2}(x-1)^2 + \frac{2}{6}(x-1)^3 \\ &= (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 \end{aligned}$$

Each polynomial provides a successively better approximation to $\ln(x)$ near $x = 1$.

5.2.6 Summary of Common Series

Common Maclaurin Series

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \qquad R = \infty$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots \qquad R = \infty$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots \qquad R = \infty$$

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots \qquad R = 1$$

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots \qquad R = 1$$

$$(1+x)^\alpha = 1 + \alpha x + \frac{\alpha(\alpha-1)}{2!}x^2 + \cdots \qquad R = 1 \text{ (binomial series)}$$

Here R denotes the radius of convergence.

5.3 Convergence of Power Series

Not every power series converges for all values of x. Understanding where a series converges is essential for correct application.

Definition 5.4 (Radius of Convergence)**.** For a power series $\sum_{k=0}^{\infty} c_k(x-a)^k$, the **radius of convergence** R is the number such that:

- The series **converges absolutely** for $|x-a| < R$
- The series **diverges** for $|x-a| > R$
- At $|x-a| = R$, convergence must be checked case by case

The interval $(a-R, a+R)$ is called the **interval of convergence**.

Theorem 5.2 (Ratio Test for Radius of Convergence)**.** *For the power series $\sum_{k=0}^{\infty} c_k(x-a)^k$, the radius of convergence is:*

$$R = \lim_{k\to\infty} \left| \frac{c_k}{c_{k+1}} \right| \tag{5.16}$$

provided this limit exists (it may be 0 or ∞).

Example 5.6 (Finding the Radius of Convergence)**.** Find the radius of convergence for $\sum_{k=1}^{\infty} \frac{x^k}{k}$ (the series for $-\ln(1-x)$).

Solution: Here $c_k = \frac{1}{k}$. Applying the ratio test:

$$R = \lim_{k\to\infty} \left| \frac{1/k}{1/(k+1)} \right| = \lim_{k\to\infty} \frac{k+1}{k} = \lim_{k\to\infty} \left(1 + \frac{1}{k}\right) = 1$$

The series converges for $|x| < 1$ and diverges for $|x| > 1$. At $x = 1$, it becomes $\sum \frac{1}{k}$ (harmonic series), which diverges. At $x = -1$, it becomes $\sum \frac{(-1)^k}{k}$, which converges conditionally.

Convergence vs. Equality

A series may converge but not equal the function it was derived from! A function f equals its Taylor series on an interval only if the *remainder term* goes to zero. For most functions encountered in applications (analytic functions), this holds within the radius of convergence, but pathological examples exist.

5.4 Integration: Foundations

Integration is the mathematical operation that "undoes" differentiation. Where differentiation measures rates of change, integration measures accumulation—areas under curves, total quantities, cumulative probabilities.

Why Integration Matters for Data Science

Integration is essential for:

- **Probability**: Computing probabilities from PDFs via $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,dx$
- **Expected values**: $\mathbb{E}[X] = \int x f(x)\,dx$
- **Normalisation**: Ensuring PDFs integrate to 1
- **Cumulative distributions**: $F(x) = \int_{-\infty}^{x} f(t)\,dt$

5.4.1 Antiderivatives and Indefinite Integrals

Definition 5.5 (Antiderivative)**.** A function $F(x)$ is an **antiderivative** of $f(x)$ if:

$$F'(x) = f(x)$$

That is, the derivative of F equals f.

Example 5.7 (Finding Antiderivatives)**.**

- $F(x) = x^3$ is an antiderivative of $f(x) = 3x^2$ because $(x^3)' = 3x^2$
- $F(x) = \sin(x)$ is an antiderivative of $f(x) = \cos(x)$ because $(\sin x)' = \cos x$
- $F(x) = e^x$ is an antiderivative of $f(x) = e^x$ because $(e^x)' = e^x$

Theorem 5.3 (Antiderivatives Differ by a Constant)**.** *If $F(x)$ is an antiderivative of $f(x)$, then so is $F(x) + C$ for any constant C. Moreover,* every *antiderivative of f has the form $F(x) + C$.*

Proof. If $F'(x) = f(x)$, then $(F(x) + C)' = F'(x) + 0 = f(x)$.

Conversely, if G is any antiderivative of f, then $(G - F)' = G' - F' = f - f = 0$. A function with zero derivative everywhere is constant, so $G(x) - F(x) = C$ for some constant C. ∎

Definition 5.6 (Indefinite Integral)**.** The **indefinite integral** of $f(x)$ is the general antiderivative:

$$\int f(x)\,dx = F(x) + C \tag{5.17}$$

where $F'(x) = f(x)$ and C is an arbitrary constant of integration.

Don't Forget the Constant

When computing indefinite integrals, always include the constant C. Forgetting it is a common error that can lead to incorrect results, especially when solving differential equations or computing definite integrals via antiderivatives.

5.4.2 Basic Integration Rules

Integration rules are derived by reversing differentiation rules.

Power Rule for Integration

For $n \neq -1$:

$$\int x^n \, dx = \frac{x^{n+1}}{n+1} + C \tag{5.18}$$

Proof. Check by differentiating: $\frac{d}{dx}\left[\frac{x^{n+1}}{n+1}\right] = \frac{(n+1)x^n}{n+1} = x^n$. ■

Constant Multiple and Sum Rules

$$\int k \cdot f(x) \, dx = k \int f(x) \, dx \tag{5.19}$$

$$\int [f(x) \pm g(x)] \, dx = \int f(x) \, dx \pm \int g(x) \, dx \tag{5.20}$$

Integration, like differentiation, is a **linear operator**.

Table of Common Integrals

$$\int x^n \, dx = \frac{x^{n+1}}{n+1} + C \quad (n \neq -1)$$

$$\int \frac{1}{x} \, dx = \ln|x| + C$$

$$\int e^x \, dx = e^x + C$$

$$\int a^x \, dx = \frac{a^x}{\ln a} + C \quad (a > 0, a \neq 1)$$

$$\int \cos(x) \, dx = \sin(x) + C$$

$$\int \sin(x) \, dx = -\cos(x) + C$$

$$\int \sec^2(x) \, dx = \tan(x) + C$$

$$\int \frac{1}{1+x^2} \, dx = \arctan(x) + C$$

$$\int \frac{1}{\sqrt{1-x^2}} \, dx = \arcsin(x) + C$$

5.5 The Fundamental Theorem of Calculus

The Fundamental Theorem of Calculus (FTC) is one of the most important results in mathematics. It establishes the deep connection between differentiation and integration—they are inverse operations.

Theorem 5.4 (Fundamental Theorem of Calculus, Part I). *If f is continuous on $[a, b]$ and we define:*

$$F(x) = \int_a^x f(t) \, dt \tag{5.21}$$

then F is differentiable on (a, b) and:

$$F'(x) = f(x) \tag{5.22}$$

In other words, the derivative of an integral with respect to its upper limit equals the integrand evaluated at that limit.

FTC Part I: Differentiation Undoes Integration

The integral $\int_a^x f(t) \, dt$ represents the "accumulated area" under f from a to x. How fast is this area growing as x increases? At rate $f(x)$—precisely the height of the function at that point.

Theorem 5.5 (Fundamental Theorem of Calculus, Part II). *If f is continuous on $[a, b]$ and F is any antiderivative of f (i.e., $F' = f$), then:*

$$\int_a^b f(x)\,dx = F(b) - F(a) \tag{5.23}$$

Evaluating Definite Integrals

To compute $\int_a^b f(x)\,dx$:

1. Find any antiderivative $F(x)$ of $f(x)$
2. Evaluate $F(b) - F(a)$

Notation: $\int_a^b f(x)\,dx = \big[F(x)\big]_a^b = F(b) - F(a)$

Example 5.8 (Using the FTC). Compute $\int_1^3 x^2\,dx$.

Solution: An antiderivative of x^2 is $F(x) = \frac{x^3}{3}$. By FTC Part II:

$$\int_1^3 x^2\,dx = \left[\frac{x^3}{3}\right]_1^3 = \frac{3^3}{3} - \frac{1^3}{3} = \frac{27}{3} - \frac{1}{3} = \frac{26}{3}$$

Example 5.9 (Computing a Probability). If a continuous random variable X has PDF $f(x) = 2x$ for $x \in [0, 1]$, find $\mathbb{P}(0.5 \leq X \leq 0.8)$.

Solution:

$$\mathbb{P}(0.5 \leq X \leq 0.8) = \int_{0.5}^{0.8} 2x\,dx = \big[x^2\big]_{0.5}^{0.8} = 0.64 - 0.25 = 0.39$$

5.6 Integration Techniques

Not all integrals can be computed directly from the basic rules. Two fundamental techniques extend our capabilities significantly.

5.6.1 Integration by Substitution

Substitution is the integration counterpart to the chain rule for differentiation.

Theorem 5.6 (Substitution Rule). *If $u = g(x)$ is a differentiable function and f is continuous, then:*

$$\int f(g(x)) \cdot g'(x)\,dx = \int f(u)\,du \tag{5.24}$$

where after integration we substitute back $u = g(x)$.

Why Substitution Works

If $F'(u) = f(u)$, then by the chain rule:

$$\frac{d}{dx}[F(g(x))] = F'(g(x)) \cdot g'(x) = f(g(x)) \cdot g'(x)$$

So $F(g(x))$ is an antiderivative of $f(g(x)) \cdot g'(x)$.

Substitution Method

1. Choose a substitution $u = g(x)$
2. Compute $du = g'(x)\,dx$, so $dx = \frac{du}{g'(x)}$
3. Rewrite the integral entirely in terms of u
4. Integrate with respect to u
5. Substitute back to express the answer in terms of x

Example 5.10 (Substitution: Polynomial Inside)**.** Compute $\int (2x+3)^5\,dx$.

Solution: Let $u = 2x+3$. Then $\frac{du}{dx} = 2$, so $dx = \frac{du}{2}$.

$$\begin{aligned}\int (2x+3)^5\,dx &= \int u^5 \cdot \frac{du}{2} = \frac{1}{2}\int u^5\,du \\ &= \frac{1}{2}\cdot\frac{u^6}{6} + C = \frac{u^6}{12} + C \\ &= \frac{(2x+3)^6}{12} + C\end{aligned}$$

Example 5.11 (Substitution: Exponential)**.** Compute $\int xe^{x^2}\,dx$.

Solution: Let $u = x^2$. Then $du = 2x\,dx$, so $x\,dx = \frac{du}{2}$.

$$\begin{aligned}\int xe^{x^2}\,dx &= \int e^u \cdot \frac{du}{2} = \frac{1}{2}\int e^u\,du \\ &= \frac{1}{2}e^u + C = \frac{1}{2}e^{x^2} + C\end{aligned}$$

Example 5.12 (Substitution: Logarithmic)**.** Compute $\int \frac{\ln(x)}{x}\,dx$.

Solution: Let $u = \ln(x)$. Then $du = \frac{1}{x}\,dx$.

$$\int \frac{\ln(x)}{x}\,dx = \int u\,du = \frac{u^2}{2} + C = \frac{(\ln x)^2}{2} + C$$

Theorem 5.7 (Substitution for Definite Integrals)**.** *When using substitution for definite integrals, change the limits of integration along with the variable:*

$$\int_a^b f(g(x))\cdot g'(x)\,dx = \int_{g(a)}^{g(b)} f(u)\,du \tag{5.25}$$

Example 5.13 (Definite Integral by Substitution)**.** Compute $\int_0^2 x(x^2+1)^3\,dx$.

Solution: Let $u = x^2+1$. Then $du = 2x\,dx$, so $x\,dx = \frac{du}{2}$.

When $x = 0$: $u = 0^2+1 = 1$. When $x = 2$: $u = 2^2+1 = 5$.

$$\begin{aligned}\int_0^2 x(x^2+1)^3\,dx &= \int_1^5 u^3\cdot\frac{du}{2} = \frac{1}{2}\left[\frac{u^4}{4}\right]_1^5 \\ &= \frac{1}{8}\left[5^4 - 1^4\right] = \frac{1}{8}(625-1) = \frac{624}{8} = 78\end{aligned}$$

5.6.2 Integration by Parts

Integration by parts is the integration counterpart to the product rule.

Theorem 5.8 (Integration by Parts)**.** *If u and v are differentiable functions, then:*

$$\int u\,dv = uv - \int v\,du \tag{5.26}$$

Proof. From the product rule: $(uv)' = u'v + uv'$. Integrating both sides:

$$uv = \int u'v\,dx + \int uv'\,dx$$

Rearranging: $\int uv'\,dx = uv - \int u'v\,dx$.

Writing $dv = v'\,dx$ and $du = u'\,dx$ gives the formula. ■

Choosing u and dv: LIATE Rule

When applying integration by parts, choose u and dv using the **LIATE** priority (choose u from earlier in the list):

1. **L**ogarithmic functions: $\ln(x)$, $\log(x)$
2. **I**nverse trigonometric functions: $\arctan(x)$, $\arcsin(x)$
3. **A**lgebraic functions: x^n, polynomials
4. **T**rigonometric functions: $\sin(x)$, $\cos(x)$
5. **E**xponential functions: e^x, a^x

Example 5.14 (Integration by Parts: $\int xe^x\,dx$)**.** Compute $\int xe^x\,dx$.

Solution: Using LIATE, let $u = x$ (algebraic) and $dv = e^x\,dx$ (exponential).

Then $du = dx$ and $v = e^x$.

Applying the formula:

$$\begin{aligned}\int xe^x\,dx &= uv - \int v\,du = xe^x - \int e^x\,dx \\ &= xe^x - e^x + C = e^x(x-1) + C\end{aligned}$$

Example 5.15 (Integration by Parts: $\int \ln(x)\,dx$)**.** Compute $\int \ln(x)\,dx$.

Solution: Let $u = \ln(x)$ and $dv = dx$.

Then $du = \frac{1}{x}\,dx$ and $v = x$.

$$\begin{aligned}\int \ln(x)\,dx &= x\ln(x) - \int x\cdot\frac{1}{x}\,dx = x\ln(x) - \int 1\,dx \\ &= x\ln(x) - x + C = x(\ln(x) - 1) + C\end{aligned}$$

Example 5.16 (Integration by Parts: $\int x^2 e^x\,dx$)**.** Compute $\int x^2 e^x\,dx$.

Solution: This requires applying integration by parts twice.

First application: Let $u = x^2$, $dv = e^x\,dx$. Then $du = 2x\,dx$, $v = e^x$.

$$\int x^2 e^x\,dx = x^2 e^x - \int 2x e^x\,dx = x^2 e^x - 2\int x e^x\,dx$$

Second application: We computed $\int x e^x\,dx = e^x(x-1) + C$ in the previous example.

$$\int x^2 e^x\,dx = x^2 e^x - 2e^x(x-1) + C = e^x(x^2 - 2x + 2) + C$$

5.7 Applications to Probability: Moment Generating Functions

Power series and integration combine beautifully in the theory of **moment generating functions** (MGFs), a powerful tool in probability theory.

Definition 5.7 (Moment Generating Function)**.** The **moment generating function** of a random variable X is:

$$M_X(t) = \mathbb{E}[e^{tX}] \tag{5.27}$$

provided this expectation exists for t in some neighbourhood of zero.

Why "Moment Generating"?

The name comes from how MGFs encode all moments of a distribution. Using the Maclaurin series for e^{tX}:

$$e^{tX} = 1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \cdots$$

Taking expectations:

$$M_X(t) = \mathbb{E}[e^{tX}] = 1 + t\,\mathbb{E}[X] + \frac{t^2}{2!}\,\mathbb{E}[X^2] + \frac{t^3}{3!}\,\mathbb{E}[X^3] + \cdots$$

The coefficients are the moments $\mathbb{E}[X^k]$ divided by $k!$. Differentiating and setting $t = 0$ extracts individual moments.

Theorem 5.9 (Moments from the MGF)**.** *If the MGF $M_X(t)$ exists in a neighbourhood of zero, then:*

$$\mathbb{E}[X^n] = M_X^{(n)}(0) = \left.\frac{d^n M_X}{dt^n}\right|_{t=0} \tag{5.28}$$

The n-th moment equals the n-th derivative of the MGF evaluated at $t = 0$.

Example 5.17 (MGF of the Exponential Distribution)**.** Find the MGF of $X \sim \text{Exponential}(\lambda)$ and use it to find $\mathbb{E}[X]$.

Solution: The PDF is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tX}] = \int_0^\infty e^{tx} \cdot \lambda e^{-\lambda x}\,dx \\ &= \lambda \int_0^\infty e^{(t-\lambda)x}\,dx \end{aligned}$$

For $t < \lambda$, this integral converges:

$$M_X(t) = \lambda \left[\frac{e^{(t-\lambda)x}}{t-\lambda} \right]_0^{\infty} = \lambda \cdot \frac{0-1}{t-\lambda} = \frac{\lambda}{\lambda - t}$$

To find $\mathbb{E}[X]$, differentiate:

$$M_X'(t) = \frac{d}{dt}\left[\frac{\lambda}{\lambda - t}\right] = \frac{\lambda}{(\lambda - t)^2}$$

Evaluating at $t = 0$:

$$\mathbb{E}[X] = M_X'(0) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

This confirms the well-known result that an Exponential(λ) random variable has mean $1/\lambda$.

Remark. MGFs are studied in detail in Chapter 7. The key connection to this chapter is that the Taylor series expansion of e^{tX} underlies the entire theory. This is a beautiful example of how power series provide both computational techniques and theoretical insight.

5.8 Practice Exercises

Taylor and Maclaurin Series

1. Find the Maclaurin series for $f(x) = e^{-x}$ by substituting $-x$ into the series for e^x.
2. Find the first four nonzero terms of the Maclaurin series for $f(x) = e^x \sin(x)$ by multiplying the series for e^x and $\sin(x)$.
3. Find the Taylor series for $f(x) = \sqrt{x}$ centred at $a = 4$, up to the $(x-4)^2$ term.
4. Use the geometric series to find a power series representation for $\frac{1}{1+x^2}$. What is its radius of convergence?
5. Verify that $\frac{d}{dx}[\sin(x)] = \cos(x)$ by differentiating the Maclaurin series for $\sin(x)$ term by term.

Integration

1. Compute $\int (3x^4 - 2x^2 + 5)\,dx$.
2. Compute $\int_0^1 e^{2x}\,dx$.
3. Use substitution to compute $\int \frac{x}{x^2+1}\,dx$.
4. Use substitution to compute $\int_0^{\pi/2} \sin^3(x)\cos(x)\,dx$.
5. Use integration by parts to compute $\int x\cos(x)\,dx$.
6. Use integration by parts to compute $\int x^2 \ln(x)\,dx$.
7. Compute $\int_0^1 xe^{-x}\,dx$ (this appears in computing $\mathbb{E}[X]$ for certain distributions).

Applications

1. If X has PDF $f(x) = 3x^2$ for $x \in [0, 1]$, compute $\mathbb{E}[X] = \int_0^1 x \cdot 3x^2 \, dx$.
2. The Gamma function is defined by $\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} \, dx$. Use integration by parts to show that $\Gamma(n) = (n-1) \cdot \Gamma(n-1)$ for $n > 1$.
3. Use the Maclaurin series for e^x to find a series expansion for the MGF of a Poisson(λ) random variable, $M_X(t) = e^{\lambda(e^t - 1)}$.

Solutions

Solutions to these exercises are available in the appendix or from your instructor.

Chapter 6

Continuous Random Variables I

Learning Objectives

By the end of this chapter, you should be able to:

- Distinguish between discrete and continuous random variables based on properties of the CDF
- State the relationship between PDF and CDF, and derive one from the other
- Compute probabilities for continuous random variables using integration
- Verify whether a function is a valid PDF
- Compute the expected value and variance of continuous random variables
- Work with the Uniform, Normal, and Exponential distributions—deriving their properties from first principles
- Apply standardisation to transform any Normal random variable to the standard Normal
- Understand and apply the memoryless property of the Exponential distribution
- Extend probability concepts (joint distributions, marginals, conditionals, Bayes' rule) to continuous random variables

Prerequisites

This chapter assumes familiarity with:

- Discrete random variables, PMFs, and CDFs (from **????**)
- Expected value and variance for discrete random variables
- Basic differentiation and integration techniques (from Chapters 4 and 5)
- The Geometric and Poisson distributions (conceptually, for comparison)

6.1 From Discrete to Continuous: The Fundamental Distinction

In earlier chapters, we studied discrete random variables—those taking values in a countable set (integers, for instance). The probability mass function (PMF) assigned positive probability to each possible outcome. However, many phenomena in nature and data science involve quantities that can take *any* value in an interval: heights, waiting times, temperatures, stock prices.

Definition 6.1 (Continuous Random Variable)**.** A random variable X is **continuous** if its cumulative distribution function (CDF) $F_X(x) = P(X \leq x)$ is a continuous function that can be expressed as

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

for some non-negative function f_X, called the **probability density function** (PDF).

The key distinction lies in the CDF's behaviour:

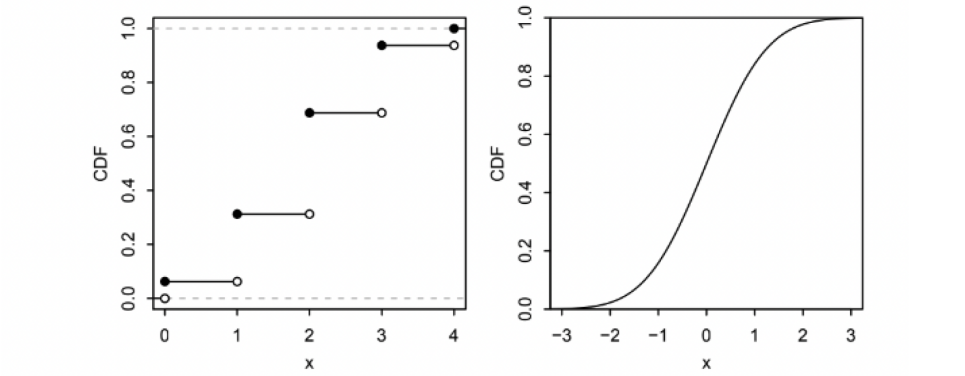


Figure 6.1: Comparison of CDFs: **Left**: CDF of a discrete random variable showing characteristic "staircase" jumps at each point mass. **Right**: CDF of a continuous random variable, which is smooth and differentiable.

Discrete vs Continuous: Key Differences

Property	Discrete	Continuous
CDF	Step function (jumps)	Smooth, differentiable
Probability function	PMF: $P(X = x)$	PDF: $f(x)$
$P(X = x)$	Can be positive	Always zero
$P(a < X \leq b)$	$\sum_{x \in (a,b]} P(X = x)$	$\int_a^b f(x)\,dx$
Total probability	$\sum_x P(X = x) = 1$	$\int_{-\infty}^{\infty} f(x)\,dx = 1$

6.2 The PDF–CDF Relationship

The relationship between the PDF and CDF is perhaps the most fundamental concept for continuous random variables. It mirrors the relationship between derivatives and integrals in calculus.

6.2.1 From CDF to PDF: Differentiation

Theorem 6.1 (PDF as Derivative of CDF). *If X is a continuous random variable with CDF $F_X(x)$ that is differentiable, then the PDF is given by*

$$f_X(x) = \frac{d}{dx} F_X(x) = F_X'(x) \tag{6.1}$$

at all points where the derivative exists.

Proof of Theorem 6.1

By the definition of the CDF and the Fundamental Theorem of Calculus:

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

Taking the derivative with respect to x:

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

The last equality follows from the Fundamental Theorem of Calculus, Part I.

6.2.2 From PDF to CDF: Integration

Theorem 6.2 (CDF as Integral of PDF). *If X is a continuous random variable with PDF $f_X(x)$, then the CDF is given by*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt = P(X \leq x) \tag{6.2}$$

The PDF–CDF Relationship

Think of the PDF as describing the "density" of probability—how concentrated probability is at different values. The CDF accumulates this density from $-\infty$ up to x.
Analogy: If the PDF represents the rate at which water flows into a tank at different times, the CDF represents the total amount of water in the tank up to time x.

6.2.3 Computing Probabilities

For continuous random variables, we compute probabilities over *intervals*, not individual points.

Theorem 6.3 (Probability Over an Interval). *For a continuous random variable X with PDF f_X and CDF F_X:*

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_{a}^{b} f_X(x)\, dx \tag{6.3}$$

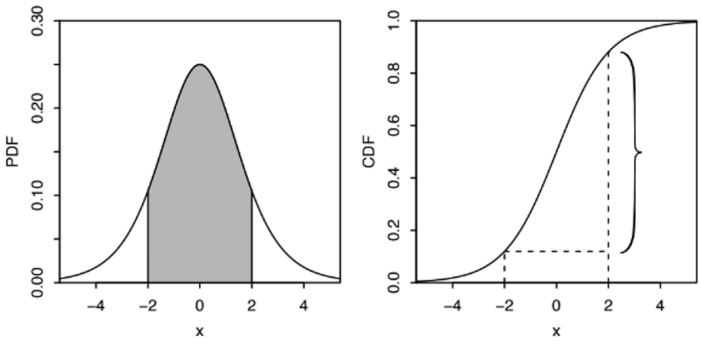


FIGURE 5.2
Logistic PDF and CDF. The probability $P(-2 < X < 2)$ is indicated by the shaded area under the PDF and the height of the curly brace on the CDF.

Figure 6.2: The probability $P(a < X \leq b)$ equals the shaded area under the PDF curve between a and b. This geometric interpretation—probability as area—is fundamental to working with continuous distributions.

Point Probabilities Are Zero

For any continuous random variable X and any specific value x:

$$P(X = x) = 0$$

This follows because:

$$P(X = x) = \lim_{\varepsilon \to 0} P(x - \varepsilon < X \leq x) = \lim_{\varepsilon \to 0} \int_{x-\varepsilon}^{x} f_X(t)\, dt = 0$$

The integral over a set of measure zero vanishes.
Consequence: For continuous random variables, the following are all equal:

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

Including or excluding endpoints does not change the probability.

6.2.4 Interpreting the PDF

A common misconception is that the PDF value $f(x)$ represents a probability. It does not.

What the PDF Value Means

The PDF $f(x)$ is a **probability density**, not a probability. Its interpretation:

- $f(x)$ can exceed 1 (unlike probabilities)
- Regions where $f(x)$ is larger have higher probability density—the random variable is *more likely* to fall near these values
- Only the *integral* of $f(x)$ over an interval gives a probability
- The PDF is normalised so that the total area under the curve equals 1

Heuristic interpretation: For small $\varepsilon > 0$,

$$P(x < X \leq x + \varepsilon) \approx f(x) \cdot \varepsilon$$

The PDF tells us probability per unit length near x.

6.2.5 Valid PDFs

Definition 6.2 (Valid PDF)**.** A function $f : \mathbb{R} \to \mathbb{R}$ is a valid probability density function if and only if:

1. **Non-negativity**: $f(x) \geq 0$ for all $x \in \mathbb{R}$
2. **Normalisation**: $\displaystyle\int_{-\infty}^{\infty} f(x)\,dx = 1$

The **support** of a continuous random variable is the set of values where $f(x) > 0$. Outside the support, $f(x) = 0$.

Example 6.1 (Verifying a Valid PDF)**.** Consider $f(x) = 2x$ for $x \in [0, 1]$ and $f(x) = 0$ otherwise. Is this a valid PDF?

Check non-negativity: For $x \in [0, 1]$, we have $2x \geq 0$. ✓

Check normalisation:

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^1 2x\,dx = \left[x^2\right]_0^1 = 1 - 0 = 1 \quad ✓$$

Yes, $f(x) = 2x$ on $[0, 1]$ is a valid PDF.

6.3 Expected Value of Continuous Random Variables

The expected value (or mean) of a continuous random variable generalises the discrete case by replacing summation with integration.

Definition 6.3 (Expected Value)**.** For a continuous random variable X with PDF $f_X(x)$, the **expected value** (or **mean**) is:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx \tag{6.4}$$

provided the integral converges absolutely.

Expected Value as Centre of Mass

Imagine the PDF as describing how mass is distributed along a beam:

- $f(x)$ represents the density of mass at position x
- $\mathbb{E}[X]$ is the centre of mass—the balance point of the distribution
- The integral $\int x f(x)\,dx$ computes the weighted average position, where each position

x is weighted by its density $f(x)$

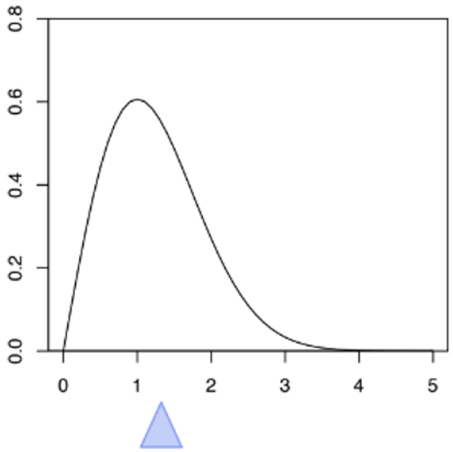


FIGURE 5.5
The expected value of a continuous r.v. is the balancing point of the PDF.

Figure 6.3: The expected value $\mathbb{E}[X]$ as the balance point (centre of mass) of the probability distribution. If the PDF were a physical object with density proportional to $f(x)$, it would balance at $\mathbb{E}[X]$.

More generally, we can compute the expected value of any function of X:

Theorem 6.4 (Law of the Unconscious Statistician (LOTUS))**.** *For a continuous random variable X with PDF $f_X(x)$ and any function g:*

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\, dx \tag{6.5}$$

This theorem is remarkably useful: to find $\mathbb{E}[g(X)]$, we do not need to first derive the distribution of $g(X)$.

6.3.1 Variance of Continuous Random Variables

Definition 6.4 (Variance)**.** The **variance** of a continuous random variable X is:

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x)\, dx \tag{6.6}$$

where $\mu = \mathbb{E}[X]$. The **standard deviation** is $\sigma = \sqrt{\text{Var}(X)}$.

Computational Formula for Variance

A more convenient formula for computing variance:

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \tag{6.7}$$

This is often easier because it requires computing only $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$, both of which follow directly from LOTUS.

6.4 The Uniform Distribution

The Uniform distribution is the simplest continuous distribution: all values in an interval are equally likely.

Definition 6.5 (Continuous Uniform Distribution)**.** A continuous random variable X has the **Uniform distribution** on the interval (a, b), written $X \sim \text{Unif}(a, b)$, if its PDF is:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases} \tag{6.8}$$

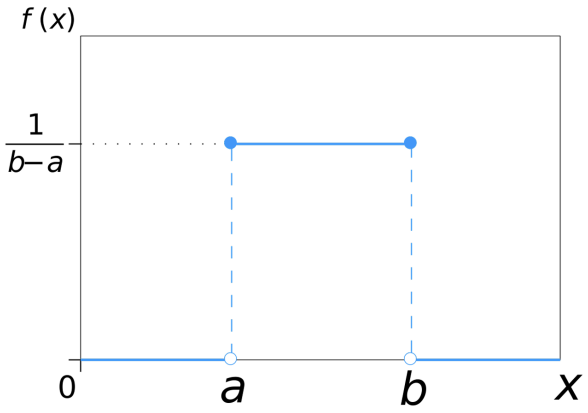


Figure 6.4: PDF of the Uniform distribution on (a, b). The constant height $\frac{1}{b-a}$ ensures that the total area (a rectangle) equals 1.

Remark. The PDF does not depend on x within the support—this is what makes the distribution "uniform". Every infinitesimal interval within (a, b) has the same probability density.

6.4.1 Verifying the Uniform PDF

Verification that Uniform PDF is Valid

Non-negativity: $\frac{1}{b-a} > 0$ since $b > a$. ✓

Normalisation: The integral over the support forms a rectangle:

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \int_a^b \frac{1}{b-a}\,dx = \frac{1}{b-a} \cdot (b-a) = 1 \quad ✓$$

6.4.2 CDF of the Uniform Distribution

Theorem 6.5 (Uniform CDF). *For $X \sim$ Unif(a, b), the CDF is:*

$$F_X(x) = \begin{cases} 0 & \text{if } x \le a \\ \dfrac{x-a}{b-a} & \text{if } a < x < b \\ 1 & \text{if } x \ge b \end{cases} \tag{6.9}$$

Derivation of Uniform CDF

For $x \leq a$: No probability mass exists below a, so $F_X(x) = 0$.

For $a < x < b$: Integrate the PDF from $-\infty$ to x:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt = \int_{a}^{x} \frac{1}{b-a}\,dt = \frac{1}{b-a} \cdot (x-a) = \frac{x-a}{b-a}$$

For $x \geq b$: All probability mass has been accumulated, so $F_X(x) = 1$.

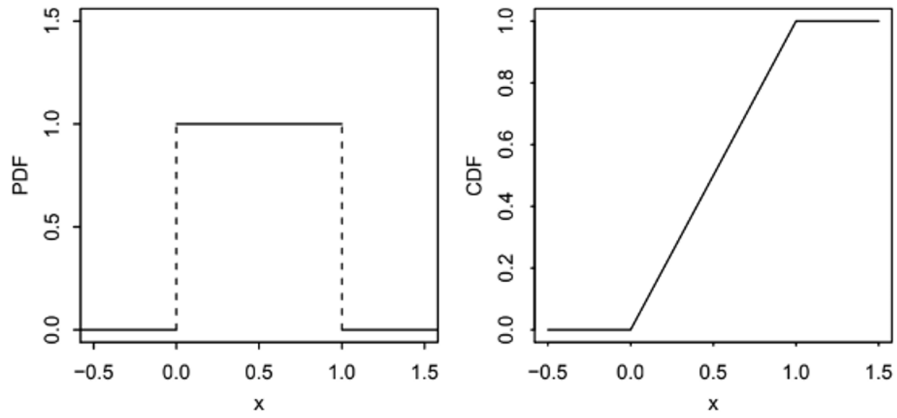


Figure 6.5: CDF of the Uniform distribution. Note the linear increase between a and b, reflecting the constant density of the PDF.

6.4.3 Mean of the Uniform Distribution

Theorem 6.6 (Mean of Uniform Distribution). *For $X \sim Unif(a, b)$:*

$$\mathbb{E}[X] = \frac{a+b}{2} \tag{6.10}$$

This result is intuitive: the mean is the midpoint of the interval.

Derivation of Uniform Mean

Applying the definition of expected value:

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx \\
&= \int_a^b x \cdot \frac{1}{b-a}\, dx \\
&= \frac{1}{b-a} \int_a^b x\, dx \\
&= \frac{1}{b-a} \left[\frac{x^2}{2}\right]_a^b \\
&= \frac{1}{b-a} \cdot \frac{b^2-a^2}{2} \\
&= \frac{1}{b-a} \cdot \frac{(b+a)(b-a)}{2} \quad \text{(difference of squares)} \\
&= \frac{a+b}{2}
\end{aligned}$$

6.4.4 Variance of the Uniform Distribution

Theorem 6.7 (Variance of Uniform Distribution). *For $X \sim Unif(a, b)$:*

$$\text{Var}(X) = \frac{(b-a)^2}{12} \tag{6.11}$$

Derivation of Uniform Variance

Using the computational formula $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$:

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

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_a^b x^2 \cdot \frac{1}{b-a}\, dx \\
&= \frac{1}{b-a} \left[\frac{x^3}{3}\right]_a^b \\
&= \frac{1}{b-a} \cdot \frac{b^3-a^3}{3} \\
&= \frac{b^3-a^3}{3(b-a)}
\end{aligned}$$

Using the factorisation $b^3 - a^3 = (b-a)(b^2+ab+a^2)$:

$$\mathbb{E}[X^2] = \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} = \frac{b^2+ab+a^2}{3}$$

Step 2: Compute $(\mathbb{E}[X])^2$:

$$(\mathbb{E}[X])^2 = \left(\frac{a+b}{2}\right)^2 = \frac{a^2+2ab+b^2}{4}$$

Step 3: Subtract:

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \frac{b^2+ab+a^2}{3} - \frac{a^2+2ab+b^2}{4} \\ &= \frac{4(b^2+ab+a^2) - 3(a^2+2ab+b^2)}{12} \\ &= \frac{4b^2+4ab+4a^2-3a^2-6ab-3b^2}{12} \\ &= \frac{b^2-2ab+a^2}{12} \\ &= \frac{(b-a)^2}{12}\end{aligned}$$

Uniform Distribution Summary

For $X \sim \text{Unif}(a,b)$:

$$\text{PDF}: \quad f(x) = \frac{1}{b-a} \quad \text{for } a < x < b$$

$$\text{CDF}: \quad F(x) = \frac{x-a}{b-a} \quad \text{for } a < x < b$$

$$\text{Mean}: \quad \mathbb{E}[X] = \frac{a+b}{2}$$

$$\text{Variance}: \quad \text{Var}(X) = \frac{(b-a)^2}{12}$$

6.5 The Normal Distribution

The Normal (or Gaussian) distribution is arguably the most important distribution in statistics. Its ubiquity stems from the Central Limit Theorem: sums and averages of many independent random variables tend towards normality, regardless of the original distribution.

Definition 6.6 (Normal Distribution)**.** A continuous random variable X has the **Normal distribution** with mean μ and variance σ^2, written $X \sim \mathcal{N}(\mu, \sigma^2)$, if its PDF is:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R} \tag{6.12}$$

Let us unpack each component of this formula:

- μ: The **mean** (or **location parameter**)—determines where the distribution is centred
- σ^2: The **variance** (with σ being the standard deviation)—determines the spread
- $\frac{1}{\sigma\sqrt{2\pi}}$: The **normalising constant**—ensures the PDF integrates to 1
- $\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$: The exponential of a negative quadratic—creates the characteristic "bell curve" shape

Remark. The support of the Normal distribution is the entire real line $(-\infty, \infty)$. Although the tails decay rapidly, technically any real value is possible.

6.5.1 The Standard Normal Distribution

Definition 6.7 (Standard Normal Distribution)**.** The **Standard Normal distribution** is the Normal distribution with $\mu = 0$ and $\sigma^2 = 1$, written $Z \sim \mathcal{N}(0, 1)$. Its PDF is:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad z \in \mathbb{R} \tag{6.13}$$

and its CDF is denoted $\Phi(z)$.

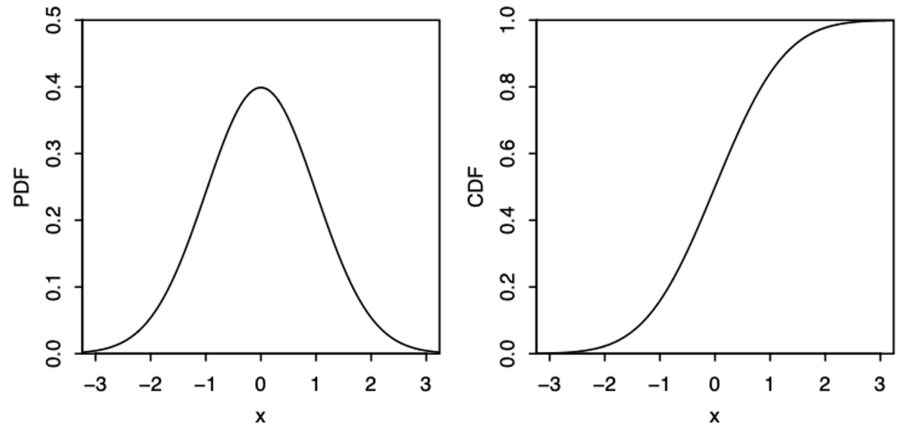


Figure 6.6: The PDF $\phi(z)$ and CDF $\Phi(z)$ of the standard Normal distribution. The PDF is symmetric about zero; the CDF is an S-shaped curve passing through $(0, 0.5)$.

Remark. The CDF $\Phi(z)$ has no closed-form expression—it must be computed numerically or looked up in tables. This is why the standard Normal is so important: we can transform any Normal random variable to the standard Normal and use pre-computed tables.

6.5.2 Parameters: Location and Scale

The parameters μ and σ have intuitive geometric interpretations:

- **Location (μ)**: Shifting μ translates the entire distribution left or right without changing its shape
- **Scale (σ)**: Increasing σ spreads the distribution out (flatter and wider); decreasing σ concentrates it (taller and narrower)

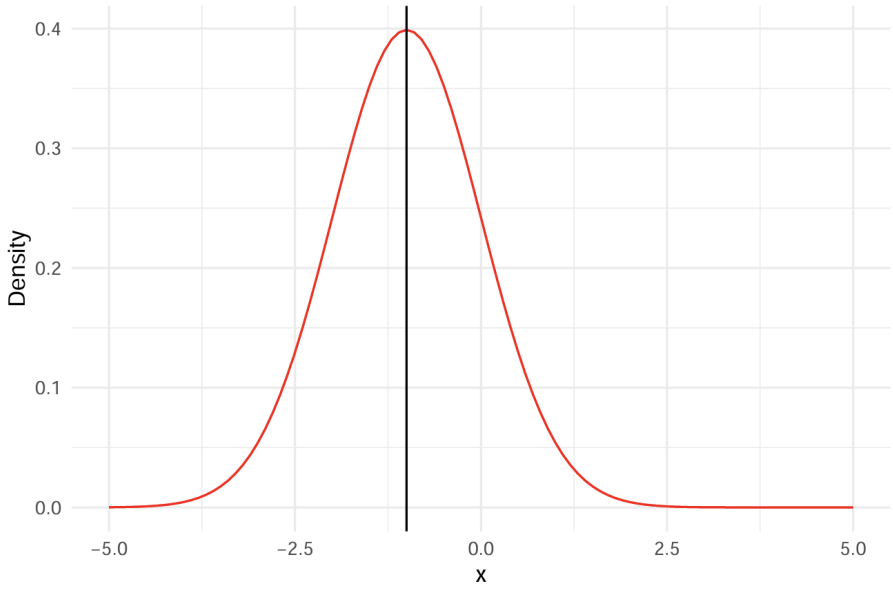


Figure 6.7: Effect of the location parameter μ: shifting the mean translates the distribution horizontally.

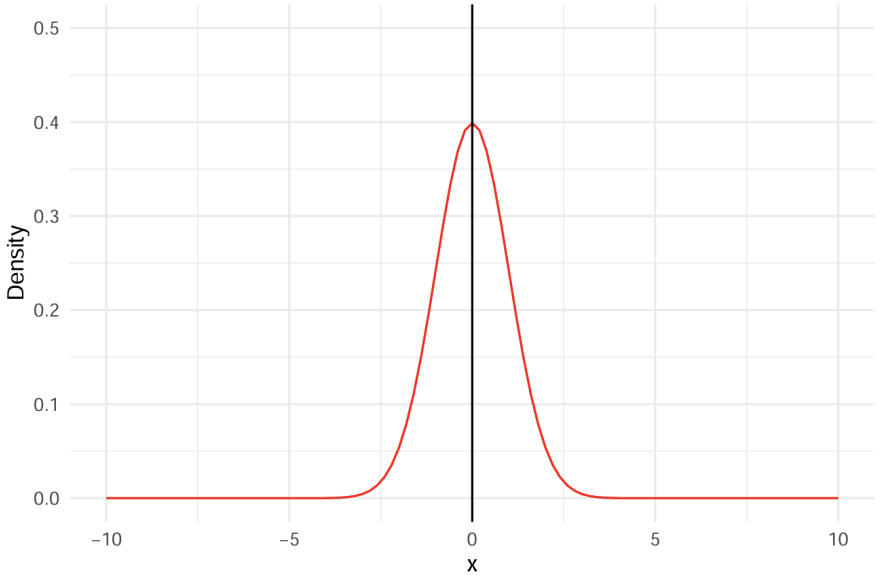


Figure 6.8: Effect of the scale parameter σ: larger σ produces a flatter, wider distribution; smaller σ produces a taller, narrower distribution. The area under the curve remains 1.

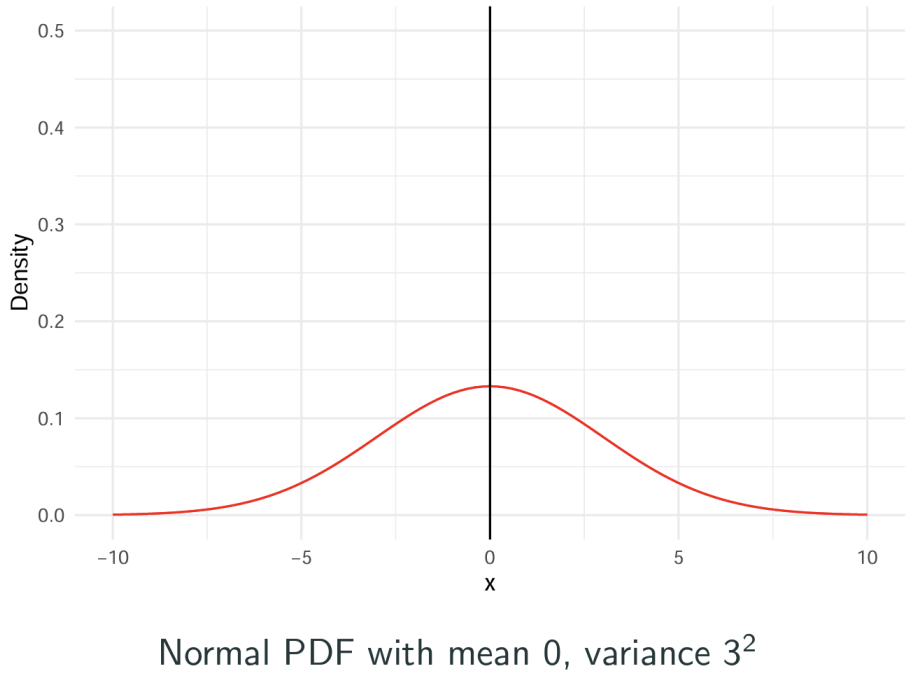


Figure 6.9: Normal distributions with different combinations of μ and σ^2.

6.5.3 Standardisation and Z-Scores

Any Normal random variable can be transformed into a standard Normal through **standardisation**.

Theorem 6.8 (Standardisation). *If $X \sim \mathcal{N}(\mu, \sigma^2)$, then*

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1) \tag{6.14}$$

Conversely, if $Z \sim \mathcal{N}(0, 1)$, then $X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$.

Why Standardisation Works

The transformation $Z = (X - \mu)/\sigma$ involves two operations:

1. **Centring**: Subtracting μ shifts the mean to zero
2. **Scaling**: Dividing by σ rescales the standard deviation to one

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

$$\mathbb{E}[Z] = \mathbb{E}\left[\frac{X - \mu}{\sigma}\right] = \frac{\mathbb{E}[X] - \mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0$$

$$\text{Var}(Z) = \text{Var}\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma^2}\text{Var}(X - \mu) = \frac{1}{\sigma^2} \cdot \sigma^2 = 1$$

Moreover, the Normal family is closed under linear transformations, so Z remains Normal.

Definition 6.8 (Z-Score)**.** The **Z-score** of an observation x from a distribution with mean μ and standard deviation σ is:

$$z = \frac{x - \mu}{\sigma} \tag{6.15}$$

The Z-score measures how many standard deviations x is from the mean.

Using Standardisation to Compute Probabilities

To find $P(X \leq x)$ for $X \sim \mathcal{N}(\mu, \sigma^2)$:

1. Standardise: $z = \frac{x-\mu}{\sigma}$
2. Look up $\Phi(z)$ in a standard Normal table or compute numerically

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

6.5.4 Properties of the Normal Distribution

Theorem 6.9 (Symmetry of the Normal)**.** *The Normal distribution is symmetric about its mean:*

1. ***PDF symmetry:*** $\phi(z) = \phi(-z)$ *for the standard Normal*
2. ***CDF symmetry:*** $\Phi(-z) = 1 - \Phi(z)$

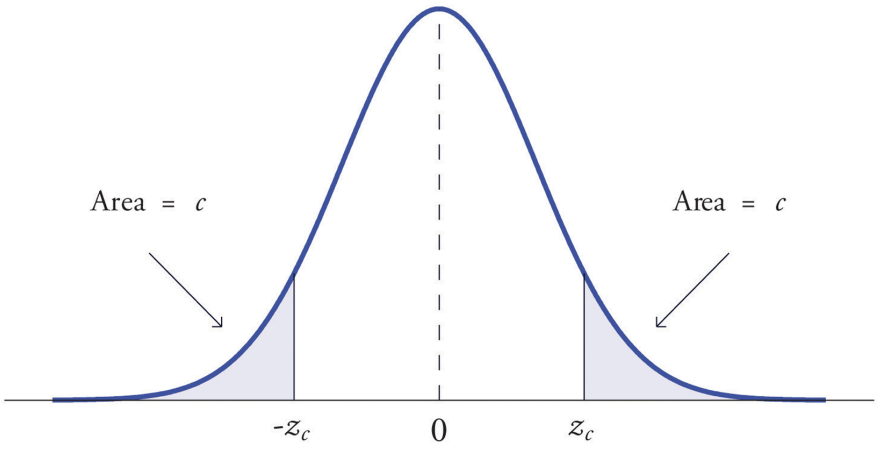


Figure 6.10: Symmetry of the standard Normal: the probability in the left tail below $-z$ equals the probability in the right tail above z.

Empirical Rule (68–95–99.7 Rule)

For $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$P(|X - \mu| < \sigma) \approx 0.68 \quad \text{(within 1 standard deviation)}$$
$$P(|X - \mu| < 2\sigma) \approx 0.95 \quad \text{(within 2 standard deviations)}$$
$$P(|X - \mu| < 3\sigma) \approx 0.997 \quad \text{(within 3 standard deviations)}$$

These benchmarks are fundamental for interpreting Normal data and constructing confidence intervals.

6.5.5 The Central Limit Theorem (Preview)

The Normal distribution's importance stems largely from the Central Limit Theorem:

Theorem 6.10 (Central Limit Theorem—Informal Statement)**.** *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2. Then as $n \to \infty$, the standardised sample mean*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

converges in distribution to the standard Normal $\mathcal{N}(0, 1)$.

Why the CLT Matters

- The original random variables X_i do *not* need to be Normally distributed
- Averages (and sums) of many independent observations tend to be approximately Normal
- This justifies using Normal-based methods (confidence intervals, hypothesis tests) even when the underlying data is not Normal
- The approximation improves as n increases

6.6 The Exponential Distribution

The Exponential distribution models waiting times between events in a Poisson process—it is the continuous analogue of the Geometric distribution.

Definition 6.9 (Exponential Distribution)**.** A continuous random variable X has the **Exponential distribution** with rate parameter $\lambda > 0$, written $X \sim \text{Expo}(\lambda)$, if its PDF is:

$$f_X(x) = \lambda e^{-\lambda x}, \quad x > 0 \tag{6.16}$$

and $f_X(x) = 0$ for $x \leq 0$.

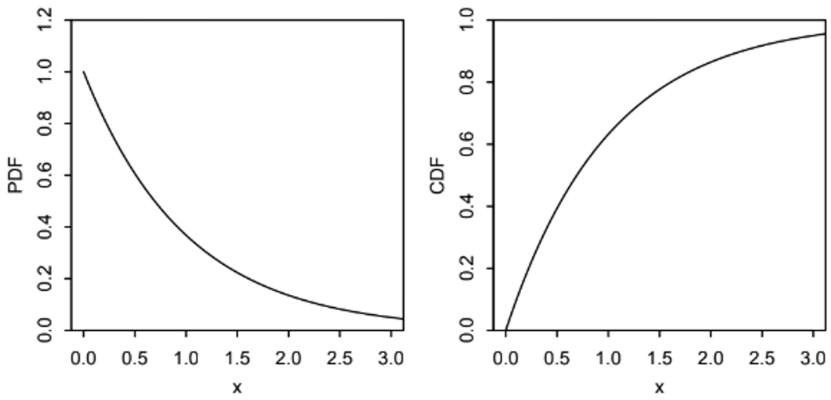


FIGURE 5.12
Expo(1) PDF and CDF.

Figure 6.11: PDF and CDF of Exponential distributions with different rate parameters λ. Larger λ means events occur more frequently, so waiting times are shorter (more probability mass near zero).

6.6.1 Connection to the Poisson Process

The Exponential distribution arises naturally from the Poisson process:

From Poisson to Exponential

Consider events occurring randomly in continuous time (customers arriving, radioactive decays, bus arrivals):

- The **Poisson distribution** counts the *number* of events in a fixed time interval
- The **Exponential distribution** measures the *time* until the next event

If events occur at rate λ (average of λ events per unit time), then:

- Number of events in time t: $N(t) \sim \text{Poisson}(\lambda t)$
- Time until first event: $T \sim \text{Expo}(\lambda)$

Derivation of Exponential from Poisson

Let T be the time until the first event in a Poisson process with rate λ. We derive its CDF:

$$P(T > t) = P(\text{no events in } [0, t]) = P(N(t) = 0)$$

Since $N(t) \sim \text{Poisson}(\lambda t)$:

$$P(N(t) = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$$

Therefore, the CDF is:

$$F_T(t) = P(T \leq t) = 1 - P(T > t) = 1 - e^{-\lambda t}, \quad t > 0$$

Differentiating to obtain the PDF:

$$f_T(t) = \frac{d}{dt}(1 - e^{-\lambda t}) = \lambda e^{-\lambda t}$$

This confirms the Exponential PDF.

6.6.2 CDF of the Exponential Distribution

Theorem 6.11 (Exponential CDF). *For $X \sim Expo(\lambda)$:*

$$F_X(x) = 1 - e^{-\lambda x}, \quad x > 0 \tag{6.17}$$

and $F_X(x) = 0$ for $x \leq 0$.

Remark. The complement $P(X > x) = e^{-\lambda x}$ is the **survival function**—the probability of "surviving" (waiting) beyond time x.

6.6.3 Mean and Variance

Theorem 6.12 (Exponential Mean and Variance). *For $X \sim Expo(\lambda)$:*

$$\mathbb{E}[X] = \frac{1}{\lambda} \tag{6.18}$$

$$\text{Var}(X) = \frac{1}{\lambda^2} \tag{6.19}$$

Intuition

The mean $1/\lambda$ is intuitive: if events occur at rate λ per unit time, then on average you wait $1/\lambda$ time units between events. For example, if buses arrive at rate 6 per hour ($\lambda = 6$), the average wait is 1/6 hours = 10 minutes.

Derivation of Exponential Mean

Using integration by parts with $u = x$ and $dv = \lambda e^{-\lambda x} dx$:

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x \cdot \lambda e^{-\lambda x}\, dx \\
&= \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x}\, dx \quad \text{(integration by parts)} \\
&= 0 + \left[-\frac{1}{\lambda} e^{-\lambda x}\right]_0^\infty \\
&= 0 - \left(-\frac{1}{\lambda}\right) \\
&= \frac{1}{\lambda}
\end{aligned}$$

6.6.4 The Memoryless Property

The Exponential distribution has a remarkable property: it is **memoryless**.

Theorem 6.13 (Memoryless Property). *A random variable X is **memoryless** if for all $s, t \geq 0$:*

$$P(X > s + t \mid X > s) = P(X > t) \tag{6.20}$$

The Exponential distribution is the only *continuous distribution with this property.*

Proof of Memorylessness

For $X \sim \text{Expo}(\lambda)$:

$$\begin{aligned} P(X > s+t \mid X > s) &= \frac{P(X > s+t \text{ and } X > s)}{P(X > s)} \\ &= \frac{P(X > s+t)}{P(X > s)} \quad (\text{since } X > s+t \Rightarrow X > s) \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} \\ &= P(X > t) \end{aligned}$$

What Memorylessness Means

Given that you have already waited s time units without an event, the distribution of the *additional* waiting time is the same as if you had just started waiting. The process has "no memory" of the time already elapsed.

Example: Suppose light bulb lifetimes are Exponentially distributed. If a bulb has been working for 1000 hours, the probability it lasts another 500 hours is the same as the probability a brand-new bulb lasts 500 hours. The bulb does not "wear out" in a probabilistic sense.

Contrast with reality: Human lifetimes are *not* memoryless. An 80-year-old does not have the same life expectancy as a newborn. This is why more flexible distributions (like the Weibull) are used to model situations with "aging" or "wear-out" effects.

Memoryless Does Not Mean Constant

Memorylessness does *not* mean that all waiting times are equal—waiting times still vary randomly according to the Exponential distribution. It means that the *conditional distribution* of remaining time, given you have already waited, equals the *unconditional distribution*.

Exponential Distribution Summary

For $X \sim \text{Expo}(\lambda)$:

$$\begin{aligned} \text{PDF}: &\quad f(x) = \lambda e^{-\lambda x} \quad \text{for } x > 0 \\ \text{CDF}: &\quad F(x) = 1 - e^{-\lambda x} \quad \text{for } x > 0 \\ \text{Mean}: &\quad \mathbb{E}[X] = \frac{1}{\lambda} \\ \text{Variance}: &\quad \text{Var}(X) = \frac{1}{\lambda^2} \\ \text{Key property}: &\quad \text{Memoryless} \end{aligned}$$

6.7 Joint, Marginal, and Conditional Distributions for Continuous Variables

The concepts of joint, marginal, and conditional distributions extend naturally from discrete to continuous random variables, with summation replaced by integration.

6.7.1 Joint CDF and PDF

Definition 6.10 (Joint CDF for Continuous Random Variables)**.** For continuous random variables X and Y, the **joint CDF** is:

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) \tag{6.21}$$

Definition 6.11 (Joint PDF)**.** The **joint PDF** is obtained by taking mixed partial derivatives of the joint CDF:

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \, \partial y} F_{X,Y}(x, y) \tag{6.22}$$

The joint PDF represents the density of probability in two-dimensional space. Integrating over a region gives the probability of (X, Y) falling in that region.

Computing the Joint PDF from Joint CDF

The notation $\frac{\partial^2}{\partial x \, \partial y}$ means we take partial derivatives *sequentially*, not separately:

1. First, differentiate $F_{X,Y}(x, y)$ with respect to x (treating y as constant)
2. Then, differentiate the result with respect to y (treating x as constant)

Example: If $F_{X,Y}(x, y) = \frac{1}{2}x^2y^3$ for (x, y) in some region:

$$\frac{\partial F}{\partial x} = xy^3$$

$$f_{X,Y}(x, y) = \frac{\partial}{\partial y}(xy^3) = 3xy^2$$

Note: This is different from computing the partial derivatives separately. The separate partial derivatives would give:

$$\frac{\partial F}{\partial x} = xy^3$$
$$\frac{\partial F}{\partial y} = \frac{3}{2}x^2y^2$$

These are *not* the joint PDF. The joint PDF requires the *mixed* partial derivative.

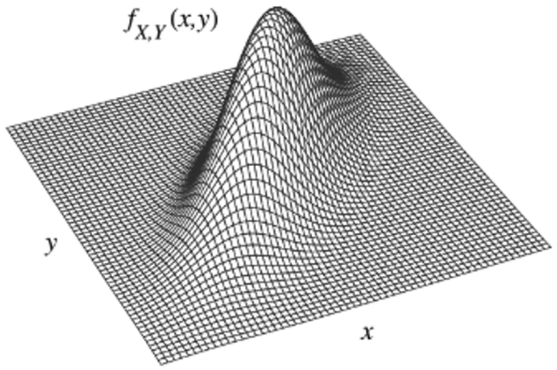


FIGURE 7.4
Joint PDF of continuous r.v.s X and Y.

Figure 6.12: Visualisation of a joint PDF for two continuous random variables. The height of the surface represents probability density. Integrating over a region gives the probability of (X, Y) falling in that region.

6.7.2 Marginal Distributions

Definition 6.12 (Marginal PDF)**.** The **marginal PDF** of X is obtained by integrating out Y from the joint PDF:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy \tag{6.23}$$

Similarly, the marginal PDF of Y is:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx \tag{6.24}$$

Marginalisation

Integrating over Y "sums out" the effect of Y, collapsing the two-dimensional distribution into a one-dimensional distribution for X alone. Geometrically, this is like projecting the joint distribution onto the x-axis.
This is analogous to the discrete case, where we summed over one variable to obtain the marginal distribution.

6.7.3 Conditional Distributions

Definition 6.13 (Conditional PDF)**.** The **conditional PDF** of Y given $X = x$ is:

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)} \tag{6.25}$$

provided $f_X(x) > 0$.

This mirrors the discrete formula: the conditional equals the joint divided by the marginal.

Conditioning on a Zero-Probability Event

For continuous random variables, $P(X = x) = 0$ for any specific x. How can we condition on an event with probability zero?
Formally, we condition on the event that X falls in a small interval $(x-\varepsilon, x+\varepsilon)$ and take the limit as $\varepsilon \to 0$. This limiting procedure is what gives rise to the conditional PDF formula.

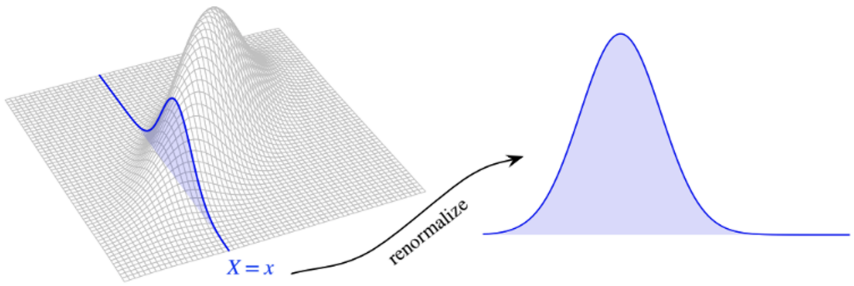


FIGURE 7.5
Conditional PDF of Y given $X = x$. The conditional PDF $f_{Y|X}(y|x)$ is obtained by renormalizing the slice of the joint PDF at the fixed value x.

Figure 6.13: Conditional distribution: fixing $X = x$ corresponds to taking a "slice" through the joint distribution. The conditional PDF of Y given $X = x$ describes the distribution along this slice (renormalised to integrate to 1).

6.7.4 Bayes' Rule and LOTP for Continuous Variables

Theorem 6.14 (Bayes' Rule for Continuous Random Variables)**.**

$$f_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y) \cdot f_Y(y)}{f_X(x)} \tag{6.26}$$

Theorem 6.15 (Law of Total Probability (LOTP) for Continuous Variables)**.**

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x \mid y) \cdot f_Y(y)\, dy \tag{6.27}$$

These are direct analogues of the discrete versions, with sums replaced by integrals.

6.7.5 Mixed Discrete-Continuous Models

In practice, we often work with models involving both discrete and continuous random variables. For example, a mixture model might use a discrete variable to select a component, and a continuous variable for the observation within that component.

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

Figure 6.14: Example of a mixed discrete-continuous model. A discrete variable Y determines which component (distribution) is active; the continuous variable X is then drawn from that component. This structure underlies mixture models and hierarchical models in statistics.

When mixing discrete and continuous variables:

- Condition discrete variables using probabilities (PMF)
- Condition continuous variables using densities (PDF)
- LOTP uses sums over discrete variables and integrals over continuous variables

6.8 Transformations of Random Variables (Introduction)

A common task is to find the distribution of $Y = g(X)$ when we know the distribution of X. For monotonic transformations, there is a systematic approach.

Theorem 6.16 (Change of Variables (Monotonic Case)). *Let X be a continuous random variable with PDF $f_X(x)$. Let $Y = g(X)$ where g is a strictly monotonic (one-to-one) function with inverse g^{-1}. Then Y has PDF:*

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right| \tag{6.28}$$

*The factor $\left|\frac{d}{dy}g^{-1}(y)\right|$ is called the **Jacobian**.*

Why the Jacobian?

The Jacobian accounts for how the transformation stretches or compresses probability. If g stretches an interval by a factor of 2, the density must be divided by 2 to preserve total probability of 1. The absolute value of the derivative of the inverse function measures this stretching/compressing factor.

Example 6.2 (Linear Transformation). Let X have PDF $f_X(x)$. Find the PDF of $Y = aX + b$ where $a \neq 0$.

Solution: The transformation $g(x) = ax + b$ has inverse $g^{-1}(y) = (y - b)/a$. The derivative is:

$$\frac{d}{dy} g^{-1}(y) = \frac{1}{a}$$

Applying the formula:

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

This generalises the standardisation result: if $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Z = (X - \mu)/\sigma$, then $Z \sim \mathcal{N}(0, 1)$.

6.9 Summary

Chapter Summary: Key Concepts

Continuous vs Discrete:

- Continuous random variables have smooth CDFs (no jumps)
- PDF is the derivative of CDF; CDF is the integral of PDF
- $P(X = x) = 0$ for all specific values; probabilities are computed over intervals

Three Fundamental Distributions:

1. **Uniform**(a, b): Constant density on (a, b); mean $= (a+b)/2$; variance $= (b-a)^2/12$
2. **Normal**(μ, σ^2): Bell curve; mean $= \mu$; variance $= \sigma^2$; standardise via $Z = (X-\mu)/\sigma$
3. **Exponential**(λ): Waiting times; mean $= 1/\lambda$; variance $= 1/\lambda^2$; memoryless

Joint Distributions for Continuous Variables:

- Joint PDF: $f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$
- Marginal: Integrate out the other variable
- Conditional: $f_{Y|X}(y|x) = f_{X,Y}(x, y)/f_X(x)$
- Bayes and LOTP: Replace sums with integrals

The tools developed in this chapter—working with PDFs, CDFs, and computing expectations—form the foundation for statistical inference, which we develop in subsequent chapters.

Chapter 7

Continuous Random Variables II

Learning Objectives

By the end of this week, you should be able to:

- Define covariance and compute it using both the definitional and computational formulas
- State and prove the key properties of covariance
- Define the correlation coefficient and interpret its value
- Explain why independence implies zero correlation, but not vice versa
- Construct examples of uncorrelated but dependent random variables
- State and interpret both the Weak and Strong Laws of Large Numbers
- Derive the mean and variance of the sample mean
- State the Central Limit Theorem precisely and explain its significance
- Apply the Normal approximation to sums and means of random variables
- Understand the EM algorithm as a method for maximum likelihood estimation with latent variables
- Derive the E-step and M-step for Gaussian mixture models

Prerequisites

This week assumes familiarity with:

- Continuous random variables, PDFs, and CDFs (Chapter 6)
- Expected value and variance for continuous random variables
- The Normal distribution and standardisation
- Independence of random variables (Chapter 3)
- Basic properties of expectation: linearity, $\mathbb{E}[g(X)]$ via LOTUS
- Maximum likelihood estimation (conceptually)

7.1 Covariance

When working with multiple random variables, we often want to understand how they relate to each other. Covariance quantifies the degree to which two random variables *vary together*—whether they tend to be simultaneously above or below their respective means.

7.1.1 Definition and Interpretation

Definition 7.1 (Covariance)**.** The **covariance** of two random variables X and Y is defined as:

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] \tag{7.1}$$

provided the expectation exists.

Let us unpack this definition term by term:

- $\mathbb{E}[X]$ and $\mathbb{E}[Y]$: The expected values (means) of the random variables, denoted μ_X and μ_Y.
- $(X - \mathbb{E}[X])$ and $(Y - \mathbb{E}[Y])$: The **deviations from the mean**—how far each observation is from its expected value. These are called *centred* random variables.
- $(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])$: The **product of deviations**. This product is:
  - *Positive* when both variables are on the same side of their means (both above or both below)
  - *Negative* when the variables are on opposite sides of their means
- $\mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big]$: The **expected value** of these products, averaging over all possible outcomes. This quantifies the *typical* behaviour of the product of deviations.

Interpreting Covariance

Think of covariance as measuring the *average tendency* of two variables to move together:

- **Positive covariance** ($\mathrm{Cov}(X, Y) > 0$): When X is above its mean, Y tends to be above its mean too (and vice versa). The variables "move together."
- **Negative covariance** ($\mathrm{Cov}(X, Y) < 0$): When X is above its mean, Y tends to be below its mean. The variables "move oppositely."
- **Zero covariance** ($\mathrm{Cov}(X, Y) = 0$): There is no consistent linear relationship. Knowing whether X is above or below its mean tells us nothing (on average) about where Y is relative to its mean.

Remark (Connection to Variance). Variance is a special case of covariance—the covariance of a random variable with itself:

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

This connection hints at why covariance and variance share many algebraic properties.

7.1.2 The Computational Formula

In practice, the definitional formula is cumbersome for calculations. We derive a more convenient form.

Theorem 7.1 (Computational Formula for Covariance)**.**

$$\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \tag{7.2}$$

In words: ***covariance is the expectation of the product minus the product of the expectations****.*

Proof of Theorem 7.1

Starting from the definition:

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] \\ &= \mathbb{E}\big[XY - X\,\mathbb{E}[Y] - Y\,\mathbb{E}[X] + \mathbb{E}[X]\,\mathbb{E}[Y]\big]\end{aligned}$$

We expand using the FOIL method (First, Outer, Inner, Last) and then apply linearity of expectation:

$$\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X\,\mathbb{E}[Y]] - \mathbb{E}[Y\,\mathbb{E}[X]] + \mathbb{E}[\mathbb{E}[X]\,\mathbb{E}[Y]]$$

Now we use the key fact that $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are *constants* (not random variables), so:

- $\mathbb{E}[X\,\mathbb{E}[Y]] = \mathbb{E}[Y] \cdot \mathbb{E}[X]$ (pulling the constant $\mathbb{E}[Y]$ out)
- $\mathbb{E}[Y\,\mathbb{E}[X]] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$ (pulling the constant $\mathbb{E}[X]$ out)
- $\mathbb{E}[\mathbb{E}[X]\,\mathbb{E}[Y]] = \mathbb{E}[X]\,\mathbb{E}[Y]$ (expectation of a constant is itself)

Substituting:

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[Y]\,\mathbb{E}[X] - \mathbb{E}[X]\,\mathbb{E}[Y] + \mathbb{E}[X]\,\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned}$$

Quick Reference: Covariance Formulas

Definitional formula (for conceptual understanding):

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

Computational formula (for calculations):

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

Mnemonic: "Expected product minus product of expectations"

Example 7.1 (Computing Covariance)**.** Let X and Y be discrete random variables with joint PMF:

	$Y=0$	$Y=1$	$P(X=x)$
$X=0$	0.1	0.2	0.3
$X=1$	0.3	0.4	0.7
$P(Y=y)$	0.4	0.6	1.0

Step 1: Compute marginal expectations.

$$\mathbb{E}[X] = 0 \times 0.3 + 1 \times 0.7 = 0.7$$
$$\mathbb{E}[Y] = 0 \times 0.4 + 1 \times 0.6 = 0.6$$

Step 2: Compute $\mathbb{E}[XY]$.

The product XY takes values:

- $XY = 0$ when $(X, Y) \in \{(0,0), (0,1), (1,0)\}$
- $XY = 1$ when $(X, Y) = (1,1)$

So:

$$\mathbb{E}[XY] = 0 \times (0.1 + 0.2 + 0.3) + 1 \times 0.4 = 0.4$$

Step 3: Apply the computational formula.

$$\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0.4 - 0.7 \times 0.6 = 0.4 - 0.42 = -0.02$$

The small negative covariance suggests a very weak tendency for X and Y to move in opposite directions.

7.1.3 Properties of Covariance

Covariance satisfies several important algebraic properties that make it a powerful analytical tool.

Theorem 7.2 (Properties of Covariance)**.** *Let X, Y, and Z be random variables with finite second moments, and let a, b, c be constants. Then:*

(i) ***Covariance with itself is variance:*** $\text{Cov}(X,X) = \text{Var}(X)$

(ii) ***Symmetry:*** $\text{Cov}(X,Y) = \text{Cov}(Y,X)$

(iii) ***Covariance with a constant:*** $\text{Cov}(X,c) = 0$

(iv) ***Scaling:*** $\text{Cov}(aX,Y) = a \cdot \text{Cov}(X,Y)$

(v) ***Bilinearity:*** $\text{Cov}(aX+bY,Z) = a \cdot \text{Cov}(X,Z) + b \cdot \text{Cov}(Y,Z)$

(vi) ***Variance of a sum:*** $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X,Y)$

(vii) ***Adding a constant:*** $\text{Cov}(X+c,Y) = \text{Cov}(X,Y)$

Proofs of Covariance Properties

Property (i): $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

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

by definition of variance.

Property (ii): $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$

Immediate from the computational formula, since multiplication is commutative:

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[YX] - \mathbb{E}[Y]\,\mathbb{E}[X] = \mathrm{Cov}(Y,X)$$

Property (iii): $\mathrm{Cov}(X, c) = 0$

A constant c has $\mathbb{E}[c] = c$, so:

$$\mathrm{Cov}(X,c) = \mathbb{E}[Xc] - \mathbb{E}[X]\,\mathbb{E}[c] = c\,\mathbb{E}[X] - \mathbb{E}[X] \cdot c = 0$$

Intuitively: a constant has no variability, so it cannot co-vary with anything.

Property (iv): $\mathrm{Cov}(aX, Y) = a \cdot \mathrm{Cov}(X, Y)$

$$\begin{aligned}\mathrm{Cov}(aX,Y) &= \mathbb{E}[aXY] - \mathbb{E}[aX]\,\mathbb{E}[Y] \\ &= a\,\mathbb{E}[XY] - a\,\mathbb{E}[X]\,\mathbb{E}[Y] \\ &= a\big(\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\big) \\ &= a \cdot \mathrm{Cov}(X,Y)\end{aligned}$$

Property (v): $\mathrm{Cov}(aX + bY, Z) = a \cdot \mathrm{Cov}(X, Z) + b \cdot \mathrm{Cov}(Y, Z)$

$$\begin{aligned}\mathrm{Cov}(aX+bY,Z) &= \mathbb{E}[(aX+bY)Z] - \mathbb{E}[aX+bY]\,\mathbb{E}[Z] \\ &= a\,\mathbb{E}[XZ] + b\,\mathbb{E}[YZ] - (a\,\mathbb{E}[X] + b\,\mathbb{E}[Y])\,\mathbb{E}[Z] \\ &= a\,\mathbb{E}[XZ] + b\,\mathbb{E}[YZ] - a\,\mathbb{E}[X]\,\mathbb{E}[Z] - b\,\mathbb{E}[Y]\,\mathbb{E}[Z] \\ &= a\big(\mathbb{E}[XZ] - \mathbb{E}[X]\,\mathbb{E}[Z]\big) + b\big(\mathbb{E}[YZ] - \mathbb{E}[Y]\,\mathbb{E}[Z]\big) \\ &= a \cdot \mathrm{Cov}(X,Z) + b \cdot \mathrm{Cov}(Y,Z)\end{aligned}$$

Property (vi): $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathrm{Cov}(X+Y, X+Y) \quad \text{(by property (i))} \\ &= \mathrm{Cov}(X, X+Y) + \mathrm{Cov}(Y, X+Y) \quad \text{(by property (v) with } a=b=1) \\ &= \mathrm{Cov}(X,X) + \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,X) + \mathrm{Cov}(Y,Y) \\ &= \mathrm{Var}(X) + 2\,\mathrm{Cov}(X,Y) + \mathrm{Var}(Y)\end{aligned}$$

Property (vii): $\mathrm{Cov}(X + c, Y) = \mathrm{Cov}(X, Y)$

$$\begin{aligned}\mathrm{Cov}(X+c,Y) &= \mathrm{Cov}(X,Y) + \mathrm{Cov}(c,Y) \quad \text{(by property (v))} \\ &= \mathrm{Cov}(X,Y) + 0 \quad \text{(by property (iii))} \\ &= \mathrm{Cov}(X,Y)\end{aligned}$$

Intuitively: shifting a distribution does not change how it co-varies with other variables.

Variance of a Sum

The formula $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$ is frequently needed. Note that variances are *not* simply additive unless $\mathrm{Cov}(X,Y) = 0$.
The simpler formula $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ holds *only* when X and Y are uncorrelated.
Common mistake: Assuming variances add without checking for correlation.

Remark (Bilinearity and Inner Products). Property (v) shows that covariance is **bilinear**: linear in each argument separately. Combined with symmetry (property ii) and the fact that $\mathrm{Cov}(X,X) \geq 0$, covariance can be viewed as an *inner product* on the space of centred, square-integrable random variables.

This perspective connects probability theory to linear algebra and explains why many results about covariance parallel results about dot products.

Corollary 7.3 (Variance of a Linear Combination). *For random variables $X_1, \ldots, X_n$ and constants $a_1, \ldots, a_n$:*

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(X_i) + 2\sum_{i<j} a_i a_j \,\mathrm{Cov}(X_i, X_j) \tag{7.3}$$

In matrix notation, if $\mathbf{X} = (X_1, \ldots, X_n)^T$ and $\mathbf{a} = (a_1, \ldots, a_n)^T$:

$$\mathrm{Var}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} \tag{7.4}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix with $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$.

7.1.4 Covariance and Independence

There is an important relationship between covariance and independence.

Theorem 7.4 (Independence Implies Zero Covariance). *If X and Y are independent random variables, then $\mathrm{Cov}(X,Y) = 0$.*

Proof of Theorem 7.4

Recall from Chapter 3 the key property of independent random variables: for independent X and Y,

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)] \cdot \mathbb{E}[h(Y)] \tag{7.5}$$

for any functions g and h (provided the expectations exist).

Taking $g(x) = x$ and $h(y) = y$:

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$$

Using the computational formula:

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0$$

Definition 7.2 (Uncorrelated Random Variables). Random variables X and Y are said to be **uncorrelated** if $\mathrm{Cov}(X,Y) = 0$, or equivalently, if $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.

The theorem above states:

$$\text{Independence} \implies \text{Uncorrelated}$$

The Converse is False

The converse does **not** hold in general:

$$\text{Uncorrelated} \not\Longrightarrow \text{Independence}$$

Two random variables can be *perfectly dependent* yet have zero covariance. This is one of the most important distinctions in probability theory.
Why? Independence requires $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)]$ for *all* functions g and h. Being uncorrelated only requires this for the specific case $g(x) = x$ and $h(y) = y$. A relationship can satisfy one condition but not the other.

7.2 Correlation

Covariance measures how two variables move together, but its magnitude depends on the scales of the variables. The correlation coefficient provides a standardised, scale-free measure.

7.2.1 Definition and Basic Properties

Definition 7.3 (Correlation Coefficient)**.** The **(Pearson) correlation coefficient** of two random variables X and Y is:

$$\rho_{X,Y} = \operatorname{Corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\cdot\operatorname{Var}(Y)}} = \frac{\operatorname{Cov}(X,Y)}{\sigma_X\sigma_Y} \tag{7.6}$$

provided $\operatorname{Var}(X) > 0$ and $\operatorname{Var}(Y) > 0$.

Remark (Alternative Expression). We can also write:

$$\rho_{X,Y} = \operatorname{Cov}\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right) = \mathbb{E}\left[\frac{X-\mu_X}{\sigma_X}\cdot\frac{Y-\mu_Y}{\sigma_Y}\right]$$

That is, correlation is the covariance of the *standardised* versions of X and Y. Since standardised variables have variance 1, the denominator in the correlation formula equals 1, and correlation simply becomes the covariance of the standardised variables.

Theorem 7.5 (Correlation is Bounded)**.** *For any random variables X and Y with finite, positive variances:*

$$-1 \le \rho_{X,Y} \le 1 \tag{7.7}$$

Proof of Theorem 7.5

The proof uses the Cauchy–Schwarz inequality for expectations.

Cauchy–Schwarz Inequality: For any random variables U and V with finite second moments:

$$|\,\mathbb{E}[UV]|^2 \le \mathbb{E}[U^2]\cdot\mathbb{E}[V^2]$$

with equality if and only if $U = cV$ almost surely for some constant c.

Let $U = X - \mu_X$ and $V = Y - \mu_Y$. Then:

$$\begin{aligned} |\operatorname{Cov}(X,Y)|^2 &= |\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]|^2 \\ &\leq \mathbb{E}[(X-\mu_X)^2] \cdot \mathbb{E}[(Y-\mu_Y)^2] \\ &= \operatorname{Var}(X) \cdot \operatorname{Var}(Y) \end{aligned}$$

Taking square roots:

$$|\operatorname{Cov}(X,Y)| \leq \sqrt{\operatorname{Var}(X) \cdot \operatorname{Var}(Y)} = \sigma_X \sigma_Y$$

Dividing by $\sigma_X \sigma_Y > 0$:

$$|\rho_{X,Y}| = \frac{|\operatorname{Cov}(X,Y)|}{\sigma_X \sigma_Y} \leq 1$$

Hence $-1 \leq \rho_{X,Y} \leq 1$.

Theorem 7.6 (When Correlation Equals ± 1)**.** $|\rho_{X,Y}| = 1$ *if and only if there exist constants* $a \neq 0$ *and* b *such that*

$$Y = aX + b \quad \textit{with probability 1}$$

Specifically:

- $\rho_{X,Y} = +1$ *when* $a > 0$ *(perfect positive linear relationship)*
- $\rho_{X,Y} = -1$ *when* $a < 0$ *(perfect negative linear relationship)*

Proof of Theorem 7.6

By Cauchy–Schwarz, equality $|\mathbb{E}[UV]|^2 = \mathbb{E}[U^2]\,\mathbb{E}[V^2]$ holds if and only if U and V are linearly dependent, i.e., $U = cV$ for some constant c.

With $U = X - \mu_X$ and $V = Y - \mu_Y$, this means:

$$X - \mu_X = c(Y - \mu_Y) \quad \Rightarrow \quad X = cY + (\mu_X - c\mu_Y)$$

Or equivalently, $Y = aX + b$ where $a = 1/c$ and $b = \mu_Y - \mu_X/c$.

The sign of ρ matches the sign of $\operatorname{Cov}(X,Y)$:

$$\operatorname{Cov}(X, aX+b) = a\operatorname{Cov}(X,X) + \operatorname{Cov}(X,b) = a\operatorname{Var}(X)$$

Since $\operatorname{Var}(X) > 0$:

- If $a > 0$, then $\operatorname{Cov}(X,Y) > 0$, so $\rho = +1$
- If $a < 0$, then $\operatorname{Cov}(X,Y) < 0$, so $\rho = -1$

Interpreting Correlation Values

$\rho_{X,Y}$	**Interpretation**
$+1$	Perfect positive linear relationship
0.7 to 0.9	Strong positive association
0.4 to 0.6	Moderate positive association
0.1 to 0.3	Weak positive association
0	No linear association
-0.1 to -0.3	Weak negative association
-0.4 to -0.6	Moderate negative association
-0.7 to -0.9	Strong negative association
-1	Perfect negative linear relationship

Key caveat: These interpretations assume the relationship is approximately linear. Non-linear relationships can produce misleading correlation values.

7.2.2 Properties of Correlation

Theorem 7.7 (Properties of Correlation). *Let X and Y be random variables with positive variances, and let a, b, c, d be constants with $a \neq 0$ and $c \neq 0$. Then:*

(i) ***Bounded****:* $-1 \leq \rho_{X,Y} \leq 1$

(ii) ***Symmetric****:* $\rho_{X,Y} = \rho_{Y,X}$

(iii) ***Scale invariance****:* $\rho_{aX+b,cY+d} = sign(ac) \cdot \rho_{X,Y}$

(iv) ***Correlation with itself****:* $\rho_{X,X} = 1$

Proof of Scale Invariance

For $U = aX + b$ and $V = cY + d$:

$$\text{Cov}(U, V) = \text{Cov}(aX + b, cY + d) = ac \cdot \text{Cov}(X, Y)$$

And:

$$\sigma_U = |a|\sigma_X$$
$$\sigma_V = |c|\sigma_Y$$

Therefore:

$$\rho_{U,V} = \frac{ac \cdot \text{Cov}(X, Y)}{|a|\sigma_X \cdot |c|\sigma_Y} = \frac{ac}{|a||c|} \cdot \rho_{X,Y} = \text{sign}(ac) \cdot \rho_{X,Y}$$

Scale Invariance

Correlation is unchanged by:

- **Shifting**: Adding constants (b, d) does not affect correlation
- **Positive scaling**: Multiplying by positive constants does not affect correlation
- **Negative scaling**: Multiplying by a negative constant flips the sign of correlation

This makes correlation a *unitless* measure. Whether heights are in centimetres or inches, the correlation between height and weight is the same.

7.2.3 Correlation Measures Linear Relationships

A crucial limitation of correlation is that it measures only *linear* association.

Correlation and Non-Linear Relationships

Two random variables can be **perfectly dependent** (one is a deterministic function of the other) yet have **zero correlation** if the relationship is non-linear.
Correlation detects linear patterns only. Before concluding that two variables are unrelated based on zero correlation, always check for non-linear relationships.

Example 7.2 (Uncorrelated but Perfectly Dependent)**.** Let $X \sim \mathcal{N}(0,1)$ and define $Y = X^2$.

Then Y is a *deterministic function* of X—knowing X tells us exactly what Y is. Yet:

Claim: $\text{Cov}(X,Y) = 0$, hence $\rho_{X,Y} = 0$.

Proof: Using the computational formula:

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

We need three quantities:

1. $\mathbb{E}[X] = 0$ (since $X \sim \mathcal{N}(0,1)$)
2. $\mathbb{E}[Y] = \mathbb{E}[X^2] = \text{Var}(X) + (\mathbb{E}[X])^2 = 1 + 0 = 1$
3. $\mathbb{E}[XY] = \mathbb{E}[X \cdot X^2] = \mathbb{E}[X^3]$

For $X \sim \mathcal{N}(0,1)$, all odd moments are zero (by symmetry of the standard Normal about zero):

$$\mathbb{E}[X^3] = 0$$

Therefore:

$$\text{Cov}(X,Y) = \mathbb{E}[X^3] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0 - 0 \cdot 1 = 0$$

So X and $Y = X^2$ are uncorrelated, despite Y being completely determined by X.

Geometric interpretation: The parabolic relationship $Y = X^2$ is symmetric about the Y-axis. Positive deviations of X from its mean (0) give the same Y values as negative deviations. The positive and negative contributions to covariance cancel exactly.

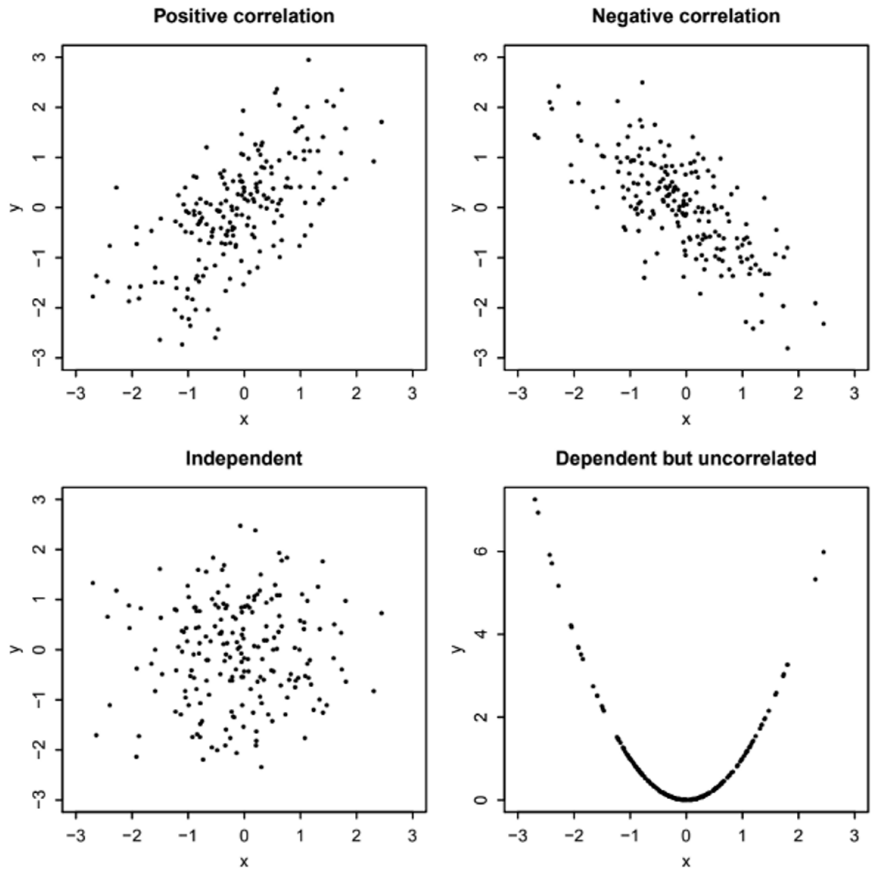


Figure 7.1: Left: A linear relationship produces high correlation. Right: A quadratic relationship ($Y = X^2$) can produce zero correlation despite perfect dependence. Correlation captures linear patterns only.

7.2.4 Independence versus Uncorrelated: Summary

Independence vs Uncorrelated

Independence is the stronger condition:

- Independence $\implies$ Uncorrelated (always)
- Uncorrelated $\not\implies$ Independence (in general)

Independence: $P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$ for all events A, B

- Knowing X tells us *nothing* about Y
- Implies $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)] \cdot \mathbb{E}[h(Y)]$ for all functions g, h

Uncorrelated: $\mathrm{Cov}(X, Y) = 0$, equivalently $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$

- Knowing X tells us nothing about Y *on average, linearly*
- Only requires $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ (one specific choice of functions)

Important exception: If (X, Y) are **jointly Normally distributed**, then:

$$\text{Uncorrelated} \iff \text{Independent}$$

This is a special property of the multivariate Normal distribution.

Example 7.3 (Verifying Independence via Covariance)**.** Suppose X and Y are discrete random variables where $X \in \{1, 2\}$ with equal probability, and $Y \in \{3, 4\}$ with equal probability. Assume X and Y are independent.

Step 1: Compute marginal expectations.

$$\mathbb{E}[X] = \frac{1}{2}(1) + \frac{1}{2}(2) = 1.5$$
$$\mathbb{E}[Y] = \frac{1}{2}(3) + \frac{1}{2}(4) = 3.5$$

Step 2: Compute $\mathbb{E}[XY]$.

By independence, each combination (x, y) has probability $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$:

$$\begin{aligned}\mathbb{E}[XY] &= \frac{1}{4}(1 \times 3) + \frac{1}{4}(1 \times 4) + \frac{1}{4}(2 \times 3) + \frac{1}{4}(2 \times 4) \\ &= \frac{1}{4}(3 + 4 + 6 + 8) = \frac{21}{4} = 5.25\end{aligned}$$

Step 3: Verify the covariance.

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 5.25 - (1.5 \times 3.5) = 5.25 - 5.25 = 0$$

As expected for independent random variables.

7.3 Law of Large Numbers

The Law of Large Numbers (LLN) is one of the fundamental theorems of probability theory. It formalises the intuition that sample averages converge to population means as the sample size grows.

7.3.1 Setup and Notation

Consider a sequence of independent and identically distributed (i.i.d.) random variables $X_1, X_2, X_3, \ldots$ with common mean $\mu = \mathbb{E}[X_i]$ and variance $\sigma^2 = \mathrm{Var}(X_i)$.

Definition 7.4 (Sample Mean)**.** The **sample mean** of n observations is:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n} \tag{7.8}$$

The sample mean $\bar{X}_n$ is itself a random variable—its value depends on which particular observations we draw. Different samples give different values of $\bar{X}_n$.

Example 7.4 (Sample Mean as a Random Variable)**.** Suppose we measure daily temperatures $X_1, X_2, \ldots, X_5$ over five days:

$$X_1 = 3°\mathrm{C}, \quad X_2 = 5°\mathrm{C}, \quad X_3 = 7°\mathrm{C}, \quad X_4 = 2°\mathrm{C}, \quad X_5 = 4°\mathrm{C}$$

The sample mean is:

$$\bar{X}_5 = \frac{3+5+7+2+4}{5} = 4.2°\mathrm{C}$$

If we had measured a *different* five days, we would get a different value of $\bar{X}_5$. This is why $\bar{X}_5$ is a random variable—it varies across different possible samples.

7.3.2 Properties of the Sample Mean

What are the expectation and variance of $\bar{X}_n$? These properties are fundamental to understanding sampling.

Theorem 7.8 (Expectation of the Sample Mean)**.** *For i.i.d. random variables $X_1, \ldots, X_n$ with mean μ:*

$$\mathbb{E}[\bar{X}_n] = \mu \tag{7.9}$$

Proof of Theorem 7.8

Using linearity of expectation:

$$\begin{aligned}
\mathbb{E}[\bar{X}_n] &= \mathbb{E}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] \\
&= \frac{1}{n}\big(\mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]\big) \\
&= \frac{1}{n}(\mu + \mu + \cdots + \mu) \quad (n \text{ terms}) \\
&= \frac{1}{n} \cdot n\mu = \mu
\end{aligned}$$

Remark (Unbiasedness). This result says that $\bar{X}_n$ is an **unbiased estimator** of μ: on average, the sample mean equals the population mean, regardless of sample size n.

Unbiasedness is a desirable property for estimators—it means we are not systematically over- or under-estimating the parameter of interest.

Theorem 7.9 (Variance of the Sample Mean). *For i.i.d. random variables $X_1, \ldots, X_n$ with variance σ^2:*

$$\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \tag{7.10}$$

The standard deviation of the sample mean is therefore:

$$SD(\bar{X}_n) = \frac{\sigma}{\sqrt{n}} \tag{7.11}$$

*This quantity is called the **standard error** of the mean.*

Proof of Theorem 7.9

Since the X_i are independent, they are uncorrelated, so variances add:

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) = n\sigma^2$$

For the sample mean, we use the property $\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$:

$$\begin{aligned}\mathrm{Var}(\bar{X}_n) &= \mathrm{Var}\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) \\ &= \frac{1}{n^2}\mathrm{Var}(X_1 + \cdots + X_n) \\ &= \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}\end{aligned}$$

Why Variance Decreases

The variance of the sample mean decreases as n increases because averaging "smooths out" individual variability. Extreme values in one observation tend to be balanced by less extreme values in others.
The $1/n$ factor (not $1/n^2$) arises because:

- The sum has variance proportional to n (variances add)
- Dividing by n squares this factor in the variance, giving $n/n^2 = 1/n$

Sample Mean: Key Properties

For i.i.d. random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2, the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies:

$$\begin{aligned}\text{Mean}: &\quad \mathbb{E}[\bar{X}_n] = \mu \\ \text{Variance}: &\quad \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \\ \text{Standard error}: &\quad \mathrm{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}\end{aligned}$$

Key insight: The variance decreases as n increases, making the sample mean increasingly concentrated around the true mean. This is the essence of the Law of Large Numbers.

7.3.3 Two Forms of Convergence

Before stating the Laws of Large Numbers, we need to clarify what "convergence" means for random variables. There are several types; two are relevant here.

Definition 7.5 (Convergence in Probability)**.** A sequence of random variables Y_n **converges in probability** to a constant c, written $Y_n \xrightarrow{P} c$, if for every $\varepsilon > 0$:

$$\lim_{n\to\infty} P(|Y_n - c| > \varepsilon) = 0$$

Equivalently, $\lim_{n\to\infty} P(|Y_n - c| \leq \varepsilon) = 1$.

Definition 7.6 (Almost Sure Convergence)**.** A sequence of random variables Y_n **converges almost surely** (or with probability 1) to c, written $Y_n \xrightarrow{\text{a.s.}} c$, if:

$$P\left(\lim_{n\to\infty} Y_n = c\right) = 1$$

Comparing Convergence Types

- **Convergence in probability**: For any tolerance ε, the probability of being more than ε away from the limit goes to zero. But for any finite n, there is still *some* probability of being far away.
- **Almost sure convergence**: The sequence *actually converges* (in the ordinary calculus sense) for almost all outcomes. There may be a set of "bad" outcomes (of probability zero) where convergence fails, but for all other outcomes, the sequence converges to c.

Almost sure convergence is stronger: $Y_n \xrightarrow{\text{a.s.}} c$ implies $Y_n \xrightarrow{P} c$, but not vice versa.

7.3.4 Statement of the Law of Large Numbers

Theorem 7.10 (Weak Law of Large Numbers (WLLN))**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and finite variance σ^2. Then for any $\varepsilon > 0$:*

$$\lim_{n\to\infty} P\big(|\bar{X}_n - \mu| > \varepsilon\big) = 0 \tag{7.12}$$

Equivalently, $\bar{X}_n \xrightarrow{P} \mu$ (convergence in probability).

What the Weak Law Says

No matter how small a margin of error ε you specify, the probability that the sample mean differs from the true mean by more than ε goes to zero as $n \to \infty$.
In practical terms: with a large enough sample, the sample mean will be arbitrarily close to the population mean, with arbitrarily high probability.

Proof of WLLN via Chebyshev's Inequality

Recall Chebyshev's inequality: for any random variable Y with mean μ_Y and variance σ_Y^2:

$$P(|Y - \mu_Y| \geq \varepsilon) \leq \frac{\sigma_Y^2}{\varepsilon^2}$$

Apply this to $Y = \bar{X}_n$, which has mean μ and variance σ^2/n:

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2/n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

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

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

Theorem 7.11 (Strong Law of Large Numbers (SLLN))**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite mean μ. Then:*

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1 \tag{7.13}$$

That is, $\bar{X}_n \xrightarrow{a.s.} \mu$ (almost sure convergence).

Weak vs Strong Law

- **Weak Law**: For any fixed ε, the probability of being wrong by more than ε goes to zero. But it does not preclude occasional large deviations for finite n.
- **Strong Law**: The sample mean *actually converges* to μ with probability 1. There is a set of "bad" outcomes (of probability zero) where convergence fails, but for every other outcome, convergence occurs.

The strong law is indeed stronger—it implies the weak law but not vice versa.

Remark (Technical Notes).

- The strong law requires only finite mean (not variance), while the Chebyshev-based proof of the weak law requires finite variance. More sophisticated proofs of the weak law also need only finite mean.
- The proof of the strong law is considerably more technical, typically using the Borel–Cantelli lemmas or martingale theory. We omit it here.
- For distributions without finite mean (e.g., Cauchy), the LLN does not hold—the sample mean does not stabilise.

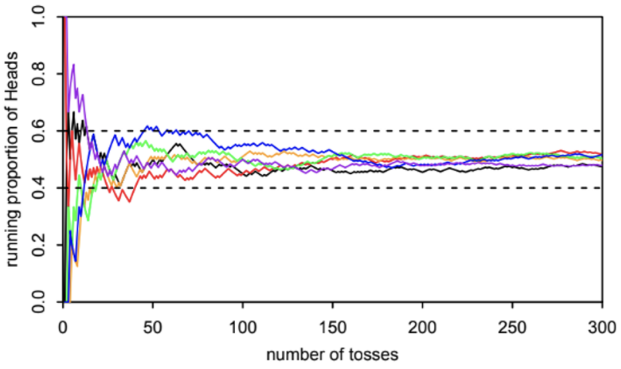


FIGURE 10.2
Running proportion of Heads in 6 sequences of fair coin tosses. Dashed lines at 0.6 and 0.4 are plotted for reference. As the number of tosses increases, the proportion of Heads approaches 1/2.

Figure 7.2: Illustration of the Law of Large Numbers: sample means from repeated experiments converge to the population mean as sample size increases. The variance of the sample mean distribution shrinks at rate $1/n$.

7.3.5 Applications of the LLN

Example 7.5 (Monte Carlo Integration)**.** To estimate $\theta = \mathbb{E}[g(X)]$ where X has a known distribution:

1. Generate i.i.d. samples $X_1, \ldots, X_n$ from the distribution of X
2. Compute $Y_i = g(X_i)$ for each sample
3. Estimate θ by $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$

By the LLN, $\hat{\theta}_n \to \theta$ as $n \to \infty$.

This is the foundation of Monte Carlo methods in statistics and machine learning.

Example 7.6 (Relative Frequency Interpretation of Probability)**.** If A is an event with $P(A) = p$, define indicator variables:

$$X_i = \begin{cases} 1 & \text{if } A \text{ occurs on trial } i \\ 0 & \text{otherwise} \end{cases}$$

Then $\mathbb{E}[X_i] = p$, and the sample mean is:

$$\bar{X}_n = \frac{\text{number of times } A \text{ occurred}}{n} = \text{relative frequency of } A$$

By the LLN, relative frequency $\to p$ as $n \to \infty$. This justifies the "long-run frequency" interpretation of probability.

7.3.6 Common Misconceptions

The Gambler's Fallacy

The LLN does **not** say that short-term deviations will be "corrected" by opposite outcomes. A sequence of 10 heads does not make tails more likely on the next flip.
Convergence occurs through **swamping**, not correction: past deviations become negligible compared to the growing number of future observations. The past tosses are not "undone"; they simply become a smaller and smaller fraction of the total.

Example 7.7 (Swamping, Not Correction)**.** Suppose you flip a fair coin and get 10 heads in a row. Your current proportion of heads is $10/10 = 100\%$.

After 10 more flips (with expected 5 heads), your expected proportion is $(10 + 5)/20 = 75\%$.

After 100 more flips (with expected 50 heads), your expected proportion is $(10+50)/110 \approx 54.5\%$.

After 1000 more flips (with expected 500 heads), your expected proportion is $(10 + 500)/1010 \approx 50.5\%$.

The initial "streak" is not corrected—it is simply swamped by subsequent data. Each flip remains 50-50 regardless of past outcomes.

7.4 Central Limit Theorem

The Central Limit Theorem (CLT) is arguably the most important theorem in probability and statistics. It explains why the Normal distribution appears so frequently in nature and provides the theoretical foundation for many statistical procedures.

7.4.1 Statement of the Central Limit Theorem

Theorem 7.12 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and finite variance $\sigma^2 > 0$. Then the **standardised sample mean**:*

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \tag{7.14}$$

converges in distribution to the standard Normal as $n \to \infty$:

$$Z_n \xrightarrow{d} \mathcal{N}(0, 1) \tag{7.15}$$

Equivalently, for any $z \in \mathbb{R}$:

$$\lim_{n\to\infty} P(Z_n \leq z) = \Phi(z)$$

where Φ is the standard Normal CDF.

Let us unpack the standardisation:

- $\bar{X}_n$ has mean μ and variance σ^2/n (from our earlier results)
- Subtracting μ centres the distribution at zero: $\mathbb{E}[\bar{X}_n - \mu] = 0$
- Dividing by $\sigma/\sqrt{n}$ (the standard deviation of $\bar{X}_n$) scales the variance to 1
- The result Z_n is a standardised random variable with mean 0 and variance 1
- The CLT says the *shape* of Z_n's distribution approaches the standard Normal bell curve

Central Limit Theorem: Key Points

What it says: The standardised sample mean converges *in distribution* to $\mathcal{N}(0,1)$.
What it does NOT require: The original X_i do *not* need to be Normally distributed. They can be Uniform, Exponential, Bernoulli, Poisson, or any distribution with finite variance.
What it DOES require:

- Independence (or weak dependence—there are extensions)
- Identical distributions (same μ and σ^2)
- Finite variance $\sigma^2 < \infty$

Practical rule of thumb: The Normal approximation is usually adequate for $n \geq 30$, though this depends on how "non-Normal" the original distribution is. Symmetric distributions need smaller n; highly skewed distributions need larger n.

Definition 7.7 (Convergence in Distribution)**.** A sequence Y_n **converges in distribution** to Y, written $Y_n \xrightarrow{d} Y$, if for all y where the CDF F_Y is continuous:

$$\lim_{n\to\infty} P(Y_n \leq y) = P(Y \leq y)$$

This is the weakest form of convergence—it only requires that the CDFs converge, not that the random variables themselves become close.

7.4.2 Equivalent Formulations

The CLT can be restated in several equivalent and practically useful ways.

Corollary 7.13 (CLT for the Sample Mean)**.** *For large n, the sample mean is approximately Normal:*

$$\bar{X}_n \dot{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{7.16}$$

where $\dot{\sim}$ denotes "approximately distributed as."

Derivation

If $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$, then for large n:

$$Z_n \dot{\sim} \mathcal{N}(0,1)$$

Rearranging: $\bar{X}_n = \mu + \frac{\sigma}{\sqrt{n}} Z_n$.

Since a linear transformation of a Normal is Normal:

$$\bar{X}_n \dot{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Corollary 7.14 (CLT for the Sum)**.** *The sum $S_n = X_1 + \cdots + X_n$ is approximately Normal:*

$$S_n \dot{\sim} \mathcal{N}(n\mu, n\sigma^2) \tag{7.17}$$

Derivation of Theorem 7.14

Since $S_n = n\bar{X}_n$ and $\bar{X}_n \dot{\sim} \mathcal{N}(\mu, \sigma^2/n)$:

$$\mathbb{E}[S_n] = n\,\mathbb{E}[\bar{X}_n] = n\mu$$

$$\mathrm{Var}(S_n) = n^2\,\mathrm{Var}(\bar{X}_n) = n^2 \cdot \frac{\sigma^2}{n} = n\sigma^2$$

Hence $S_n \dot{\sim} \mathcal{N}(n\mu, n\sigma^2)$.

7.4.3 Visualising the CLT

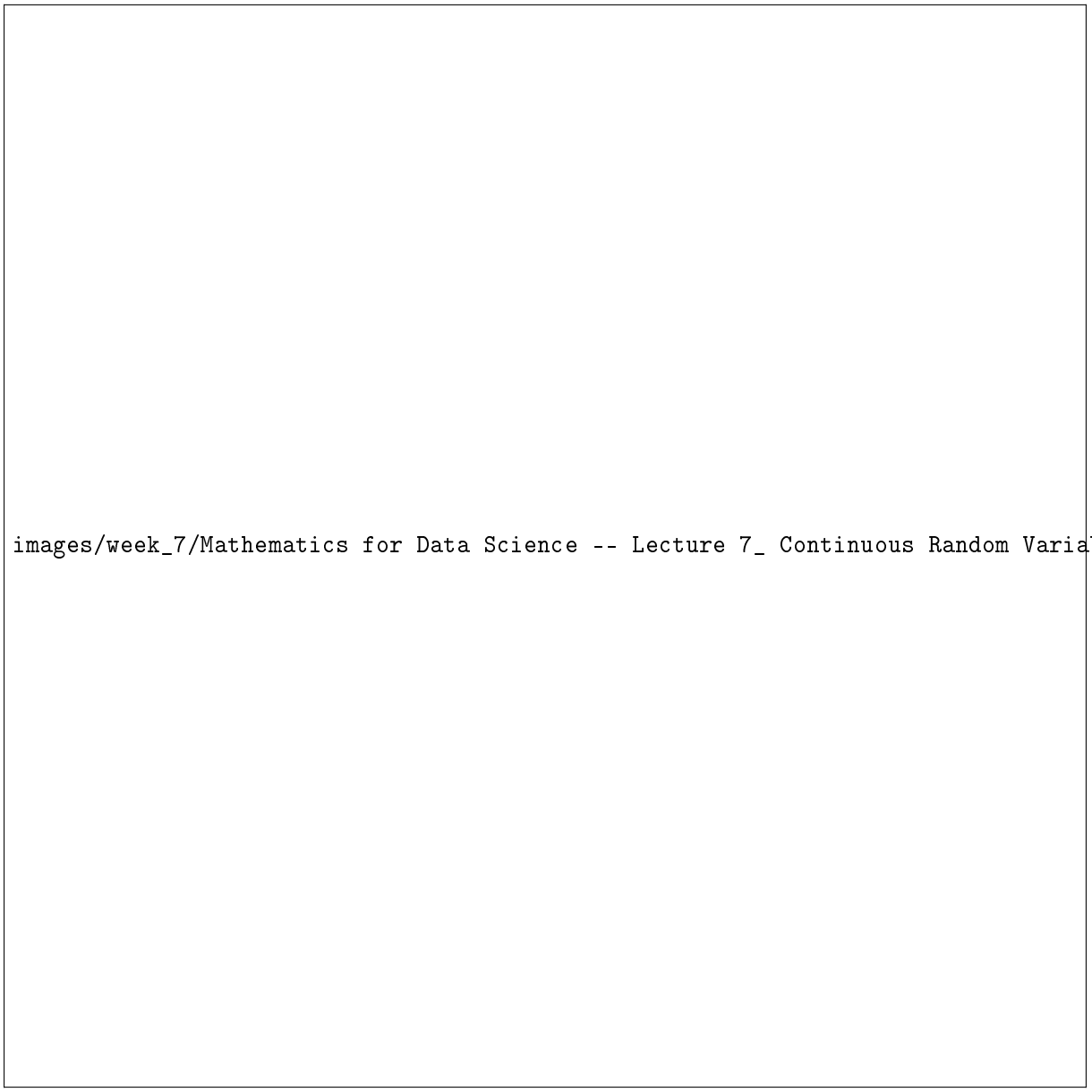


Figure 7.3: The Central Limit Theorem in action. Top row: sampling distributions for different sample sizes from a Geometric distribution. As n increases, the distribution of $\bar{X}_n$ becomes increasingly Normal, regardless of the skewed shape of the original distribution.

Understanding the CLT Visualisation

- $n = 1$: Each "sample mean" is just a single observation, so the distribution of $\bar{X}_1$ is identical to the original distribution (here, Geometric—highly right-skewed).
- n **small**: We start averaging multiple observations. The distribution becomes more symmetric and concentrated.
- n **moderate** (~ 30): The distribution is close to Normal, despite the original being highly skewed.
- **Large** n: The distribution of sample means is nearly indistinguishable from Normal.
- **Variance shrinking**: Notice that the distribution becomes narrower as n increases, reflecting $\text{Var}(\bar{X}_n) = \sigma^2/n$.

7.4.4 Normal Approximation to the Binomial

An important and historically significant application of the CLT is approximating the Binomial distribution.

Theorem 7.15 (Normal Approximation to Binomial)**.** *If $X \sim$ Binomial(n, p), then for large n:*

$$X \dot{\sim} \mathcal{N}(np, np(1-p)) \tag{7.18}$$

Derivation via CLT

Recall that $X \sim \text{Binomial}(n, p)$ can be written as a sum of i.i.d. Bernoulli trials:

$$X = Y_1 + Y_2 + \cdots + Y_n$$

where $Y_i \sim \text{Bernoulli}(p)$ are i.i.d. with:

$$\mathbb{E}[Y_i] = p$$
$$\text{Var}(Y_i) = p(1-p)$$

By the CLT for sums (Theorem 7.14):

$$X = \sum_{i=1}^{n} Y_i \dot{\sim} \mathcal{N}(n \cdot p, n \cdot p(1-p)) = \mathcal{N}(np, np(1-p))$$

Remark (Rule of Thumb for Normal Approximation to Binomial). The approximation is generally considered adequate when both:

- $np \geq 10$ (enough expected successes)
- $n(1-p) \geq 10$ (enough expected failures)

Some sources use 5 instead of 10 as the threshold.

Remark (Continuity Correction). Since the Binomial is discrete and the Normal is continuous, accuracy can be improved using a **continuity correction**:

$$P(X \leq k) \approx \Phi\left(\frac{k + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

The +0.5 adjustment accounts for the fact that the continuous Normal must approximate a discrete distribution.

Example 7.8 (Normal Approximation to Binomial)**.** Suppose $X \sim \text{Binomial}(100, 0.3)$. Find $P(X \leq 25)$.

Setup:

$$\mu = np = 100 \times 0.3 = 30$$
$$\sigma^2 = np(1-p) = 100 \times 0.3 \times 0.7 = 21$$
$$\sigma = \sqrt{21} \approx 4.58$$

Check conditions: $np = 30 \geq 10$ and $n(1-p) = 70 \geq 10$. Approximation is appropriate.

Without continuity correction:

$$P(X \leq 25) \approx P\left(Z \leq \frac{25 - 30}{4.58}\right) = P(Z \leq -1.09) \approx 0.138$$

With continuity correction:

$$P(X \leq 25) \approx P\left(Z \leq \frac{25.5 - 30}{4.58}\right) = P(Z \leq -0.98) \approx 0.164$$

The exact value (from Binomial tables or software) is approximately 0.163, so the continuity correction improves accuracy.

7.4.5 Why Does the CLT Work?

Why Sums Become Normal

The CLT is remarkable because it works for *any* starting distribution (with finite variance). The key insight is that summing many independent contributions has a "smoothing" effect:

1. **Extreme values average out**: Very large or very small individual values become rare when averaged with many others.
2. **Moderate values accumulate**: Most sums end up near the middle because there are exponentially more ways to achieve middle values than extreme values.
3. **Symmetry emerges**: Even if the original distribution is skewed, the sum of many independent copies tends to be symmetric because positive and negative deviations from the mean tend to cancel.

Sketch of Proof via Characteristic Functions

The rigorous proof uses characteristic functions (or moment generating functions):

1. The characteristic function of a sum of independent RVs is the product of their characteristic functions.
2. For i.i.d. X_i with mean μ and variance σ^2, the characteristic function of X_i near 0 behaves like $\phi(t) = 1 + i\mu t - \frac{\sigma^2 t^2}{2} + O(t^3)$.

3. The characteristic function of the standardised sum converges to $e^{-t^2/2}$, which is the characteristic function of $\mathcal{N}(0,1)$.
4. By Lévy's continuity theorem, convergence of characteristic functions implies convergence in distribution.

The full proof requires careful analysis of the higher-order terms.

7.4.6 Extensions and Generalisations

Remark (Beyond i.i.d.). The CLT has been generalised in many directions:

- **Lindeberg–Feller CLT**: Allows non-identical distributions, provided no single term dominates the sum.
- **Lyapunov CLT**: Gives conditions based on moments rather than distributions.
- **Martingale CLT**: Extends to dependent sequences with martingale structure.
- **Multivariate CLT**: Vector-valued random variables converge to multivariate Normal.

These generalisations are important in advanced probability and statistics.

When the CLT Fails

The CLT requires finite variance. For "heavy-tailed" distributions like the Cauchy, which has no finite variance (or even mean), the CLT does not apply.
For Cauchy random variables, the sample mean has the *same* distribution as a single observation—averaging does not concentrate the distribution at all!

7.5 The Expectation-Maximisation (EM) Algorithm

The Expectation-Maximisation (EM) algorithm is a powerful iterative method for finding maximum likelihood estimates when the model involves **latent (hidden) variables**. It is particularly useful for mixture models, where observations arise from one of several possible distributions but we do not know which.

7.5.1 Motivation: Mixture Models and Latent Variables

Consider the Old Faithful geyser data: eruption durations appear to come from two distinct populations—short eruptions and long eruptions—corresponding to different geological processes. However, we do not directly observe which type each eruption is. The eruption type is a **latent variable**.

Definition 7.8 (Latent Variable)**.** A **latent variable** is a variable that affects the observed data but is not directly observed. It is "hidden" or "unobserved."

Definition 7.9 (Gaussian Mixture Model)**.** A **Gaussian mixture model** (GMM) with K components assumes that each observation x is generated by:

1. First, selecting a component $k \in \{1, \ldots, K\}$ with probability π_k (the **mixing proportions**, with $\sum_{k=1}^{K} \pi_k = 1$)
2. Then, drawing x from $\mathcal{N}(\mu_k, \sigma_k^2)$

The marginal density of X is:

$$f(x) = \sum_{k=1}^{K} \pi_k \cdot \phi(x; \mu_k, \sigma_k^2) \tag{7.19}$$

where $\phi(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ is the Normal PDF.

Mixture Models

A mixture model says: "Each data point comes from one of K different populations, but we do not know which." The populations may have different means, variances, or even different distributional shapes.
Examples:

- Customer segments (high-value vs low-value)
- Genetic subpopulations
- Topic models in text analysis (each document is a mixture of topics)

For the Old Faithful example with $K = 2$ components:

- $Z_i \in \{1, 2\}$: Latent variable indicating eruption type (unobserved)
- $\pi_1 = P(Z_i = 1)$: Probability of short eruption type
- $\pi_2 = P(Z_i = 2) = 1 - \pi_1$: Probability of long eruption type
- $X_i \mid Z_i = k \sim \mathcal{N}(\mu_k, \sigma_k^2)$: Duration given type

The parameters to estimate are: $\theta = (\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

7.5.2 Why Standard MLE Fails

For a sample $x_1, \ldots, x_n$, the likelihood is:

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} \Big[\pi_1 \phi(x_i; \mu_1, \sigma_1^2) + \pi_2 \phi(x_i; \mu_2, \sigma_2^2)\Big] \tag{7.20}$$

The log-likelihood is:

$$\ell(\theta) = \sum_{i=1}^{n} \log\Big[\pi_1 \phi(x_i; \mu_1, \sigma_1^2) + \pi_2 \phi(x_i; \mu_2, \sigma_2^2)\Big] \tag{7.21}$$

The Log-Sum Problem

Standard MLE requires taking derivatives and setting them to zero. The problem is that the log-likelihood contains log(sum), which does not simplify nicely.
Differentiating ℓ with respect to μ_1:

$$\frac{\partial \ell}{\partial \mu_1} = \sum_{i=1}^{n} \frac{\pi_1 \cdot \frac{\partial}{\partial \mu_1}\phi(x_i;\mu_1,\sigma_1^2)}{\pi_1\phi(x_i;\mu_1,\sigma_1^2) + \pi_2\phi(x_i;\mu_2,\sigma_2^2)}$$

Setting this to zero yields equations where all parameters are intertwined—there is no closed-form solution.
Contrast with simple MLE: For a single Normal distribution, $\ell = \sum \log \phi(x_i;\mu,\sigma^2)$, and the log pulls through to give clean, separable equations.

7.5.3 The Key Insight: Responsibilities

The EM algorithm is based on a clever observation. Looking at the derivative above, we notice a recurring term:

$$\gamma_{ik} = \frac{\pi_k \phi(x_i;\mu_k,\sigma_k^2)}{\sum_{j=1}^{K} \pi_j \phi(x_i;\mu_j,\sigma_j^2)} \tag{7.22}$$

This is recognisable via Bayes' theorem as the **posterior probability** that observation i belongs to component k:

$$\gamma_{ik} = P(Z_i = k \mid X_i = x_i, \theta)$$

Definition 7.10 (Responsibilities)**.** The **responsibility** γ_{ik} is the posterior probability that data point i was generated by component k, given the observed data and current parameter estimates:

$$\gamma_{ik} = P(Z_i = k \mid x_i) = \frac{\pi_k \phi(x_i;\mu_k,\sigma_k^2)}{\sum_{j=1}^{K} \pi_j \phi(x_i;\mu_j,\sigma_j^2)} \tag{7.23}$$

Deriving Responsibilities via Bayes' Theorem

By Bayes' theorem:

$$P(Z_i = k \mid X_i = x_i) = \frac{P(X_i = x_i \mid Z_i = k) \cdot P(Z_i = k)}{P(X_i = x_i)}$$

The terms are:

- $P(X_i = x_i \mid Z_i = k) = \phi(x_i;\mu_k,\sigma_k^2)$ (likelihood given component)
- $P(Z_i = k) = \pi_k$ (prior probability of component)
- $P(X_i = x_i) = \sum_{j=1}^{K} \pi_j \phi(x_i;\mu_j,\sigma_j^2)$ (marginal likelihood)

Substituting gives the responsibility formula.

What Responsibilities Mean

The responsibility γ_{ik} is a "soft assignment" of observation i to component k:

- If $\gamma_{i1} = 0.9$, we are 90% confident observation i came from component 1
- Responsibilities sum to 1 across components: $\sum_k \gamma_{ik} = 1$ for each i
- If we *knew* the true assignments (the Z_i), MLE would be straightforward—just fit a Normal to each subset. The responsibilities provide probabilistic "guesses" of these assignments.

7.5.4 The EM Algorithm: Structure

The EM algorithm exploits this insight by alternating between two steps:

1. **E-step (Expectation)**: Given current parameter estimates, compute the expected value of the latent variables—i.e., the responsibilities.
2. **M-step (Maximisation)**: Given the responsibilities, update parameters to maximise the expected complete-data log-likelihood.

The key idea: if we knew the latent variables, MLE would be easy. If we knew the parameters, computing the latent variable posteriors would be easy. EM alternates between these, using each to inform the other.

EM Algorithm: High-Level Structure

Goal: Find MLE for parameters θ when some variables Z are unobserved.
Initialisation: Choose starting values $\theta^{(0)}$.
Iterate until convergence:
E-Step: Compute $Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{Z|X,\theta^{(t)}}[\log L(\theta; X, Z)]$
This is the expected complete-data log-likelihood, where the expectation is over the latent variables Z given observed data X and current parameters $\theta^{(t)}$.
M-Step: Set $\theta^{(t+1)} = \arg\max_\theta Q(\theta \mid \theta^{(t)})$

7.5.5 EM for Gaussian Mixture Models

For Gaussian mixture models, the E-step and M-step take particularly clean forms.

The EM Algorithm for Gaussian Mixtures

Initialisation: Choose initial values for parameters $\pi_k^{(0)}$, $\mu_k^{(0)}$, $(\sigma_k^2)^{(0)}$ for $k = 1, \ldots, K$.

Iterate until convergence:

E-Step: For each observation i and component k, compute the responsibility:

$$\gamma_{ik}^{(t)} = \frac{\pi_k^{(t)} \phi(x_i; \mu_k^{(t)}, (\sigma_k^2)^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \phi(x_i; \mu_j^{(t)}, (\sigma_j^2)^{(t)})} \tag{7.24}$$

M-Step: Update parameters using weighted averages based on responsibilities:

First, compute the "effective number" of points in each component:

$$N_k^{(t)} = \sum_{i=1}^{n} \gamma_{ik}^{(t)} \tag{7.25}$$

Then update:

$$\mu_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{i=1}^{n} \gamma_{ik}^{(t)} x_i \tag{7.26}$$

$$(\sigma_k^2)^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{i=1}^{n} \gamma_{ik}^{(t)} (x_i - \mu_k^{(t+1)})^2 \tag{7.27}$$

$$\pi_k^{(t+1)} = \frac{N_k^{(t)}}{n} \tag{7.28}$$

Convergence check: Evaluate the log-likelihood:

$$\ell^{(t+1)} = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k^{(t+1)} \phi(x_i; \mu_k^{(t+1)}, (\sigma_k^2)^{(t+1)}) \right]$$

Stop when $|\ell^{(t+1)} - \ell^{(t)}| < \varepsilon$ for some tolerance ε.

EM for GMM: Summary

E-Step (soft assignment):

$$\gamma_{ik} = \frac{\pi_k \phi(x_i; \mu_k, \sigma_k^2)}{\sum_{j=1}^{K} \pi_j \phi(x_i; \mu_j, \sigma_j^2)}$$

M-Step (weighted parameter updates):

$$N_k = \sum_{i=1}^{n} \gamma_{ik} \quad \text{(effective count)}$$

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} x_i \quad \text{(weighted mean)}$$

$$\hat{\sigma}_k^2 = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} (x_i - \hat{\mu}_k)^2 \quad \text{(weighted variance)}$$

$$\hat{\pi}_k = \frac{N_k}{n} \quad \text{(fraction of responsibility)}$$

7.5.6 Derivation of the M-Step Updates

Why do the M-step updates take this form? They come from maximising the expected complete-data log-likelihood.

Derivation of $\hat{\mu}_k$

If we observed the latent variables Z_i, the complete-data log-likelihood would be:

$$\ell_c(\theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbf{1}_{Z_i=k}\left[\log \pi_k + \log \phi(x_i; \mu_k, \sigma_k^2)\right]$$

Taking expectation over Z given X and current parameters, and using $\mathbb{E}[\mathbf{1}_{Z_i=k}] = \gamma_{ik}$:

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik}\left[\log \pi_k - \frac{1}{2}\log(2\pi\sigma_k^2) - \frac{(x_i-\mu_k)^2}{2\sigma_k^2}\right]$$

Differentiating with respect to μ_k and setting to zero:

$$\frac{\partial Q}{\partial \mu_k} = \sum_{i=1}^{n} \gamma_{ik} \cdot \frac{x_i - \mu_k}{\sigma_k^2} = 0$$

$$\sum_{i=1}^{n} \gamma_{ik}(x_i - \mu_k) = 0$$

$$\sum_{i=1}^{n} \gamma_{ik} x_i = \mu_k \sum_{i=1}^{n} \gamma_{ik}$$

$$\mu_k = \frac{\sum_{i=1}^{n} \gamma_{ik} x_i}{\sum_{i=1}^{n} \gamma_{ik}} = \frac{\sum_{i=1}^{n} \gamma_{ik} x_i}{N_k}$$

This is a *weighted average* of the observations, where the weight for observation i is its responsibility γ_{ik}.

Interpreting the M-Step

The M-step updates have natural interpretations:

- $\hat{\mu}_k$: Weighted mean of observations, where weights reflect how much each observation "belongs" to component k. If responsibilities were hard (0 or 1), this would be the sample mean of points assigned to component k.
- $\hat{\sigma}_k^2$: Weighted variance around the new mean.
- $N_k = \sum_i \gamma_{ik}$: The "effective number" of points in component k. Not an integer, because points are softly assigned.
- $\hat{\pi}_k = N_k/n$: The fraction of total "responsibility mass" assigned to component k.

7.5.7 Convergence Properties

Theorem 7.16 (EM Monotonically Increases Likelihood)**.** *Each iteration of the EM algorithm increases (or maintains) the observed-data log-likelihood:*

$$\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$$

with equality only at a stationary point.

Why EM Works: The Lower Bound Perspective

The EM algorithm can be understood through the lens of *lower bounds*:

1. The E-step constructs a lower bound $Q(\theta \mid \theta^{(t)})$ on the log-likelihood that *touches* the log-likelihood at $\theta^{(t)}$.
2. The M-step maximises this lower bound, finding $\theta^{(t+1)}$.
3. Because Q is a lower bound that equals ℓ at $\theta^{(t)}$, we have:

$$\ell(\theta^{(t+1)}) \geq Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta^{(t)} \mid \theta^{(t)}) = \ell(\theta^{(t)})$$

Each iteration "lifts" the lower bound, which in turn lifts the log-likelihood.

EM Convergence Caveats

1. **Local, not global**: EM converges to a *local* maximum (or saddle point), not necessarily the global maximum. The final solution depends on initialisation.
2. **Multiple initialisations**: In practice, run EM multiple times with different starting values and choose the solution with highest log-likelihood.
3. **Convergence can be slow**: Near the optimum, convergence is often linear (each iteration reduces the error by a constant factor), which can be slower than Newton-type methods that converge quadratically.
4. **Degeneracies**: With insufficient data or poor initialisation, a component may "collapse" onto a single point, causing $\sigma_k \to 0$ and $\ell \to \infty$. Regularisation or variance constraints can help.

7.5.8 Practical Considerations

1. **Initialisation strategies**:
   - **Random**: Sample initial means from the data points
   - **K-means**: Use K-means clustering to get initial hard assignments, then estimate parameters from each cluster
   - **K-means++**: A more careful random initialisation that spreads out initial centres
   - **Heuristic**: Use domain knowledge (e.g., for bimodal data, initialise means near the modes)
2. **Choosing K**: The number of components is typically chosen using:
   - **Information criteria**: AIC, BIC (penalise model complexity)
   - **Cross-validation**: Hold out data and evaluate predictive performance
   - **Domain knowledge**: Sometimes the number of groups is known a priori
3. **Numerical stability**: When computing responsibilities, work with **log-probabilities** to avoid underflow. The "log-sum-exp" trick is essential:

$$\log\left(\sum_k e^{a_k}\right) = m + \log\left(\sum_k e^{a_k - m}\right)$$

where $m = \max_k a_k$.

Remark (EM Beyond Mixture Models). The EM algorithm applies to any model with latent variables, not just mixtures:

- Hidden Markov models
- Factor analysis
- Missing data problems
- Variational autoencoders (variational EM)

The general principle—alternate between inferring latent variables and updating parameters—is widely applicable.

7.6 Summary

Week 7 Summary

Covariance and Correlation:

- $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$ measures joint variability
- $\rho_{X,Y} = \text{Cov}(X, Y)/(\sigma_X \sigma_Y) \in [-1, 1]$ is the standardised, scale-free version
- Independence $\implies$ uncorrelated, but uncorrelated $\not\Longrightarrow$ independent
- Correlation measures *linear* association only; non-linear relationships can produce $\rho = 0$
- Exception: For jointly Normal (X, Y), uncorrelated $\iff$ independent

Law of Large Numbers:

- Sample mean $\bar{X}_n$ has $\mathbb{E}[\bar{X}_n] = \mu$ (unbiased) and $\text{Var}(\bar{X}_n) = \sigma^2/n$
- WLLN: $P(|\bar{X}_n - \mu| > \varepsilon) \to 0$ as $n \to \infty$ (convergence in probability)
- SLLN: $\bar{X}_n \to \mu$ almost surely (stronger)
- Convergence is by *swamping*, not correction (avoid gambler's fallacy)

Central Limit Theorem:

- Standardised sample mean converges in distribution to $\mathcal{N}(0, 1)$:

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

- Holds regardless of the original distribution (given finite variance)
- Enables Normal approximations: $\bar{X}_n \dot{\sim} \mathcal{N}(\mu, \sigma^2/n)$
- Applications: confidence intervals, hypothesis tests, approximating Binomial

EM Algorithm:

- Iterative method for MLE with latent variables
- E-step: Compute responsibilities (posterior probabilities of latent assignments)
- M-step: Update parameters using weighted averages
- Monotonically increases likelihood; converges to local maximum
- Use multiple initialisations to find global maximum

The concepts in this week—covariance, the LLN, and the CLT—form the theoretical backbone of statistical inference. The CLT, in particular, justifies the widespread use of Normal-based confidence intervals and hypothesis tests, even when the underlying data are not Normal. The EM algorithm demonstrates how elegant iterative methods can solve problems that appear intractable with direct optimisation, and its "E-step / M-step" structure appears throughout modern machine learning.

Part III

Linear Algebra

Chapter 8

Linear Algebra I

Learning Objectives

By the end of this week, you should be able to:

- Define and distinguish between scalars, vectors, matrices, and tensors
- Use compact notation to express data structures in terms of real number spaces
- Perform matrix operations: transpose, addition, scalar multiplication, and matrix multiplication
- Verify conformability conditions for matrix multiplication
- Express systems of linear equations in matrix form
- Understand the identity matrix and its role in defining matrix inverses
- Compute and interpret vector norms (L_1, L_2, L_∞)
- Derive the ordinary least squares (OLS) estimator using matrix calculus
- Understand and derive the Ridge, Lasso, and Elastic Net regression estimators
- Explain the geometric and statistical intuition behind regularisation

Prerequisites

This week assumes familiarity with:

- Basic calculus: differentiation, partial derivatives (**??**)
- Elementary algebra and equation manipulation
- Linear regression concepts (from statistics courses)
- Summation notation

8.1 Data Structures in Linear Algebra

Linear algebra provides the language and machinery for working with structured numerical data. Before diving into operations, we must understand the fundamental objects we manipulate.

8.1.1 Scalars, Vectors, Matrices, and Tensors

Definition 8.1 (Scalar)**.** A **scalar** is a single numerical value. We denote scalars with lowercase letters:

$$x = 5, \quad \alpha = 0.01, \quad c = -3.14$$

Scalars are 0-dimensional objects—they have no notion of direction, only magnitude.

Definition 8.2 (Vector)**.** A **vector** is an ordered list of scalars. We denote vectors with bold lowercase letters. By convention, vectors are *column vectors*:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

A vector with n elements is called an **n-dimensional vector** or an **n-vector**.

Geometric Interpretation of Vectors

A vector can be visualised as an arrow in space:

- A 2-dimensional vector $\mathbf{x} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$ represents an arrow from the origin to the point $(3, 4)$ in the plane.
- A 3-dimensional vector specifies a point (or direction) in 3D space.
- An n-dimensional vector specifies a point in n-dimensional space—harder to visualise but mathematically identical.

In data science, each element of a vector often represents a different *feature* or *variable* for a single observation.

Definition 8.3 (Matrix)**.** A **matrix** is a rectangular array of scalars arranged in rows and columns. We denote matrices with bold uppercase letters:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

This matrix has m rows and n columns, so we say it has **dimensions** $m \times n$ (read "m by n"). A generic element in row i and column j is denoted a_{ij}.

Index Ordering Convention

Matrix indices follow the convention [row, column], so a_{ij} refers to row i, column j. This matches array indexing in languages like R and Julia, but note that Python uses 0-based indexing.
A common source of confusion: when we write dimensions as "$m \times n$", we mean m rows and n columns. The first number is always rows.

In data science contexts:

- **Rows** typically represent *observations* (data points, samples, individuals)
- **Columns** typically represent *variables* (features, attributes, predictors)

So a dataset with 100 observations and 5 features would be represented as a 100×5 matrix.

Definition 8.4 (Tensor)**.** A **tensor** is a multi-dimensional array that generalises scalars, vectors, and matrices:

- A scalar is a 0th-order tensor (0 indices needed)
- A vector is a 1st-order tensor (1 index needed: x_i)
- A matrix is a 2nd-order tensor (2 indices needed: a_{ij})
- A 3rd-order tensor requires 3 indices: a_{ijk}

A 3rd-order tensor can be visualised as a "cube" or "stack" of matrices. Elements are indexed by three subscripts: a_{ijk} refers to the element at position (i, j, k).

Example 8.1 (Tensors in Practice)**.** Tensors appear naturally in many data science applications:

- **Colour images**: A single RGB image is a 3rd-order tensor with dimensions (height $\times$ width $\times$ 3 colour channels).
- **Video**: A video is a 4th-order tensor (time $\times$ height $\times$ width $\times$ channels).
- **Batch processing**: In deep learning, a batch of images is a 4th-order tensor (batch size $\times$ height $\times$ width $\times$ channels).

8.1.2 Compact Notation

We use set membership notation to compactly express the type and dimensionality of mathematical objects.

Compact Notation for Data Structures

$$\begin{aligned}
\text{Scalar:} &\quad x \in \mathbb{R} \quad (\text{or equivalently } x \in \mathbb{R}^1) \\
\text{Vector:} &\quad \mathbf{x} \in \mathbb{R}^n \quad (n\text{-dimensional column vector}) \\
\text{Matrix:} &\quad \mathbf{A} \in \mathbb{R}^{m \times n} \quad (m \text{ rows}, n \text{ columns}) \\
\text{Tensor:} &\quad \mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_k} \quad (k\text{th-order tensor})
\end{aligned}$$

The symbol $\mathbb{R}$ denotes the set of all real numbers. The notation $\mathbf{x} \in \mathbb{R}^n$ means "$\mathbf{x}$ is an element of n-dimensional real space"—in other words, $\mathbf{x}$ is a vector of n real numbers.

Remark (Row vs Column Vectors). By convention, when we write $\mathbf{x} \in \mathbb{R}^n$ without further specification, we mean a *column vector*. To denote a row vector explicitly, we would write $\mathbf{x}^\top \in \mathbb{R}^{1\times n}$ (a $1 \times n$ matrix).

8.2 The Transpose Operation

The transpose is one of the most fundamental matrix operations, interchanging rows and columns.

Definition 8.5 (Transpose)**.** The **transpose** of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, denoted $\mathbf{A}^\top$ (or sometimes $\mathbf{A}'$), is the matrix $\mathbf{A}^\top \in \mathbb{R}^{n\times m}$ obtained by interchanging rows and columns:

$$(\mathbf{A}^\top)_{ij} = a_{ji}$$

That is, the element in row i, column j of $\mathbf{A}^\top$ equals the element in row j, column i of $\mathbf{A}$.

Example 8.2 (Transpose of a Matrix)**.**

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \quad \Longrightarrow \quad \mathbf{A}^\top = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{bmatrix}$$

With numerical values:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}^\top = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Visualising the Transpose

Think of the transpose as "flipping" the matrix along its main diagonal (the diagonal from top-left to bottom-right). Equivalently:

1. Take the first column of $\mathbf{A}$ and lay it out as the first row of $\mathbf{A}^\top$.
2. Take the second column of $\mathbf{A}$ and lay it out as the second row of $\mathbf{A}^\top$.
3. Continue for all columns.

The transpose applies equally to vectors. Since vectors are column matrices:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad \Longrightarrow \quad \mathbf{x}^\top = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$$

A scalar can be viewed as a 1×1 matrix, so its transpose is itself.

Theorem 8.1 (Properties of the Transpose)**.** *For matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *of appropriate dimensions, and scalar* c*:*

$$(\mathbf{A}^\top)^\top = \mathbf{A} \tag{8.1}$$

$$(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top \tag{8.2}$$

$$(c\mathbf{A})^\top = c\mathbf{A}^\top \tag{8.3}$$

$$(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top \tag{8.4}$$

Order Reversal in Products

Equation (8.4) is crucial and frequently tested: the transpose of a product *reverses* the order of the factors. This extends to longer products:

$$(\mathbf{ABC})^\top = \mathbf{C}^\top \mathbf{B}^\top \mathbf{A}^\top$$

For vectors, this gives us: $(\mathbf{a}^\top \mathbf{b})^\top = \mathbf{b}^\top \mathbf{a}$. Since the dot product is a scalar, this tells us $\mathbf{a}^\top \mathbf{b} = \mathbf{b}^\top \mathbf{a}$—the dot product is commutative.

Definition 8.6 (Symmetric Matrix)**.** A square matrix $\mathbf{A}$ is **symmetric** if $\mathbf{A}^\top = \mathbf{A}$, i.e., $a_{ij} = a_{ji}$ for all i, j.

Symmetric matrices arise frequently in statistics and machine learning. For any matrix $\mathbf{X}$, the products $\mathbf{X}^\top \mathbf{X}$ and $\mathbf{X}\mathbf{X}^\top$ are always symmetric (verify this using the transpose properties).

8.3 Matrix Addition and Scalar Multiplication

8.3.1 Matrix Addition

Definition 8.7 (Matrix Addition)**.** For two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ of *identical dimensions*, their sum $\mathbf{C} = \mathbf{A} + \mathbf{B}$ is defined element-wise:

$$c_{ij} = a_{ij} + b_{ij} \quad \text{for all } i = 1, \ldots, m \text{ and } j = 1, \ldots, n$$

Example 8.3 (Matrix Addition)**.**

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 4 \\ 5 & 6 & 6 \end{bmatrix}$$

Dimension Requirement

Matrix addition is only defined when both matrices have *exactly the same dimensions*. You cannot add a 2×3 matrix to a 3×2 matrix—they must match in both rows and columns.

8.3.2 Scalar Multiplication

Definition 8.8 (Scalar Multiplication)**.** For a scalar $c \in \mathbb{R}$ and matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the product $c\mathbf{A}$ is defined element-wise:

$$(c\mathbf{A})_{ij} = c \cdot a_{ij} \quad \text{for all } i, j$$

Example 8.4 (Scalar Multiplication)**.**

$$3 \times \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 9 & 12 \end{bmatrix}$$

Scalar multiplication "scales" every element of the matrix by the same factor. Geometrically, for vectors, this stretches or shrinks the arrow (and reverses direction if the scalar is negative).

8.4 Matrix Multiplication

Matrix multiplication is perhaps the most important operation in linear algebra, but it is *not* element-wise multiplication. Understanding this operation deeply is essential.

8.4.1 The Dot Product (Inner Product)

Before tackling general matrix multiplication, we start with the simpler case of multiplying two vectors.

Definition 8.9 (Dot Product / Inner Product)**.** For two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ of equal length, their **dot product** (or **inner product**) is:

$$\mathbf{a}^\top \mathbf{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

The result is a scalar.

Example 8.5 (Computing a Dot Product)**.**

$$\mathbf{a}^\top \mathbf{b} = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = 1 \cdot 4 + 2 \cdot 5 + 3 \cdot 6 = 4 + 10 + 18 = 32$$

Geometric Interpretation of the Dot Product

The dot product has a beautiful geometric interpretation. For vectors $\mathbf{a}$ and $\mathbf{b}$:

$$\mathbf{a}^\top \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos\theta$$

where $\|\mathbf{a}\|$ and $\|\mathbf{b}\|$ are the lengths (norms) of the vectors and θ is the angle between them. This tells us:

- **Positive dot product**: The vectors point in roughly the same direction ($\theta < 90$).
- **Zero dot product**: The vectors are *orthogonal* (perpendicular, $\theta = 90$).
- **Negative dot product**: The vectors point in roughly opposite directions ($\theta > 90$).

The dot product measures "how much" one vector points in the direction of another.

8.4.2 Matrix-Matrix Multiplication

Definition 8.10 (Matrix Multiplication)**.** For matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times p}$, their product $\mathbf{C} = \mathbf{AB}$ is a matrix $\mathbf{C} \in \mathbb{R}^{m \times p}$ where each element is:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{in} b_{nj}$$

Equivalently, c_{ij} is the dot product of row i of $\mathbf{A}$ with column j of $\mathbf{B}$.

Conformability and Output Dimensions

Conformability condition: For **AB** to be defined, the number of columns of **A** must equal the number of rows of **B**.
Mnemonic: Think of the dimensions as $(m \times \mathbf{n}) \cdot (\mathbf{n} \times p) = (m \times p)$.

- The **inner dimensions** (n) must match—this is the conformability condition.
- The **outer dimensions** (m and p) give the shape of the result.

Example 8.6 (Matrix Multiplication: 2×3 by 3×2)**.**

$$\mathbf{AB} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{bmatrix}$$

Check: $(2 \times 3) \cdot (3 \times 2) = (2 \times 2)$. Inner dimensions match (both 3), output is 2×2.

Numerical example:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{bmatrix} = \begin{bmatrix} 1(7) + 2(9) + 3(11) & 1(8) + 2(10) + 3(12) \\ 4(7) + 5(9) + 6(11) & 4(8) + 5(10) + 6(12) \end{bmatrix} = \begin{bmatrix} 58 & 64 \\ 139 & 154 \end{bmatrix}$$

Algorithm for Matrix Multiplication

To compute element c_{ij} of the product $\mathbf{C} = \mathbf{AB}$:

1. Identify the ith row of **A**: $(a_{i1}, a_{i2}, \ldots, a_{in})$
2. Identify the jth column of **B**: $(b_{1j}, b_{2j}, \ldots, b_{nj})^\top$
3. Compute their dot product: $c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$

Repeat for all $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, p\}$.

Visual heuristic: To find c_{ij}, "slide" row i of **A** across to column j of **B**, multiply corresponding elements, and sum.

Quick Reference: Small Matrix Products

2×2 **matrices**:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{pmatrix}$$

3×3 **matrices**:

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} j & k & l \\ m & n & o \\ p & q & r \end{pmatrix} = \begin{pmatrix} aj + bm + cp & ak + bn + cq & al + bo + cr \\ dj + em + fp & dk + en + fq & dl + eo + fr \\ gj + hm + ip & gk + hn + iq & gl + ho + ir \end{pmatrix}$$

Matrix Multiplication is Not Commutative

In general, $\mathbf{AB} \neq \mathbf{BA}$. In fact:

- If $\mathbf{AB}$ is defined, $\mathbf{BA}$ might not even be defined (dimensions may not conform).
- Even when both products are defined, they may have different dimensions.
- Even when both products have the same dimensions, the matrices are typically different.

This non-commutativity is a major departure from scalar arithmetic and a common source of errors.

8.4.3 Matrix-Vector Multiplication

Matrix-vector multiplication is a special case of matrix-matrix multiplication, but it is so common that it deserves explicit attention.

For $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{x} \in \mathbb{R}^n$, the product $\mathbf{y} = \mathbf{Ax}$ is a vector $\mathbf{y} \in \mathbb{R}^m$ where:

$$y_i = \sum_{j=1}^{n} a_{ij} x_j = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n$$

Example 8.7 (Matrix-Vector Multiplication)**.**

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix} = \begin{bmatrix} 1(7) + 2(8) + 3(9) \\ 4(7) + 5(8) + 6(9) \end{bmatrix} = \begin{bmatrix} 50 \\ 122 \end{bmatrix}$$

Check: $(2 \times 3) \cdot (3 \times 1) = (2 \times 1)$—the result is a 2-vector.

Two Interpretations of Ax

There are two equivalent ways to think about matrix-vector multiplication:
Row interpretation: Each element of $\mathbf{y}$ is the dot product of a row of $\mathbf{A}$ with $\mathbf{x}$:

$$y_i = (\text{row } i \text{ of } \mathbf{A}) \cdot \mathbf{x}$$

Column interpretation: The result is a *linear combination* of the columns of $\mathbf{A}$, with coefficients given by $\mathbf{x}$:

$$\mathbf{Ax} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n$$

where $\mathbf{a}_j$ is the jth column of $\mathbf{A}$.
The column interpretation is particularly important for understanding linear systems and the geometry of matrix transformations.

8.5 Systems of Linear Equations

One of the most important applications of matrices is representing and solving systems of linear equations.

8.5.1 Matrix Representation

Consider a system of m linear equations in n unknowns:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

This system can be written compactly in matrix form as:

$$\mathbf{Ax} = \mathbf{b} \tag{8.5}$$

where:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

Components of a Linear System $\mathbf{Ax}$

- $\mathbf{A} \in \mathbb{R}^{m \times n}$: The **coefficient matrix** containing the coefficients of the equations
- $\mathbf{x} \in \mathbb{R}^n$: The **unknown vector** we want to solve for
- $\mathbf{b} \in \mathbb{R}^m$: The **right-hand side vector** (constants)

In regression contexts: $\mathbf{A}$ is the design matrix (data), $\mathbf{x}$ is the coefficient vector, and $\mathbf{b}$ is the response vector.

8.5.2 Geometric Interpretation

For a system of 2 equations in 2 unknowns, each equation represents a line in the plane. The solution is the intersection point. Three possibilities arise:

1. **Unique solution**: The lines intersect at exactly one point.
2. **No solution**: The lines are parallel (inconsistent system).
3. **Infinitely many solutions**: The lines are identical (dependent system).

In higher dimensions, each equation represents a hyperplane, and solutions are intersections of these hyperplanes.

8.5.3 Gaussian Elimination

Gaussian elimination is the standard algorithm for solving systems of linear equations. It systematically transforms the system into an equivalent one that is easier to solve.

Definition 8.11 (Elementary Row Operations)**.** The following operations on a matrix do not change the solution set of the corresponding linear system:

1. **Row swap**: Interchange two rows ($R_i \leftrightarrow R_j$).
2. **Row scaling**: Multiply a row by a non-zero constant ($R_i \to c \cdot R_i$, $c \neq 0$).
3. **Row addition**: Add a multiple of one row to another ($R_i \to R_i + c \cdot R_j$).

Definition 8.12 (Augmented Matrix)**.** The **augmented matrix** $[\mathbf{A} \mid \mathbf{b}]$ combines the coefficient matrix and right-hand side:

$$[\mathbf{A} \mid \mathbf{b}] = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array}\right]$$

Gaussian Elimination Algorithm

Goal: Transform the augmented matrix to **row echelon form** (REF), where:

- All rows consisting entirely of zeros are at the bottom.
- The leading entry (pivot) of each non-zero row is to the right of the pivot in the row above.
- All entries below each pivot are zero.

Forward elimination (achieving REF):

1. Start with column 1. Find the first non-zero entry (pivot). If needed, swap rows to bring it to position (1,1).
2. Use row operations to eliminate all entries below the pivot.
3. Move to column 2, row 2. Repeat: find pivot, eliminate entries below.
4. Continue until the matrix is in REF.

Back substitution: Starting from the last equation, solve for each variable and substitute back into earlier equations.

For **reduced row echelon form** (RREF), continue with:

1. Scale each pivot to 1.
2. Use row operations to eliminate all entries *above* each pivot as well.

Example 8.8 (Gaussian Elimination)**.** Solve the system:

$$\begin{aligned} x + 2y + z &= 9 \\ 2x + 4y + 3z &= 21 \\ 3x + 5y + 4z &= 25 \end{aligned}$$

Step 1: Form the augmented matrix.

$$\left[\begin{array}{ccc|c} 1 & 2 & 1 & 9 \\ 2 & 4 & 3 & 21 \\ 3 & 5 & 4 & 25 \end{array}\right]$$

Step 2: Eliminate entries below pivot in column 1.

- $R_2 \to R_2 - 2R_1$: $\left[\begin{array}{ccc|c} 1 & 2 & 1 & 9 \\ 0 & 0 & 1 & 3 \\ 3 & 5 & 4 & 25 \end{array}\right]$
- $R_3 \to R_3 - 3R_1$: $\left[\begin{array}{ccc|c} 1 & 2 & 1 & 9 \\ 0 & 0 & 1 & 3 \\ 0 & -1 & 1 & -2 \end{array}\right]$

Step 3: Swap R_2 and R_3 to get a pivot in position (2,2).

$$\left[\begin{array}{ccc|c} 1 & 2 & 1 & 9 \\ 0 & -1 & 1 & -2 \\ 0 & 0 & 1 & 3 \end{array}\right]$$

Step 4: Back substitution.

- From row 3: $z = 3$
- From row 2: $-y + z = -2 \Rightarrow -y + 3 = -2 \Rightarrow y = 5$
- From row 1: $x + 2y + z = 9 \Rightarrow x + 10 + 3 = 9 \Rightarrow x = -4$

Solution: $(x, y, z) = (-4, 5, 3)$.

8.6 The Identity Matrix and Matrix Inverses

8.6.1 The Identity Matrix

Definition 8.13 (Identity Matrix)**.** The **identity matrix** $\mathbf{I}_n$ is the $n \times n$ square matrix with 1s on the main diagonal and 0s elsewhere:

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

For example:

$$\mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The identity matrix is the matrix analogue of the number 1 in scalar arithmetic. It is the **multiplicative identity**:

Theorem 8.2 (Properties of the Identity Matrix). *For any matrix* $\mathbf{A} \in \mathbb{R}^{m \times n}$*:*

$$\mathbf{I}_m \mathbf{A} = \mathbf{A} \tag{8.6}$$
$$\mathbf{A}\mathbf{I}_n = \mathbf{A} \tag{8.7}$$

For any vector $\mathbf{x} \in \mathbb{R}^n$*:*

$$\mathbf{I}_n \mathbf{x} = \mathbf{x}$$

Multiplying by the identity matrix "does nothing"—it returns the original matrix or vector unchanged.

8.6.2 The Matrix Inverse

Definition 8.14 (Matrix Inverse). For a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, if there exists a matrix $\mathbf{A}^{-1}$ such that:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$$

then $\mathbf{A}^{-1}$ is called the **inverse** of $\mathbf{A}$, and $\mathbf{A}$ is said to be **invertible** (or **non-singular**).

Existence of the Inverse

- **Only square matrices can have inverses.** A 3×4 matrix cannot be inverted.
- **Not all square matrices have inverses.** A matrix without an inverse is called *singular* or *non-invertible*.
- A square matrix is invertible if and only if its determinant is non-zero (equivalently, if its columns/rows are linearly independent—see Chapter 9).

The inverse is theoretically important for solving linear systems. If $\mathbf{A}\mathbf{x} = \mathbf{b}$ and $\mathbf{A}$ is invertible:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

However, in practice, directly computing $\mathbf{A}^{-1}$ is often avoided because it is computationally expensive and numerically unstable. Instead, we use algorithms like Gaussian elimination or matrix factorisations.

Theorem 8.3 (Properties of the Inverse). *For invertible matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *of compatible dimensions:*

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A} \tag{8.8}$$
$$(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top \tag{8.9}$$
$$(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \quad \textit{(order reverses)} \tag{8.10}$$

8.7 Vector Norms

Norms provide a way to measure the "size" or "length" of vectors. Different norms emphasise different aspects of a vector.

Definition 8.15 (Vector Norm). A **norm** on $\mathbb{R}^n$ is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ satisfying:

1. **Non-negativity**: $\|\mathbf{x}\| \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{0}$.
2. **Absolute homogeneity**: $\|c\mathbf{x}\| = |c|\|\mathbf{x}\|$ for any scalar c.
3. **Triangle inequality**: $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

Common Vector Norms

For $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top \in \mathbb{R}^n$:
L_1 norm (Manhattan / Taxicab norm):

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| = |x_1| + |x_2| + \cdots + |x_n|$$

L_2 norm (Euclidean norm):

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

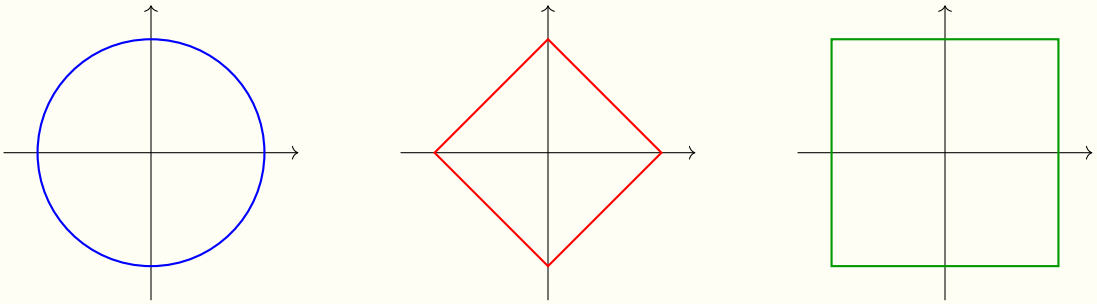
L_∞ norm (Maximum / Chebyshev norm):

$$\|\mathbf{x}\|_\infty = \max_i |x_i|$$

Geometric Interpretation of Norms

Consider the set of all vectors with norm equal to 1 (the "unit ball"). The shape of this set differs by norm:

- **L_2 norm**: The unit ball is a circle (2D) or sphere (3D)—all points at Euclidean distance 1 from the origin.
- **L_1 norm**: The unit ball is a diamond (2D) or octahedron (3D). The L_1 distance is like navigating a city grid where you can only move along axes (hence "taxicab" or "Manhattan" distance).
- **L_∞ norm**: The unit ball is a square (2D) or cube (3D). The norm equals the largest coordinate in absolute value.



L_2 L_1 L_∞

Example 8.9 (Computing Norms). For $\mathbf{x} = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$:

$$\|\mathbf{x}\|_1 = |3| + |-4| = 7$$
$$\|\mathbf{x}\|_2 = \sqrt{3^2 + (-4)^2} = \sqrt{9+16} = 5$$
$$\|\mathbf{x}\|_\infty = \max(|3|, |-4|) = 4$$

Note that the L_2 norm of $(3, -4)$ equals 5—this is the familiar 3-4-5 right triangle from Pythagorean theorem.

The choice of norm matters significantly in applications like regularisation (see Section 8.9) and optimisation.

8.8 Linear Regression in Matrix Form

Linear regression is one of the most important applications of linear algebra in data science. We now derive the ordinary least squares (OLS) estimator using matrix calculus.

8.8.1 The Linear Model

For n observations, each with p predictor variables, the linear regression model assumes:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n \tag{8.11}$$

where y_i is the response for observation i, the x_{ij} are predictor values, the β_j are unknown coefficients, and ε_i is the error term.

Matrix Form of Linear Regression

We can write the entire system compactly as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{8.12}$$

where:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Dimensions:

- $\mathbf{y} \in \mathbb{R}^n$: Response vector
- $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$: Design matrix (the column of 1s handles the intercept)
- $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$: Coefficient vector
- $\boldsymbol{\varepsilon} \in \mathbb{R}^n$: Error vector

Remark (Why the Column of Ones?). The first column of $\mathbf{X}$ consists entirely of 1s. When multiplied by β_0, this produces the constant intercept term in each equation. This trick allows us to treat the intercept uniformly with the other coefficients.

8.8.2 The Least Squares Objective

Our goal is to find the coefficient vector $\boldsymbol{\beta}$ that minimises the sum of squared errors (residuals):

$$\text{SSE} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{8.13}$$

where $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ is the predicted value.

In matrix notation, the sum of squared errors can be written elegantly:

$$\text{SSE} = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{8.14}$$

Equivalence of Scalar and Matrix SSE

To see why $\sum_i \varepsilon_i^2 = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}$, note that:

$$\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_n^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

This is the dot product of $\boldsymbol{\varepsilon}$ with itself, which equals the sum of squared elements.

8.8.3 Expanding the Objective Function

To find the minimum, we need to expand the objective function and take its derivative. Let us expand $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.

Expansion of the SSE

Using the transpose properties from Theorem 8.1:

$$\begin{aligned} \text{SSE} &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\boldsymbol{\beta} - (\mathbf{X}\boldsymbol{\beta})^\top \mathbf{y} + (\mathbf{X}\boldsymbol{\beta})^\top (\mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

For the cross terms, note that:

- $(\mathbf{X}\boldsymbol{\beta})^\top = \boldsymbol{\beta}^\top \mathbf{X}^\top$ (transpose of a product reverses order)
- $\mathbf{y}^\top \mathbf{X}\boldsymbol{\beta}$ is a scalar, and the transpose of a scalar is itself
- Therefore $\mathbf{y}^\top \mathbf{X}\boldsymbol{\beta} = (\mathbf{y}^\top \mathbf{X}\boldsymbol{\beta})^\top = \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y}$

So both cross terms are equal, and:

$$\text{SSE} = \mathbf{y}^\top \mathbf{y} - 2\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} \tag{8.15}$$

Equivalently: $\text{SSE} = \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}$

8.8.4 Deriving the OLS Estimator

To minimise SSE, we differentiate with respect to $\boldsymbol{\beta}$ and set the derivative equal to zero.

Matrix Calculus: Key Results

We need two results from matrix calculus (see the "Matrix Cookbook" for comprehensive reference):

For a vector $\mathbf{a}$ and matrices of appropriate dimensions:

1. $\frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{a}^\top \boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^\top \mathbf{a}) = \mathbf{a}$
2. $\frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^\top \mathbf{A}\boldsymbol{\beta}) = 2\mathbf{A}\boldsymbol{\beta}$ (when $\mathbf{A}$ is symmetric)

Note that $\mathbf{X}^\top\mathbf{X}$ is symmetric: $(\mathbf{X}^\top\mathbf{X})^\top = \mathbf{X}^\top(\mathbf{X}^\top)^\top = \mathbf{X}^\top\mathbf{X}$.

Differentiating (8.15) with respect to $\boldsymbol{\beta}$:

$$\begin{aligned}\frac{\partial \text{SSE}}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}}\left(\mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\right) \\ &= 0 - 2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

Setting equal to zero:

$$-2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

Rearranging:

$$\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{y} \tag{8.16}$$

These are called the **normal equations**. If $\mathbf{X}^\top\mathbf{X}$ is invertible, we can solve for $\boldsymbol{\beta}$:

OLS Estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \tag{8.17}$$

This is the **ordinary least squares (OLS) estimator**—the vector of coefficients that minimises the sum of squared errors.

Remark (When Does $(\mathbf{X}^\top\mathbf{X})^{-1}$ Exist?). The matrix $\mathbf{X}^\top\mathbf{X}$ is invertible if and only if $\mathbf{X}$ has full column rank, meaning its columns are linearly independent (see Chapter 9). This fails when:

- There are more predictors than observations ($p + 1 > n$).
- Some predictors are perfectly collinear (one is a linear combination of others).

In these cases, the OLS solution is not unique, and regularisation methods become essential.

Remark (Connection to Partial Derivatives). The matrix formulation encapsulates what would otherwise require solving $p + 1$ separate partial derivative equations. Each component of the vector equation $\frac{\partial \text{SSE}}{\partial \boldsymbol{\beta}} = \mathbf{0}$ corresponds to one partial derivative:

$$\frac{\partial \text{SSE}}{\partial \beta_j} = 0 \quad \text{for } j = 0, 1, \ldots, p$$

This is why regression coefficients have the interpretation "holding other variables constant"—each partial derivative isolates the effect of one variable.

8.9 Penalised Regression

Ordinary least squares can perform poorly when:

- The number of predictors is large relative to the sample size.
- Predictors are highly correlated (multicollinearity).
- We want a simpler, more interpretable model.

Penalised regression (also called **regularised regression**) addresses these issues by adding a penalty term to the objective function that discourages large coefficient values.

8.9.1 The General Framework

The general penalised regression objective is:

$$\min_{\boldsymbol{\beta}} \left\{ \underbrace{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}_{\text{SSE (data fit)}} + \underbrace{\lambda \cdot \text{Penalty}(\boldsymbol{\beta})}_{\text{Regularisation}} \right\} \tag{8.18}$$

where $\lambda \geq 0$ is the **regularisation parameter** (or **tuning parameter**) that controls the trade-off between fitting the data and keeping coefficients small.

The Bias-Variance Trade-off

Regularisation introduces *bias* (coefficients are systematically shrunk toward zero) in exchange for reduced *variance* (estimates are more stable across different samples). When $\lambda = 0$, we recover OLS. As $\lambda \to \infty$, all coefficients shrink toward zero.
The optimal λ balances these effects, typically chosen via cross-validation.

8.9.2 Ridge Regression (L_2 Penalty)

Definition 8.16 (Ridge Regression)**. Ridge regression** uses the L_2 norm (squared) as the penalty:

$$\min_{\boldsymbol{\beta}} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{8.19}$$

Note: The intercept β_0 is typically *not* penalised, though the penalty is sometimes written as $\lambda \|\boldsymbol{\beta}\|_2^2 = \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}$ for simplicity.

Derivation of the Ridge Estimator

Writing the penalty in matrix form as $\lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}$, the objective becomes:

$$L(\boldsymbol{\beta}) = \mathbf{y}^\top \mathbf{y} - 2\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}$$

Taking the derivative with respect to $\boldsymbol{\beta}$:

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} + 2\lambda \boldsymbol{\beta}$$

Setting to zero:

$$-2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta} = \mathbf{0}$$

Rearranging:

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{y}$$

Ridge Estimator

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} \tag{8.20}$$

Compare with OLS: $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$
The only difference is the addition of $\lambda\mathbf{I}$ to $\mathbf{X}^\top\mathbf{X}$.

Remark (Ridge Solves the Invertibility Problem). Adding $\lambda\mathbf{I}$ to $\mathbf{X}^\top\mathbf{X}$ guarantees invertibility for any $\lambda > 0$, even when $\mathbf{X}^\top\mathbf{X}$ itself is singular. This is because adding a positive constant to the diagonal eigenvalues ensures they are all positive.

What Does Ridge Do?

Ridge regression:

- **Shrinks** all coefficients toward zero, but never exactly to zero.
- Works well when many predictors contribute small effects.
- Handles multicollinearity by stabilising coefficient estimates.
- Does **not** perform variable selection—all predictors remain in the model.

8.9.3 Lasso Regression (L_1 Penalty)

Definition 8.17 (Lasso Regression)**.** The **Lasso** (Least Absolute Shrinkage and Selection Operator) uses the L_1 norm as the penalty:

$$\min_{\boldsymbol{\beta}} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{8.21}$$

No Closed-Form Solution

Unlike Ridge regression, Lasso does *not* have a closed-form solution because the L_1 penalty $|\beta_j|$ is not differentiable at $\beta_j = 0$. Lasso must be solved using iterative optimisation algorithms such as coordinate descent or LARS (Least Angle Regression).

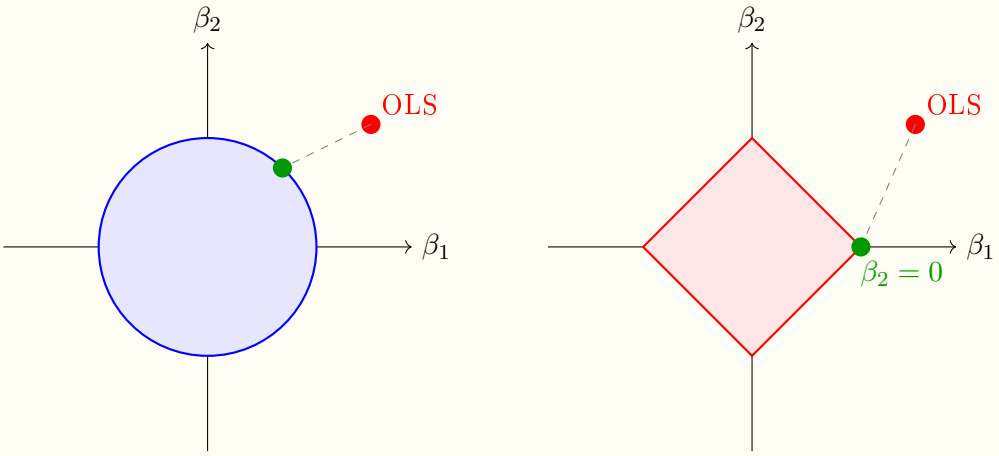
Why Does Lasso Produce Sparse Solutions?

The L_1 penalty has the remarkable property of setting some coefficients *exactly to zero*, effectively performing automatic variable selection.
Geometric explanation: Consider the constraint region (the set of $\boldsymbol{\beta}$ satisfying the penalty constraint) in 2D:

- For L_2 (Ridge): The constraint region is a circle.

- For L_1 (Lasso): The constraint region is a diamond with corners on the axes.

The OLS solution lies somewhere in $\boldsymbol{\beta}$-space. We seek the point in the constraint region closest to this OLS solution (in the sense of minimising SSE). With the diamond-shaped L_1 constraint, this optimal point is likely to land at a corner—where one or more coordinates are exactly zero.



Ridge (L_2) Lasso (L_1)

The Lasso solution lands at a corner where $\beta_2 = 0$, demonstrating automatic variable selection.

Ridge vs Lasso: Summary

Property	Ridge (L_2)	Lasso (L_1)
Penalty	$\lambda \sum_j \beta_j^2$	$\lambda \sum_j \lvert\beta_j\rvert$
Closed-form solution	Yes	No
Shrinks coefficients	To small values	To exactly zero
Variable selection	No	Yes
Handles multicollinearity	Well	Arbitrarily selects one
Good when	Many small effects	Few large effects

8.9.4 Elastic Net ($L_1 + L_2$ Penalty)

Definition 8.18 (Elastic Net)**. Elastic Net** combines the L_1 and L_2 penalties:

$$\min_{\boldsymbol{\beta}} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\} \tag{8.22}$$

Equivalently, with a single λ and mixing parameter $\alpha \in [0, 1]$:

$$\min_{\boldsymbol{\beta}} \left\{ \text{SSE} + \lambda \left[\alpha \|\boldsymbol{\beta}\|_1 + (1 - \alpha) \|\boldsymbol{\beta}\|_2^2 \right] \right\} \tag{8.23}$$

where $\alpha = 1$ gives Lasso, $\alpha = 0$ gives Ridge, and intermediate values give a blend.

Why Elastic Net?

Elastic Net addresses limitations of Lasso:

- **Grouped selection**: When predictors are highly correlated, Lasso tends to select one arbitrarily and ignore others. Elastic Net tends to select groups of correlated variables together.
- **Stability**: The L_2 component stabilises the solution when $p > n$.
- **Flexibility**: The mixing parameter α allows tuning the balance between sparsity (Lasso-like) and grouping (Ridge-like).

8.9.5 Choosing the Regularisation Parameter

The regularisation parameter λ (and α for Elastic Net) must be chosen carefully:

- **Too small** λ: Minimal regularisation, overfitting risk.
- **Too large** λ: Excessive shrinkage, underfitting risk.

The standard approach is **cross-validation**:

1. Split data into k folds (typically $k = 5$ or 10).
2. For each candidate λ, train on $k - 1$ folds and evaluate on the held-out fold.
3. Repeat for all folds and average the prediction error.
4. Select the λ with lowest cross-validated error (or the largest λ within one standard error of the minimum—the "one-standard-error rule").

8.10 Summary and Connections

Week 8 Key Results

Data structures:

- Scalar ($\mathbb{R}$), Vector ($\mathbb{R}^n$), Matrix ($\mathbb{R}^{m\times n}$), Tensor (higher-order arrays)

Matrix operations:

- Transpose: $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$
- Multiplication: $(m \times n)(n \times p) = (m \times p)$; inner dimensions must match

Linear systems: $\mathbf{Ax} = \mathbf{b}$ solved via Gaussian elimination
OLS estimator:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Ridge estimator:

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

Lasso: No closed form; promotes sparsity via L_1 penalty

Looking ahead: In Chapter 9, we will explore deeper concepts including linear independence, rank, determinants, eigenvalues, and singular value decomposition—tools that reveal the geometric and algebraic structure of matrices.

Chapter 9

Linear Algebra II

Learning Objectives

After completing this chapter, you should be able to:

1. Determine whether a set of vectors is linearly dependent or independent
2. Understand span, basis, and the dimension of a vector space
3. Calculate determinants for 2×2 and 3×3 matrices
4. Identify when a matrix is invertible and compute inverses
5. Find eigenvalues and eigenvectors of a matrix
6. Perform eigendecomposition and understand its conditions
7. Apply Singular Value Decomposition (SVD) to non-square matrices
8. Connect PCA to eigendecomposition of the covariance matrix

Prerequisites

This chapter assumes familiarity with:

- Matrix multiplication and transpose operations (Chapter **??**)
- Systems of linear equations
- Basic vector operations (addition, scalar multiplication)
- The concept of a vector space

9.1 Linear Dependence and Independence

Linear dependence captures the idea of *redundant information*: at least one vector in a set can be expressed as a linear combination of the others. Conversely, a set of vectors is **linearly independent** if no vector in the set is redundant—each provides unique directional information.

Definition 9.1 (Linear Dependence)**.** A set of vectors $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is **linearly dependent** if and only if there exist scalars $c_1, c_2, \ldots, c_n$, *not all zero*, such that:

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \mathbf{0} \tag{9.1}$$

This is called a **non-trivial solution** to the homogeneous equation.

Definition 9.2 (Linear Independence)**.** A set of vectors $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is **linearly independent** if the *only* solution to Equation (9.1) is the **trivial solution**: $c_1 = c_2 = \cdots = c_n = 0$.

Key Insight

Linear Dependence: We can find non-zero coefficients that make a weighted sum of vectors equal zero. This means at least one vector can be "cancelled out" by the others—it lies in the same subspace they span.
Linear Independence: The only way to get zero is to set all coefficients to zero. No vector can be expressed as a combination of the others—each points in a genuinely new direction.

9.1.1 Geometric Intuition

The geometric interpretation makes linear dependence intuitive:

- **Two vectors in $\mathbb{R}^2$**: They are linearly dependent if and only if they are *collinear* (one is a scalar multiple of the other). They point along the same line.
- **Three vectors in $\mathbb{R}^3$**: They are linearly dependent if they are *coplanar*—all three lie in the same plane. The third vector doesn't "lift off" into a new dimension.
- **General principle**: Linearly dependent vectors fail to span as many dimensions as there are vectors. They have redundancy.

Example 9.1 (Linearly Dependent Vectors)**.** Consider the vectors:

$$\mathbf{v}_1 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$$

These are linearly dependent because $\mathbf{v}_2 = 2\mathbf{v}_1$. We can write:

$$2\mathbf{v}_1 - 1\mathbf{v}_2 = \mathbf{0}$$

with $c_1 = 2$ and $c_2 = -1$, both non-zero.

Example 9.2 (Linearly Independent Vectors)**.** Consider:

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$$

These are linearly independent. The equation $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 = \mathbf{0}$ gives:

$$\begin{bmatrix} 4c_2 \\ 3c_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

which requires $c_1 = 0$ and $c_2 = 0$. No scalar multiple of one vector can produce the other—they point in perpendicular directions.

Example 9.3 (Adding a Redundant Vector)**.** Now consider adding a third vector to the independent set:

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 4 \\ 0 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

This set is **linearly dependent**. Why? In $\mathbb{R}^2$, you can have at most 2 linearly independent vectors. The third vector $\mathbf{v}_3$ must be expressible as a combination of $\mathbf{v}_1$ and $\mathbf{v}_2$:

$$\mathbf{v}_3 = \frac{2}{3}\mathbf{v}_1 + \frac{1}{2}\mathbf{v}_2$$

Partial Dependence Implies Full Dependence

If *any* subset of vectors is linearly dependent, then the entire set is linearly dependent. For a matrix, if even one row (or column) can be expressed as a linear combination of others, the entire matrix has dependent rows (or columns).

9.1.2 Proving Linear (In)dependence

Procedure for Testing Linear Dependence

1. Set up the equation: $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \mathbf{0}$
2. Write this as a system of linear equations (one equation per component)
3. Solve the system for $c_1, c_2, \ldots, c_n$
4. If the *only* solution is $c_1 = c_2 = \cdots = c_n = 0$: **linearly independent**
5. If a non-trivial solution exists: **linearly dependent**

Example: Proving Linear Dependence

Example 9.4 (Finding a Non-Trivial Solution)**.** Consider:

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} -1 \\ -2 \\ -3 \end{bmatrix}$$

Step 1: Set up $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = \mathbf{0}$:

$$\begin{aligned} c_1 + 2c_2 - c_3 &= 0 \\ 2c_1 + 4c_2 - 2c_3 &= 0 \\ 3c_1 + 6c_2 - 3c_3 &= 0 \end{aligned}$$

Step 2: Observe that every equation is a multiple of the first. This is a homogeneous system with infinitely many solutions.

Step 3: Find a non-trivial solution. Let $c_2 = 1$ and $c_3 = 0$. Then from the first equation: $c_1 + 2(1) - 0 = 0$, so $c_1 = -2$.

Conclusion: The solution $(c_1, c_2, c_3) = (-2, 1, 0)$ is non-trivial, so the vectors are **linearly dependent**.

Geometric observation: $\mathbf{v}_2 = 2\mathbf{v}_1$ and $\mathbf{v}_3 = -\mathbf{v}_1$. All three vectors lie on the same line through the origin.

Example: Proving Linear Independence

Example 9.5 (Only Trivial Solution Exists)**.** Consider the standard basis vectors in $\mathbb{R}^3$:

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Step 1: Set up $c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + c_3\mathbf{e}_3 = \mathbf{0}$:

$$\begin{aligned} c_1 &= 0 \\ c_2 &= 0 \\ c_3 &= 0 \end{aligned}$$

Conclusion: The only solution is the trivial solution, so the vectors are **linearly independent**.

These vectors point along the three coordinate axes—each in a completely different direction. No combination of two can produce the third.

Example: Mixed Case

Example 9.6 (Detecting Partial Dependence)**.** Consider:

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 2 \\ 6 \\ 8 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

Setting up $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = \mathbf{0}$:

$$\begin{aligned} c_1 + 2c_2 &= 0 \\ 3c_1 + 6c_2 + c_3 &= 0 \\ 4c_1 + 8c_2 + 2c_3 &= 0 \end{aligned}$$

From the first equation: $c_1 = -2c_2$. Substituting into the second: $3(-2c_2) + 6c_2 + c_3 = 0$, giving $c_3 = 0$. The third equation then holds for any c_2.

Taking $c_2 = 1$ gives $(c_1, c_2, c_3) = (-2, 1, 0)$.

Conclusion: Linearly dependent. Notice that $\mathbf{v}_2 = 2\mathbf{v}_1$, making $\mathbf{v}_1$ and $\mathbf{v}_2$ dependent regardless of $\mathbf{v}_3$.

9.2 Span and Basis

Definition 9.3 (Span)**.** The **span** of a set of vectors $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is the set of all possible linear combinations of those vectors:

$$\text{span}(S) = \{a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n : a_1, a_2, \ldots, a_n \in \mathbb{R}\}$$

The span represents all points "reachable" by scaling and adding the vectors in S.

Example 9.7 (Span in $\mathbb{R}^2$)**.** If $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, then:

$$\text{span}(\{\mathbf{v}_1, \mathbf{v}_2\}) = \mathbb{R}^2$$

Any point (x, y) in the plane can be written as $x\mathbf{v}_1 + y\mathbf{v}_2$.

Definition 9.4 (Spanning Set)**.** A set S is a **spanning set** for a vector space V if every vector in V can be expressed as a linear combination of vectors in S. We say "S spans V".

Definition 9.5 (Basis)**.** A **basis** for a vector space V is a set of vectors that is:

1. **Linearly independent**: No redundancy
2. **Spanning**: Covers all of V

A basis is a *minimal* spanning set—removing any vector would make it fail to span V.

Definition 9.6 (Dimension)**.** The **dimension** of a vector space V, denoted $\dim(V)$, is the number of vectors in any basis for V. This is well-defined because all bases for V have the same size.

Fundamental Facts About Dimension

- $\mathbb{R}^n$ has dimension n
- Any set of more than n vectors in $\mathbb{R}^n$ must be linearly dependent
- Any basis for $\mathbb{R}^n$ contains exactly n vectors
- The standard basis for $\mathbb{R}^n$ is $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ where $\mathbf{e}_i$ has a 1 in position i and 0s elsewhere

Why Exactly n Independent Vectors in $\mathbb{R}^n$?

Think geometrically in $\mathbb{R}^3$:

- One vector spans a *line* (1D)
- Two independent vectors span a *plane* (2D)
- Three independent vectors span all of *space* (3D)
- A fourth vector must lie in the 3D space already spanned—it's redundant

Each independent vector adds one new dimension. Once you have n independent vectors in $\mathbb{R}^n$, you've filled all available dimensions.

9.3 The Determinant

The **determinant** is a scalar value computed from a square matrix that encodes essential information about the matrix's properties.

Definition 9.7 (Determinant)**.** For a square matrix A, the determinant, denoted $\det(A)$ or $|A|$, is a scalar that can be computed from the matrix entries according to specific rules.

9.3.1 Geometric Interpretation

Before diving into formulas, understanding what the determinant *means* is valuable:

Geometric Meaning of the Determinant

The determinant of a matrix A represents:

- In 2D: The **signed area** of the parallelogram formed by the column vectors
- In 3D: The **signed volume** of the parallelepiped formed by the column vectors
- In general: The factor by which A scales volumes under the linear transformation it represents

The sign indicates whether the transformation preserves orientation ($\det > 0$) or reverses it ($\det < 0$).

9.3.2 Applications of the Determinant

1. **Linear Transformations**: $|\det(A)|$ gives the volume scaling factor. A negative determinant indicates the transformation includes a reflection.
2. **Invertibility**: A matrix is invertible if and only if $\det(A) \neq 0$.
3. **Linear Independence**: Columns (or rows) are linearly independent if and only if $\det(A) \neq 0$.
4. **Eigenvalues**: The characteristic polynomial, used to find eigenvalues, involves the determinant (see Section 9.5).
5. **Area/Volume Calculations**: The determinant directly computes areas and volumes of geometric shapes.
6. **Solving Systems**: Cramer's rule uses determinants to solve systems of linear equations.

9.3.3 Computing Determinants

1×1 Determinant

For $A = [a]$:

$$\det(A) = a$$

2×2 Determinant

For $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$:

$$\det(A) = ad - bc$$

Memory aid: Product of main diagonal minus product of anti-diagonal.

Example 9.8 (2×2 Determinant)**.**

$$\det \begin{bmatrix} 3 & 2 \\ 1 & 4 \end{bmatrix} = (3)(4) - (2)(1) = 12 - 2 = 10$$

This tells us the parallelogram formed by vectors $\begin{bmatrix} 3 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 2 \\ 4 \end{bmatrix}$ has area 10.

3×3 Determinant (Expansion by First Row)

For $A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$:

$$\det(A) = a(ei - fh) - b(di - fg) + c(dh - eg)$$

The pattern is: take each element of the first row, multiply by the determinant of the 2×2 **minor** (the submatrix obtained by deleting that element's row and column), and alternate signs: $+, -, +$.

Example 9.9 (3×3 Determinant)**.**

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

$$\begin{aligned} \det(A) &= 1 \cdot \det \begin{bmatrix} 5 & 6 \\ 8 & 9 \end{bmatrix} - 2 \cdot \det \begin{bmatrix} 4 & 6 \\ 7 & 9 \end{bmatrix} + 3 \cdot \det \begin{bmatrix} 4 & 5 \\ 7 & 8 \end{bmatrix} \\ &= 1(45 - 48) - 2(36 - 42) + 3(32 - 35) \\ &= 1(-3) - 2(-6) + 3(-3) \\ &= -3 + 12 - 9 = 0 \end{aligned}$$

The determinant is zero, indicating the columns are linearly dependent. Indeed, the third column equals the average of the first two columns scaled by 2.

General Formula: Cofactor Expansion

For an $n \times n$ matrix, the determinant can be computed by expanding along any row i:

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} M_{ij}$$

where M_{ij} is the **minor**: the determinant of the $(n-1) \times (n-1)$ submatrix obtained by deleting row i and column j.

The term $(-1)^{i+j}$ creates the alternating sign pattern:

$$\begin{bmatrix} + & - & + & - & \cdots \\ - & + & - & + & \cdots \\ + & - & + & - & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Computational Complexity

Computing determinants via cofactor expansion has factorial complexity $O(n!)$, making it impractical for large matrices. In practice, row reduction to triangular form (where the determinant is the product of diagonal entries) is used, giving $O(n^3)$ complexity.

9.4 Matrix Inverse

Definition 9.8 (Matrix Inverse)**.** For a square matrix A, its **inverse** A^{-1} (if it exists) satisfies:

$$A^{-1}A = AA^{-1} = I$$

where I is the identity matrix.

9.4.1 Conditions for Invertibility

A Square Matrix A is Invertible if and only if:

1. $\det(A) \neq 0$
2. The columns of A are linearly independent
3. The rows of A are linearly independent
4. A has full rank (rank equals the matrix dimension)
5. The only solution to $A\mathbf{x} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$

All these conditions are equivalent.

The logical chain is:

$$\text{Linear Independence} \iff \det(A) \neq 0 \iff \text{Invertible}$$

Why These Conditions Are Connected

If a matrix has linearly dependent columns, the transformation it represents "collapses" some dimension—information is lost. You cannot reverse a process that loses information, hence no inverse exists.
Geometrically, a zero determinant means the transformation squashes n-dimensional space into a lower-dimensional subspace (a plane, line, or point). Such a transformation cannot be undone.

9.4.2 Computing the Inverse

2×2 Inverse Formula

For $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ with $\det(A) = ad - bc \neq 0$:

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Recipe: Swap diagonal entries, negate off-diagonal entries, divide by determinant.

Example 9.10 (Computing a 2×2 Inverse)**.** For $A = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$:

Step 1: Compute $\det(A) = (3)(3) - (0)(0) = 9 \neq 0$. The inverse exists.

Step 2: Apply the formula:

$$A^{-1} = \frac{1}{9}\begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \end{bmatrix}$$

Verification:

$$\begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I \quad \checkmark$$

Example 9.11 (Non-Invertible Matrix)**.** For $A = \begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix}$:

$$\det(A) = (1)(0) - (0)(2) = 0$$

The determinant is zero, so A is **not invertible**. Notice that the second column is all zeros—it's linearly dependent with any vector (being the zero vector scaled).

General Inverse Formula

For an $n \times n$ matrix:

$$A^{-1} = \frac{1}{\det(A)}\text{adj}(A)$$

where $\text{adj}(A)$ is the **adjugate matrix**—the transpose of the matrix of cofactors. The cofactor $C_{ij} = (-1)^{i+j}M_{ij}$ where M_{ij} is the (i, j) minor.

In practice, Gaussian elimination is used to compute inverses efficiently.

9.4.3 Proof: Linearly Dependent Matrix Has No Inverse

Proof by Contradiction

Claim: If A has linearly dependent columns, then A is not invertible.

Proof: Suppose A has linearly dependent columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$. By definition of linear dependence, there exists a non-zero vector $\mathbf{c} = (c_1, \ldots, c_n)^T$ such that:

$$c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_n\mathbf{a}_n = \mathbf{0}$$

This can be written as $A\mathbf{c} = \mathbf{0}$.

Now suppose, for contradiction, that A is invertible. Then we could multiply both sides by A^{-1}:

$$A^{-1}(A\mathbf{c}) = A^{-1}\mathbf{0} \implies I\mathbf{c} = \mathbf{0} \implies \mathbf{c} = \mathbf{0}$$

But this contradicts our assumption that $\mathbf{c} \neq \mathbf{0}$.

Therefore, A cannot be invertible. ■

9.5 Eigenvalues and Eigenvectors

Eigenvalues and eigenvectors reveal the fundamental "directions" and "scaling factors" of a linear transformation.

Definition 9.9 (Eigenvector and Eigenvalue)**.** For a square matrix A, a non-zero vector $\mathbf{v}$ is an **eigenvector** of A if:

$$A\mathbf{v} = \lambda\mathbf{v} \tag{9.2}$$

for some scalar λ. The scalar λ is called the corresponding **eigenvalue**.

Key Insight

An eigenvector is a direction that is *preserved* by the transformation A—the vector may be stretched or compressed (by factor λ) but its direction doesn't change.
The eigenvalue-eigenvector pair effectively "summarises" the matrix's action along one direction: the complex matrix multiplication $A\mathbf{v}$ reduces to simple scalar multiplication $\lambda\mathbf{v}$.

The Meaning of "Eigen"

"Eigen" is German for "own" or "characteristic". Eigenvectors are the matrix's "own" or "characteristic" directions—the directions intrinsic to the transformation, along which its behaviour is simplest.

Example 9.12 (Verifying an Eigenvector)**.** For $A = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$:

$$A\mathbf{v} = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1+4 \\ 4+6 \end{bmatrix} = \begin{bmatrix} 5 \\ 10 \end{bmatrix} = 5 \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 5\mathbf{v}$$

Thus $\mathbf{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ is an eigenvector with eigenvalue $\lambda = 5$.

9.5.1 Finding Eigenvalues: The Characteristic Equation

How do we find eigenvalues systematically? We derive the **characteristic equation**.

Derivation of the Characteristic Equation

Starting from $A\mathbf{v} = \lambda\mathbf{v}$, rearrange:

$$\begin{aligned} A\mathbf{v} - \lambda\mathbf{v} &= \mathbf{0} \\ A\mathbf{v} - \lambda I\mathbf{v} &= \mathbf{0} \quad (\text{since } \lambda\mathbf{v} = \lambda I\mathbf{v}) \\ (A - \lambda I)\mathbf{v} &= \mathbf{0} \end{aligned}$$

This is a homogeneous system. We want a *non-trivial* solution (since eigenvectors must be non-zero).

A homogeneous system $(A - \lambda I)\mathbf{v} = \mathbf{0}$ has non-trivial solutions if and only if $(A - \lambda I)$ is singular (non-invertible), which occurs if and only if:

$$\det(A - \lambda I) = 0 \tag{9.3}$$

This is the **characteristic equation** of A.

The Characteristic Equation

$$\det(A - \lambda I) = 0$$

Solutions to this equation are the eigenvalues of A.
Expanding the determinant yields the **characteristic polynomial**, a polynomial of degree n in λ for an $n \times n$ matrix. An $n \times n$ matrix has exactly n eigenvalues (counting multiplicities, and including complex eigenvalues).

Example 9.13 (Finding Eigenvalues)**.** For $A = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$:

Step 1: Form $A - \lambda I$:

$$A - \lambda I = \begin{bmatrix} 1 - \lambda & 2 \\ 4 & 3 - \lambda \end{bmatrix}$$

Step 2: Compute the determinant and set to zero:

$$\det(A - \lambda I) = (1 - \lambda)(3 - \lambda) - (2)(4) = 0$$

$$3 - \lambda - 3\lambda + \lambda^2 - 8 = 0$$

$$\lambda^2 - 4\lambda - 5 = 0$$

Step 3: Solve the characteristic polynomial:

$$(\lambda - 5)(\lambda + 1) = 0$$

giving $\lambda_1 = 5$ and $\lambda_2 = -1$.

9.5.2 Finding Eigenvectors

Once eigenvalues are known, eigenvectors are found by solving $(A - \lambda I)\mathbf{v} = \mathbf{0}$ for each eigenvalue.

Example 9.14 (Finding Eigenvectors)**.** Continuing from Example 9.13 with $A = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$:

For $\lambda_1 = 5$:

$$(A - 5I)\mathbf{v} = \begin{bmatrix} -4 & 2 \\ 4 & -2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

From the first row: $-4v_1 + 2v_2 = 0 \implies v_2 = 2v_1$.

Taking $v_1 = 1$ gives $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

For $\lambda_2 = -1$:

$$(A + I)\mathbf{v} = \begin{bmatrix} 2 & 2 \\ 4 & 4 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

From the first row: $2v_1 + 2v_2 = 0 \implies v_2 = -v_1$.

Taking $v_1 = 1$ gives $\mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

9.5.3 The Eigenspace

Definition 9.10 (Eigenspace)**.** The **eigenspace** E_λ corresponding to eigenvalue λ is the set of all vectors satisfying $A\mathbf{v} = \lambda\mathbf{v}$:

$$E_\lambda = \{\mathbf{v} : A\mathbf{v} = \lambda\mathbf{v}\} = \ker(A - \lambda I)$$

This is a subspace of $\mathbb{R}^n$ that includes the zero vector and all eigenvectors for λ, plus all their linear combinations.

Note that while the zero vector is in every eigenspace, we do not call it an eigenvector (by convention, eigenvectors must be non-zero).

Example 9.15 (Eigenspace)**.** For $A = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$ with $\lambda = 5$:

$$E_5 = \text{span}\left\{\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right\}$$

This is the line through the origin in the direction of $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Every non-zero vector on this line is an eigenvector for $\lambda = 5$.

Eigenvectors Are Not Unique

For a given eigenvalue, any non-zero scalar multiple of an eigenvector is also an eigenvector. The eigenvector specifies a *direction*, not a specific vector. We often normalise eigenvectors to have unit length.

9.6 Eigendecomposition

When a matrix has enough linearly independent eigenvectors, it can be decomposed into a particularly elegant form.

Definition 9.11 (Eigendecomposition (Diagonalisation))**.** A square matrix A is **diagonalisable** if it can be written as:

$$A = Q\Lambda Q^{-1} \tag{9.4}$$

where:

- Λ is a **diagonal matrix** with the eigenvalues of A on the diagonal
- Q is a matrix whose columns are the corresponding eigenvectors of A

Conditions for Eigendecomposition

An $n \times n$ matrix A is diagonalisable if and only if it has n linearly independent eigenvectors. Sufficient (but not necessary) conditions:

- A has n distinct eigenvalues (guarantees n independent eigenvectors)
- A is symmetric ($A = A^T$) — always diagonalisable with orthogonal eigenvectors

Example 9.16 (Eigendecomposition). For $A = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$ with eigenvalues $\lambda_1 = 5$, $\lambda_2 = -1$ and eigenvectors $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$:

$$A = \underbrace{\begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} 5 & 0 \\ 0 & -1 \end{bmatrix}}_{\Lambda} \underbrace{\begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix}^{-1}}_{Q^{-1}}$$

The columns of Q are the eigenvectors; the diagonal of Λ contains the eigenvalues in corresponding order.

Why Eigendecomposition?

Eigendecomposition reveals the "essence" of a matrix:

- Q^{-1} rotates to the eigenvector coordinate system
- Λ performs simple scaling along each eigenvector direction
- Q rotates back to the original coordinate system

This makes many computations vastly simpler. For instance:

$$A^k = Q\Lambda^k Q^{-1}$$

Computing Λ^k is trivial: just raise each diagonal entry to the power k.

9.7 Singular Value Decomposition

Eigendecomposition only applies to square matrices. **Singular Value Decomposition (SVD)** extends this idea to *any* matrix, including rectangular ones.

Definition 9.12 (Singular Value Decomposition). For any $m \times n$ matrix A, the SVD is:

$$A = U\Sigma V^T \tag{9.5}$$

where:

- U is an $m \times m$ orthogonal matrix (columns are **left singular vectors**)
- Σ is an $m \times n$ diagonal matrix (diagonal entries are **singular values**, non-negative and typically ordered $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$)
- V^T is an $n \times n$ orthogonal matrix (rows are **right singular vectors**)

Where Do U

- The columns of U are eigenvectors of AA^T
- The columns of V are eigenvectors of A^TA
- The singular values in Σ are the **square roots** of the eigenvalues of A^TA (equivalently, AA^T)

The key insight: while A may not be square, both AA^T and A^TA are square and symmetric, so they can be eigendecomposed.

Geometric Interpretation of SVD

SVD decomposes any linear transformation into three steps:

1. V^T: Rotate/reflect in the domain (input space)
2. Σ: Scale along orthogonal axes (possibly changing dimension)
3. U: Rotate/reflect in the codomain (output space)

The singular values tell you *how much* the transformation stretches along each direction. Large singular values indicate directions of high "importance"; small or zero singular values indicate directions that are compressed or eliminated.

9.7.1 SVD Reveals Matrix Structure

Information from SVD

- **Rank**: The number of non-zero singular values equals the rank of A—the number of linearly independent columns (or rows)
- **Redundancy**: Many small singular values suggest the data lies in a lower-dimensional subspace; the matrix can be well-approximated by a lower-rank matrix
- **Condition number**: The ratio σ_1/σ_n indicates numerical stability; large ratios mean the matrix is nearly singular

Example 9.17 (SVD and Rank)**.** If a 5×3 matrix A has SVD with $\Sigma = \text{diag}(4, 2, 0)$:

- $\text{rank}(A) = 2$ (two non-zero singular values)
- The columns of A span a 2-dimensional subspace
- The third column is a linear combination of the first two

9.8 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a fundamental dimensionality reduction technique. It finds new variables (principal components) that capture the maximum variance in the data. Remarkably, PCA is equivalent to eigendecomposition of the covariance matrix.

9.8.1 Motivation: Dimensionality Reduction

Consider a data matrix X with n observations and p variables. Often:

- Many variables are correlated (redundant information)
- We want to summarise the data with fewer variables

- We want to visualise high-dimensional data

The PCA Question

If we could keep only *one* number per observation, which linear combination of variables would preserve the most information?
Answer: The combination that maximises variance. High variance means observations are spread out—there's more "signal" to distinguish them.

9.8.2 PCA as Variance Maximisation

Let X be an $n \times p$ data matrix, assumed to be **centred** (column means are zero).

Definition 9.13 (Principal Component)**.** The **first principal component** is a linear combination of the columns of X:

$$\mathbf{z}_1 = \phi_{11}x_1 + \phi_{21}x_2 + \cdots + \phi_{p1}x_p = X\boldsymbol{\phi}_1$$

where $\boldsymbol{\phi}_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})^T$ is chosen to maximise the variance of $\mathbf{z}_1$.

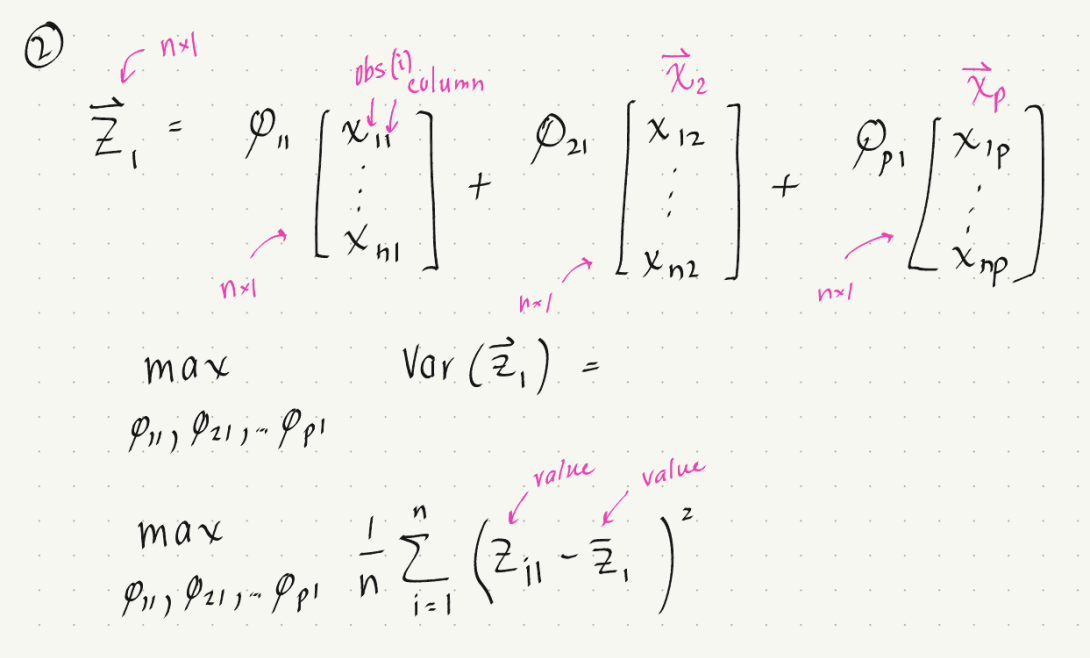


Figure 9.1: Visual representation of forming a principal component as a linear combination of original variables. Each variable x_j is scaled by its loading ϕ_{j1} and the scaled vectors are summed element-wise to produce $\mathbf{z}_1$.

Each element z_{i1} of the principal component vector is:

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip}$$

This is a weighted sum of observation i's values across all p variables.

The Variance Maximisation Problem

We seek loadings $\boldsymbol{\phi}_1$ that maximise:

$$\begin{aligned}\text{Var}(\mathbf{z}_1) &= \frac{1}{n}\sum_{i=1}^{n}(z_{i1} - \bar{z}_1)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} z_{i1}^2 \quad \text{(since data is centred, } \bar{z}_1 = 0)\end{aligned}$$

To prevent unbounded solutions (we could make variance arbitrarily large by scaling $\boldsymbol{\phi}_1$), we impose the constraint:

$$\|\boldsymbol{\phi}_1\|^2 = \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

This is a constrained optimisation problem solvable via Lagrange multipliers (see Chapter **??**), but there's an elegant alternative: eigendecomposition.

9.8.3 PCA via Eigendecomposition of the Covariance Matrix

Definition 9.14 (Sample Covariance Matrix)**.** For centred data matrix X, the sample covariance matrix is:

$$S = \frac{1}{n}X^T X$$

This is a $p \times p$ symmetric matrix where:

- Diagonal entries s_{jj} are the variances of each variable
- Off-diagonal entries s_{jk} are the covariances between variables j and k

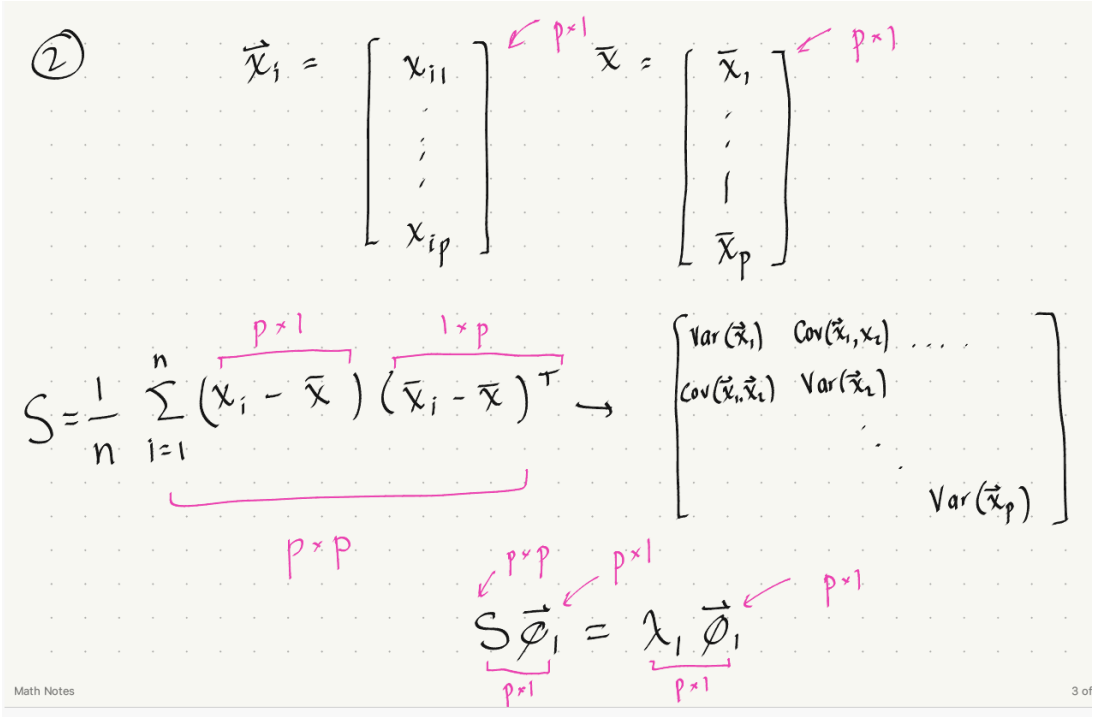


Figure 9.2: Structure of the covariance matrix. Diagonal elements represent variances of individual variables; off-diagonal elements capture pairwise covariances, measuring how variables move together.

Theorem 9.1 (PCA via Eigendecomposition)**.** *The principal components of X are given by the eigenvectors of the covariance matrix S. Specifically:*

- *The first principal component loadings $\boldsymbol{\phi}_1$ are the eigenvector corresponding to the **largest** eigenvalue λ_1*
- *The k-th principal component loadings $\boldsymbol{\phi}_k$ are the eigenvector corresponding to the k-th largest eigenvalue λ_k*
- *The variance explained by component k equals the eigenvalue λ_k*

PCA Algorithm via Eigendecomposition

1. Centre the data: subtract column means from X
2. Compute the covariance matrix: $S = \frac{1}{n} X^T X$
3. Find eigenvalues and eigenvectors of S: $S\boldsymbol{\phi}_k = \lambda_k \boldsymbol{\phi}_k$
4. Order eigenvectors by decreasing eigenvalue
5. Principal components: $\mathbf{z}_k = X\boldsymbol{\phi}_k$

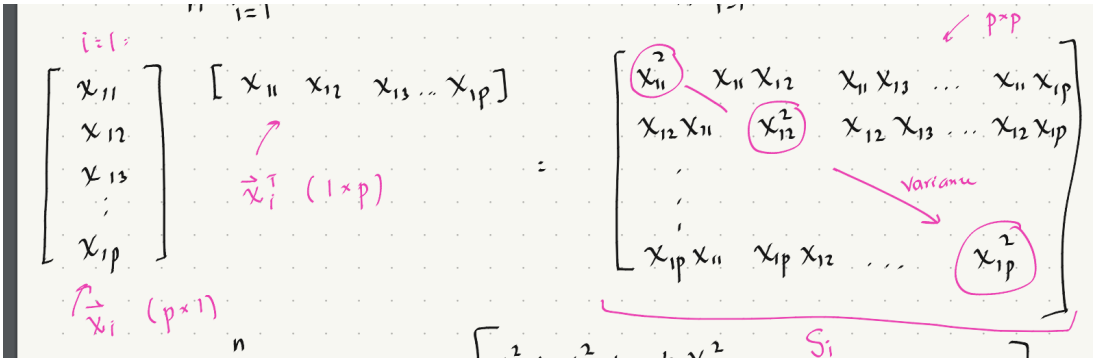


Figure 9.3: Eigendecomposition of the covariance matrix yields the principal component loadings (eigenvectors) and the variance explained by each component (eigenvalues).

Why Does Eigendecomposition Solve the Maximisation Problem?

The covariance matrix S encodes how variables vary together. Its eigenvectors point in directions of "natural variation" in the data:

- The eigenvector for the largest eigenvalue points in the direction of maximum variance
- Subsequent eigenvectors point in directions of decreasing variance, orthogonal to previous ones

Since S is symmetric, its eigenvectors are orthogonal. This means principal components are **uncorrelated**—each captures genuinely new information.

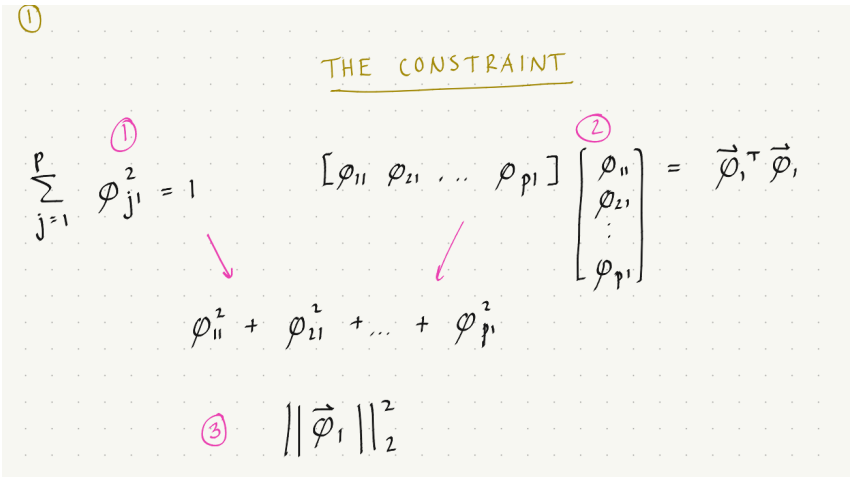


Figure 9.4: Geometric interpretation of PCA. The first principal component (PC1) aligns with the direction of maximum variance in the data cloud. The second principal component (PC2) is orthogonal to PC1 and captures the most remaining variance.

9.8.4 Computing Principal Components

Once we have the loadings (eigenvectors), the principal component scores are:

$$\mathbf{z}_k = X\boldsymbol{\phi}_k$$

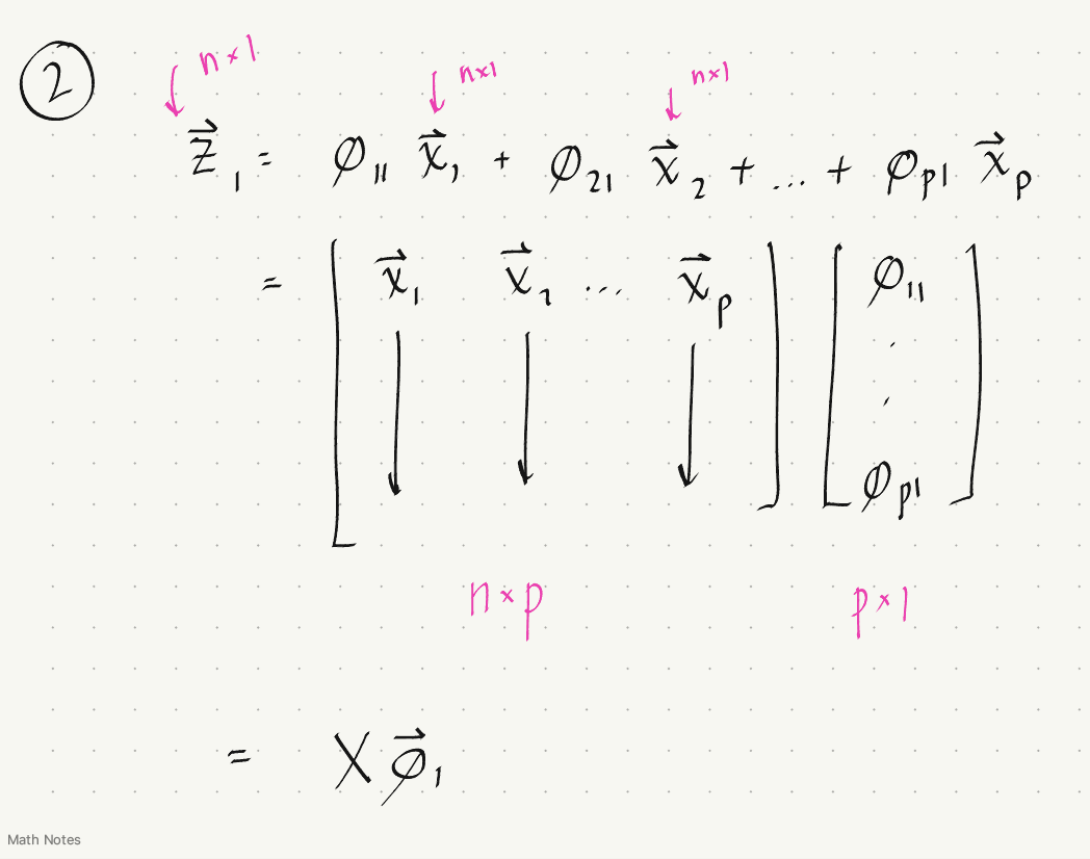


Figure 9.5: Computing principal component scores by projecting the data matrix onto the loading vector.

Each principal component $\mathbf{z}_k$ is an n-dimensional vector—one score per observation. The full set of principal components transforms the original $n \times p$ data into a new $n \times p$ representation where variables are uncorrelated.

9.8.5 Subsequent Principal Components

Higher-Order Components

The k-th principal component maximises variance subject to:

1. Unit norm: $\|\boldsymbol{\phi}_k\| = 1$
2. Orthogonality to previous loadings: $\boldsymbol{\phi}_k^T \boldsymbol{\phi}_j = 0$ for $j < k$

Equivalently, it is the eigenvector corresponding to the k-th largest eigenvalue of S:

$$\begin{aligned} S\boldsymbol{\phi}_1 = \lambda_1 \boldsymbol{\phi}_1 \quad &\rightarrow \quad \mathbf{z}_1 = X\boldsymbol{\phi}_1 \\ S\boldsymbol{\phi}_2 = \lambda_2 \boldsymbol{\phi}_2 \quad &\rightarrow \quad \mathbf{z}_2 = X\boldsymbol{\phi}_2 \\ &\vdots \end{aligned}$$

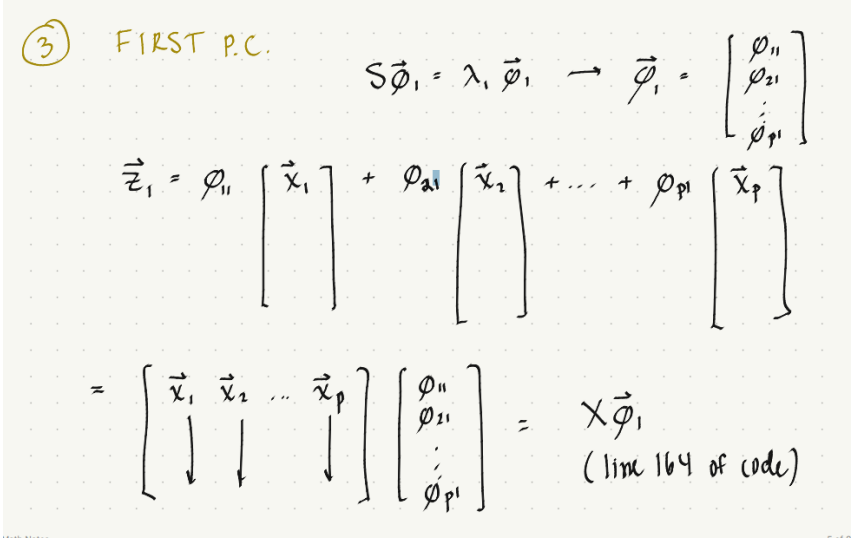


Figure 9.6: Multiple principal components ordered by variance explained. Each subsequent component captures progressively less variance while remaining orthogonal to all previous components.

9.8.6 Dimensionality Reduction

The power of PCA lies in **dimensionality reduction**: often, a small number of principal components capture most of the variance.

Choosing the Number of Components

Common strategies:

- **Variance threshold**: Keep enough components to explain (e.g.) 95% of total variance
- **Scree plot**: Plot eigenvalues and look for an "elbow"
- **Kaiser criterion**: Keep components with eigenvalue > 1 (for standardised data)

The proportion of variance explained by the first k components is:

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

9.8.7 Reconstructing Data from Principal Components

Given principal components, we can approximate the original data:

Data Reconstruction

Using all p components gives exact reconstruction:

$$X = Z\Phi^T$$

where $Z = [\mathbf{z}_1|\mathbf{z}_2|\cdots|\mathbf{z}_p]$ and $\Phi = [\boldsymbol{\phi}_1|\boldsymbol{\phi}_2|\cdots|\boldsymbol{\phi}_p]$.

Using only the first $k < p$ components gives a rank-k approximation:

$$\hat{X}_k = Z_k \Phi_k^T = \sum_{j=1}^{k} \mathbf{z}_j \boldsymbol{\phi}_j^T$$

This is the **best** rank-k approximation to X in the least-squares sense.

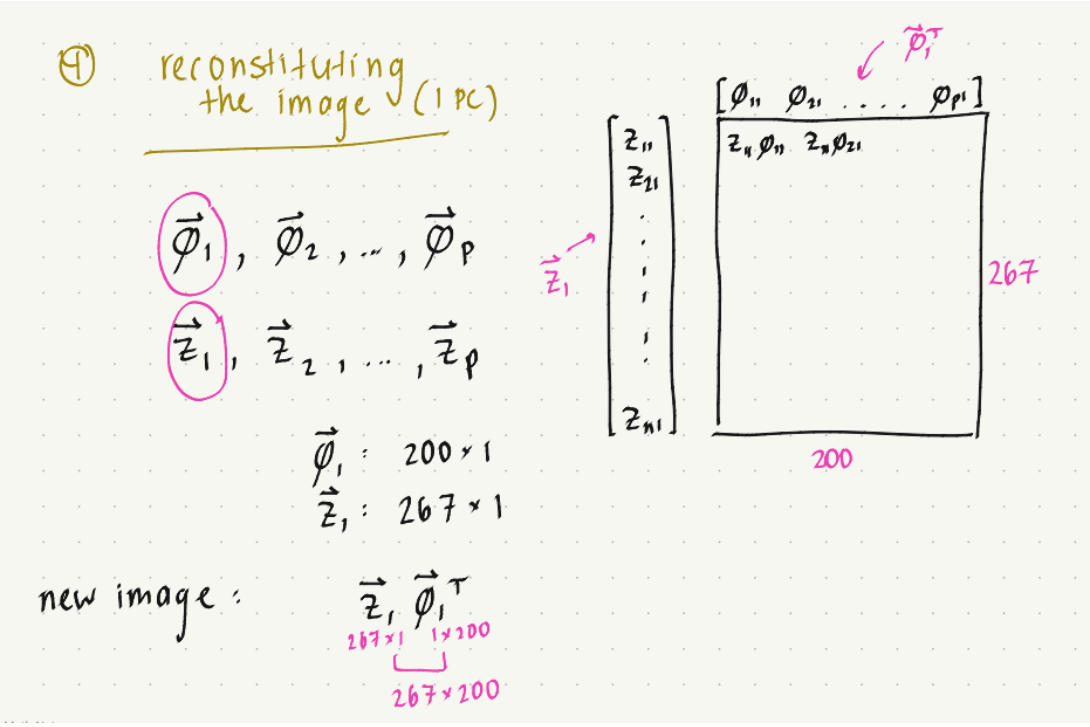


Figure 9.7: Reconstruction from principal components. The outer product $\mathbf{z}_k \boldsymbol{\phi}_k^T$ creates a rank-1 matrix; summing these gives progressively better approximations to the original data.

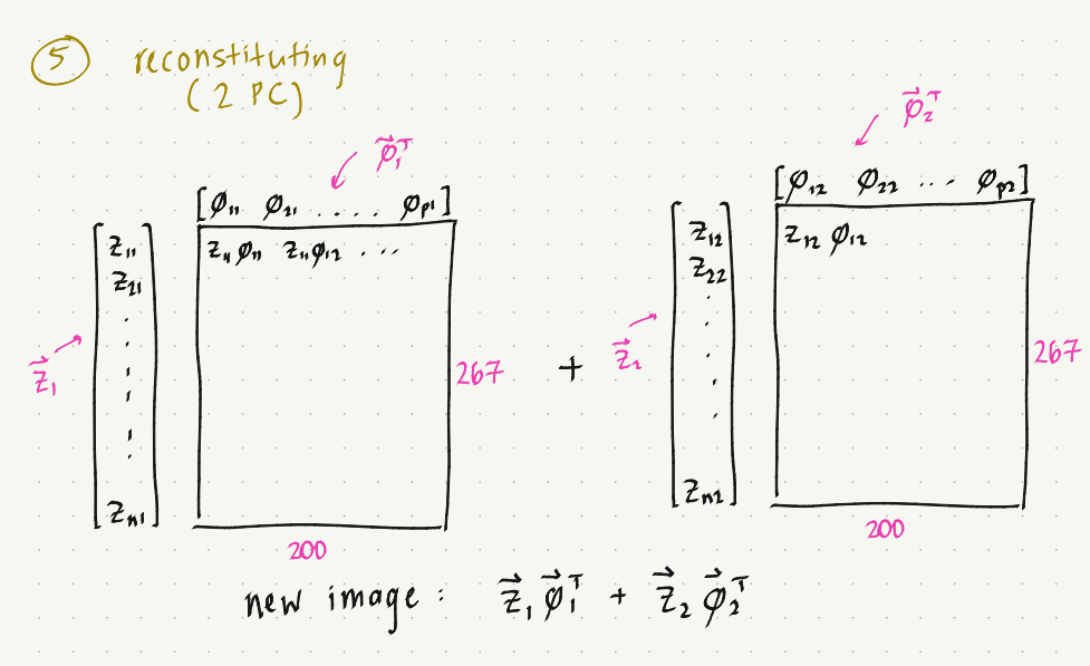


Figure 9.8: Each additional principal component adds detail to the reconstruction, capturing finer structure in the data.

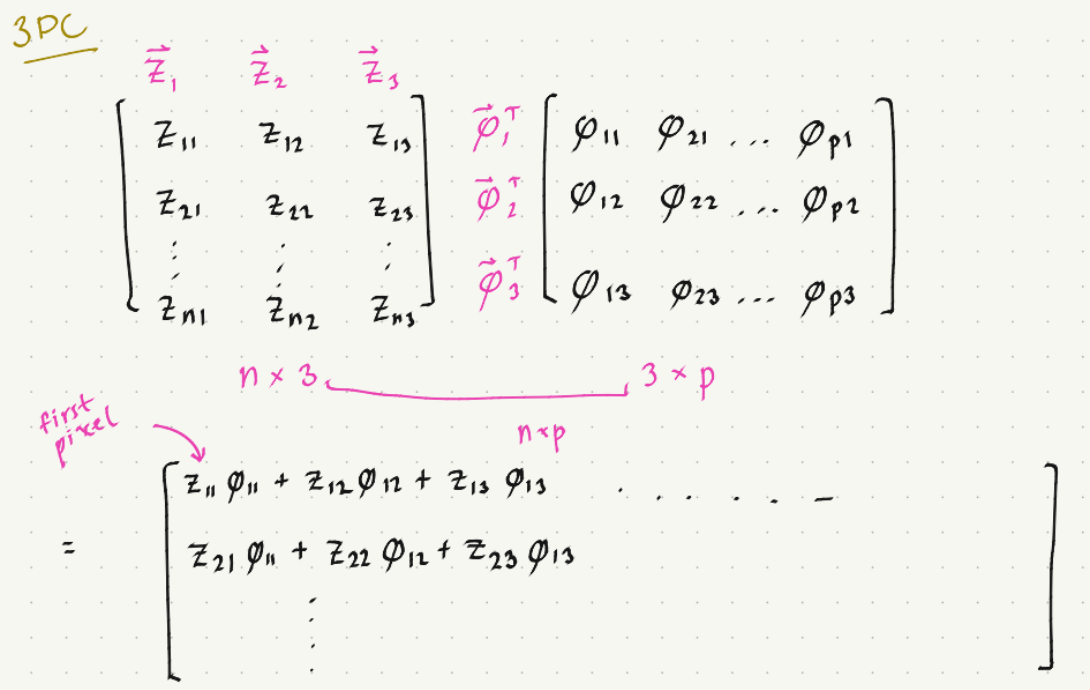


Figure 9.9: Image compression example: reconstructing an image using increasing numbers of principal components. Few components capture the main structure; more components recover fine details.

9.8.8 Applications of PCA

- **Data visualisation**: Project high-dimensional data to 2D or 3D for plotting
- **Noise reduction**: Low-rank approximations filter out noise in minor components

- **Feature extraction**: Principal components serve as new features for downstream models
- **Compression**: Store data using fewer numbers (principal component scores)
- **Multicollinearity**: Address correlated predictors by using uncorrelated components

Applications span genomics, facial recognition, computer vision, finance, climate science, and social sciences.

PCA Limitations

- PCA finds *linear* combinations; it cannot capture nonlinear structure
- PCA maximises variance, which may not align with predictive power or class separation
- Results depend on scaling: standardise variables first if they have different units
- Interpretability can be challenging: principal components are often mixtures of all original variables

9.9 Summary: Key Results

Chapter Summary

Linear Dependence:

- Vectors are linearly dependent if a non-trivial solution exists to $\sum c_i \mathbf{v}_i = \mathbf{0}$
- Dependent vectors contain redundant information; one can be expressed as a combination of others

Key Equivalences for Square Matrices:

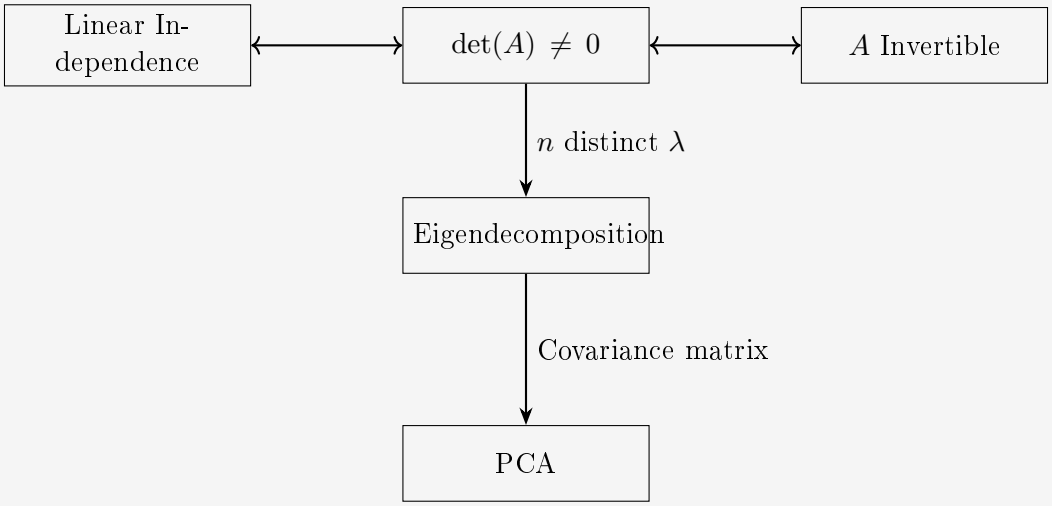
$$\text{Linearly Independent Columns} \iff \det(A) \neq 0 \iff A \text{ is Invertible}$$

Eigenvalues and Eigenvectors:

- Definition: $A\mathbf{v} = \lambda\mathbf{v}$
- Eigenvalues found from: $\det(A - \lambda I) = 0$
- Eigenvectors found from: $(A - \lambda I)\mathbf{v} = \mathbf{0}$

Eigendecomposition: $A = Q\Lambda Q^{-1}$ (requires n independent eigenvectors)
SVD: $A = U\Sigma V^T$ (works for any matrix)
PCA: The principal components are eigenvectors of the covariance matrix; eigenvalues give variance explained.

Connections Between Concepts



Part IV

Optimisation

Chapter 10

Optimisation

Learning Objectives

By the end of this chapter, you should be able to:

- Find critical points of functions using first derivatives
- Apply the second derivative test to classify critical points in one variable
- Construct and interpret the Hessian matrix for multivariable functions
- Use eigenvalue analysis of the Hessian to classify critical points
- Formulate constrained optimisation problems using Lagrange multipliers
- State and interpret the Karush-Kuhn-Tucker (KKT) conditions for inequality constraints
- Implement gradient descent and understand its convergence properties
- Connect optimisation techniques to Maximum Likelihood Estimation

Prerequisites

This chapter assumes familiarity with:

- Differentiation rules: product, quotient, and chain rules (from Chapter 4)
- Partial derivatives and the gradient vector
- Eigenvalues and eigenvectors of matrices (from **??**)
- Maximum Likelihood Estimation (from Chapter 4)

10.1 Why Optimisation Matters

Optimisation is the mathematical foundation of modern data science and machine learning. Almost every learning algorithm can be framed as finding parameters that minimise (or maximise) some objective function:

- **Maximum Likelihood Estimation**: Maximise the probability of observed data
- **Linear regression**: Minimise the sum of squared residuals
- **Neural networks**: Minimise a loss function via gradient descent
- **Support Vector Machines**: Maximise the margin subject to constraints
- **Portfolio optimisation**: Maximise returns subject to risk constraints

This chapter develops the mathematical machinery for finding optima, starting with single-variable calculus and building to constrained multivariable problems.

10.2 Unconstrained Optimisation: Single Variable

We begin with the simplest case: finding maxima and minima of a function $f : \mathbb{R} \to \mathbb{R}$.

10.2.1 Critical Points and the First Derivative

Definition 10.1 (Critical Point)**.** A point $x = c$ is a **critical point** of f if either:

1. $f'(c) = 0$ (the derivative is zero), or
2. $f'(c)$ does not exist (but $f(c)$ is defined)

Theorem 10.1 (Fermat's Theorem)**.** *If f has a local maximum or minimum at c, and $f'(c)$ exists, then $f'(c) = 0$.*

Why Critical Points?

At a local maximum or minimum, the tangent line must be horizontal. A positive slope means we could increase f by moving right; a negative slope means we could increase f by moving left. Only at a zero slope is there no "improving direction."

Converse is False

Not every critical point is an extremum. The function $f(x) = x^3$ has $f'(0) = 0$, but $x = 0$ is an **inflection point**, not a maximum or minimum. We need additional tests to classify critical points.

10.2.2 The Second Derivative Test

The second derivative tells us about the *curvature* of a function, which determines whether a critical point is a maximum, minimum, or neither.

Definition 10.2 (Concavity)**.** A function f is:

- **Concave up** on an interval if $f''(x) > 0$ on that interval (curves upward like a cup)
- **Concave down** on an interval if $f''(x) < 0$ on that interval (curves downward like a cap)

Theorem 10.2 (Second Derivative Test (Single Variable)). *Suppose f'' is continuous near c and $f'(c) = 0$. Then:*

1. *If $f''(c) > 0$, then f has a **local minimum** at c*
2. *If $f''(c) < 0$, then f has a **local maximum** at c*
3. *If $f''(c) = 0$, the test is **inconclusive***

Second Derivative Test Summary

At a critical point where $f'(c) = 0$:

$$\begin{aligned} f''(c) > 0 &\implies \text{local minimum (concave up: cup catches the point)} \\ f''(c) < 0 &\implies \text{local maximum (concave down: cap covers the point)} \\ f''(c) = 0 &\implies \text{inconclusive (need higher-order tests)} \end{aligned}$$

Physical Intuition

Think of $f''(c)$ as measuring "how the slope is changing" at c:

- $f''(c) > 0$: the slope is *increasing*. Since $f'(c) = 0$, the slope goes from negative to positive — we're at a minimum.
- $f''(c) < 0$: the slope is *decreasing*. Since $f'(c) = 0$, the slope goes from positive to negative — we're at a maximum.

Example 10.1 (Classifying Critical Points). Consider $f(x) = x^4 - 4x^3 + 4x^2$.

Step 1: Find critical points.

$$f'(x) = 4x^3 - 12x^2 + 8x = 4x(x^2 - 3x + 2) = 4x(x-1)(x-2)$$

Setting $f'(x) = 0$: critical points at $x = 0$, $x = 1$, and $x = 2$.

Step 2: Compute second derivative.

$$f''(x) = 12x^2 - 24x + 8$$

Step 3: Classify each critical point.

$$\begin{aligned} f''(0) &= 8 > 0 \quad \Rightarrow \text{local minimum} \\ f''(1) &= 12 - 24 + 8 = -4 < 0 \quad \Rightarrow \text{local maximum} \\ f''(2) &= 48 - 48 + 8 = 8 > 0 \quad \Rightarrow \text{local minimum} \end{aligned}$$

10.3 The Gradient Vector

For functions of multiple variables, the gradient generalises the concept of derivative.

Definition 10.3 (Gradient). For a scalar function $f : \mathbb{R}^n \to \mathbb{R}$, the **gradient** is the vector of partial derivatives:

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

Properties of the Gradient

The gradient $\nabla f(\mathbf{x})$ at a point $\mathbf{x}$:

1. **Points in the direction of steepest ascent** of f
2. **Has magnitude equal to the rate of steepest ascent**
3. **Is perpendicular to level curves/surfaces** of f

The negative gradient $-\nabla f$ points in the direction of steepest *descent*.

Example 10.2 (Gradient of a Quadratic Form)**.** Consider $f(\mathbf{z}) = \mathbf{z}^T\mathbf{z}$ where $\mathbf{z} \in \mathbb{R}^m$. This is the squared Euclidean norm.

Expanding: $f(\mathbf{z}) = z_1^2 + z_2^2 + \cdots + z_m^2$.

The gradient is:

$$\nabla_{\mathbf{z}} f(\mathbf{z}) = \begin{pmatrix} \frac{\partial f}{\partial z_1} \\ \frac{\partial f}{\partial z_2} \\ \vdots \\ \frac{\partial f}{\partial z_m} \end{pmatrix} = \begin{pmatrix} 2z_1 \\ 2z_2 \\ \vdots \\ 2z_m \end{pmatrix} = 2\mathbf{z}$$

This is analogous to the univariate result $\frac{d}{dx}(x^2) = 2x$: the gradient of $\|\mathbf{z}\|^2$ is $2\mathbf{z}$.

Gradient of Matrix-Valued Functions

When $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ takes a matrix A as input and outputs a scalar, the gradient is itself a matrix of the same shape:

$$\nabla_A f(A) = \begin{pmatrix} \frac{\partial f}{\partial A_{11}} & \frac{\partial f}{\partial A_{12}} & \cdots & \frac{\partial f}{\partial A_{1n}} \\ \frac{\partial f}{\partial A_{21}} & \frac{\partial f}{\partial A_{22}} & \cdots & \frac{\partial f}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial A_{m1}} & \frac{\partial f}{\partial A_{m2}} & \cdots & \frac{\partial f}{\partial A_{mn}} \end{pmatrix}$$

The (i, j)-entry is $(\nabla_A f)_{ij} = \frac{\partial f}{\partial A_{ij}}$: the rate of change of the output with respect to each input element.

10.4 The Hessian Matrix

The Hessian matrix generalises the second derivative to multiple dimensions. It captures the *curvature* of a function in all directions.

Definition 10.4 (Hessian Matrix)**.** For a twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, the **Hessian matrix** is the $n \times n$ matrix of second-order partial derivatives:

$$H(f) = \nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

The (i, j)-entry is $H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$.

Symmetry of the Hessian

If f has continuous second partial derivatives (which is typically the case in applications), then by Schwarz's theorem:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

Therefore the Hessian is a **symmetric matrix**: $H = H^T$.

10.4.1 The Second Derivative Test in Multiple Variables

For multivariable functions, we classify critical points using the eigenvalues of the Hessian.

Theorem 10.3 (Second Derivative Test (Multivariable)). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable, and suppose $\mathbf{c}$ is a critical point where $\nabla f(\mathbf{c}) = \mathbf{0}$. Let H be the Hessian evaluated at $\mathbf{c}$. Then:*

1. *If all eigenvalues of H are **positive**, then $\mathbf{c}$ is a **local minimum***
2. *If all eigenvalues of H are **negative**, then $\mathbf{c}$ is a **local maximum***
3. *If H has both positive and negative eigenvalues, then $\mathbf{c}$ is a **saddle point***
4. *If any eigenvalue is zero, the test is **inconclusive***

Hessian Classification Summary

At a critical point $\mathbf{c}$ where $\nabla f(\mathbf{c}) = \mathbf{0}$:

Eigenvalues of H	Classification
All $\lambda_i > 0$ (positive definite)	Local minimum
All $\lambda_i < 0$ (negative definite)	Local maximum
Mixed signs (indefinite)	Saddle point
Any $\lambda_i = 0$ (semidefinite)	Inconclusive

Why Eigenvalues?

The eigenvalues of the Hessian measure the curvature of f in the directions of the eigenvectors. A positive eigenvalue means the function curves upward in that direction; negative means it curves downward. For a minimum, we need upward curvature in *all* directions.

10.4.2 Special Case: Two Variables

For functions of two variables $f(x, y)$, there is a simpler criterion using the determinant.

Theorem 10.4 (Second Derivative Test for $f(x, y)$). *Let (a, b) be a critical point of $f(x, y)$. Define:*

$$D = \det(H) = \frac{\partial^2 f}{\partial x^2} \cdot \frac{\partial^2 f}{\partial y^2} - \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2$$

evaluated at (a, b). Let $f_{xx} = \frac{\partial^2 f}{\partial x^2}$ at (a, b). Then:

1. *If $D > 0$ and $f_{xx} > 0$:* ***local minimum***
2. *If $D > 0$ and $f_{xx} < 0$:* ***local maximum***
3. *If $D < 0$:* ***saddle point***
4. *If $D = 0$:* ***inconclusive***

Why D

For a 2×2 symmetric matrix, the eigenvalues satisfy:

$$\lambda_1 + \lambda_2 = \text{Tr}(H) = f_{xx} + f_{yy}$$
$$\lambda_1 \cdot \lambda_2 = \det(H) = f_{xx}f_{yy} - f_{xy}^2$$

When $D > 0$, the eigenvalues have the same sign. When $D < 0$, they have opposite signs (saddle). The sign of f_{xx} determines whether both are positive (minimum) or both negative (maximum).

Example 10.3 (Classifying Critical Points in Two Variables)**.** Consider $f(x, y) = x^3 - 3xy + y^3$.

Step 1: Find critical points.

$$\frac{\partial f}{\partial x} = 3x^2 - 3y = 0 \implies y = x^2$$
$$\frac{\partial f}{\partial y} = -3x + 3y^2 = 0 \implies x = y^2$$

Substituting $y = x^2$ into $x = y^2$: $x = (x^2)^2 = x^4$, so $x^4 - x = 0$, giving $x(x^3 - 1) = 0$.

Critical points: $(0, 0)$ and $(1, 1)$.

Step 2: Compute Hessian.

$$H = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix} = \begin{pmatrix} 6x & -3 \\ -3 & 6y \end{pmatrix}$$

Step 3: Classify each critical point.

At $(0, 0)$: $H = \begin{pmatrix} 0 & -3 \\ -3 & 0 \end{pmatrix}$, $D = (0)(0) - (-3)^2 = -9 < 0 \Rightarrow$ **saddle point**.

At $(1, 1)$: $H = \begin{pmatrix} 6 & -3 \\ -3 & 6 \end{pmatrix}$, $D = (6)(6) - (-3)^2 = 36 - 9 = 27 > 0$ and $f_{xx} = 6 > 0 \Rightarrow$ **local minimum**.

10.5 Constrained Optimisation: Lagrange Multipliers

Often we want to optimise a function subject to constraints. For example, maximising utility subject to a budget constraint, or finding the closest point on a surface to a given point.

10.5.1 The Geometric Intuition

Consider minimising $f(x, y) = x + y$ subject to $x^2 + y^2 = 1$ (the unit circle).

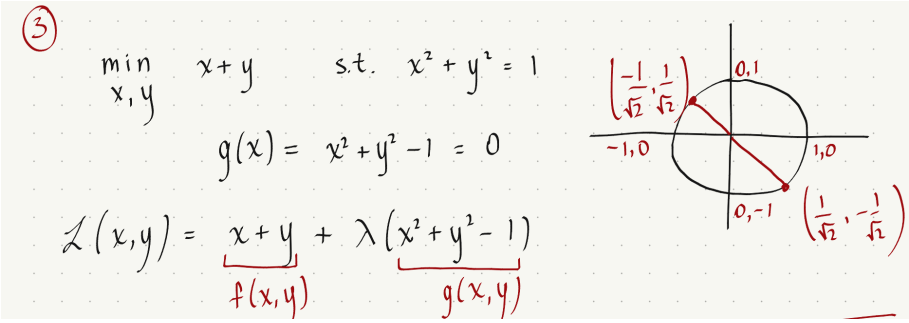


Figure 10.1: Minimising $f(x, y) = x + y$ subject to $x^2 + y^2 = 1$. The minimum occurs at $(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ where a level curve of f is tangent to the constraint circle.

Why Tangency?

Level curves of $f(x, y) = x + y$ are lines of the form $x + y = c$. As we decrease c, these lines move toward the lower-left. The minimum value of c for which the line still touches the circle is where the line is *tangent* to the circle. At tangency, the gradient of f and the gradient of the constraint point in the same (or opposite) direction.

10.5.2 The Method of Lagrange Multipliers

Definition 10.5 (The Lagrangian)**.** For the constrained optimisation problem

$$\text{optimise } f(\mathbf{x}) \quad \text{subject to } g(\mathbf{x}) = 0$$

the **Lagrangian** is:

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda \cdot g(\mathbf{x}) \tag{10.1}$$

where λ is called the **Lagrange multiplier**.

Theorem 10.5 (Lagrange Multiplier Theorem)**.** *Suppose $\mathbf{x}^*$ is a local extremum of f subject to $g(\mathbf{x}) = 0$, and $\nabla g(\mathbf{x}^*) \neq \mathbf{0}$. Then there exists $\lambda^* \in \mathbb{R}$ such that:*

$$\nabla f(\mathbf{x}^*) = \lambda^* \nabla g(\mathbf{x}^*)$$

Equivalently, the gradients $\nabla_{\mathbf{x}} \mathcal{L}$ and $\nabla_{\lambda} \mathcal{L}$ both vanish at $(\mathbf{x}^*, \lambda^*)$.

Procedure for Lagrange Multipliers

To solve optimise $f(x, y)$ subject to $g(x, y) = 0$:

1. Form the Lagrangian: $\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda \cdot g(x, y)$

2. Compute the first-order conditions:

$$\frac{\partial \mathcal{L}}{\partial x} = 0$$
$$\frac{\partial \mathcal{L}}{\partial y} = 0$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \quad \text{(recovers the constraint)}$$

3. Solve this system of equations for x, y, and λ
4. Classify the solutions (e.g., using bordered Hessian or direct evaluation)

The Sign Convention

Different textbooks use different conventions for the Lagrangian:

- $\mathcal{L} = f - \lambda g$ (used here)
- $\mathcal{L} = f + \lambda g$ (also common)

The sign of λ differs, but the method works identically. Be consistent within a problem.

10.5.3 Worked Examples

Example 10.4 (Maximising xy Subject to a Linear Constraint)**.** Maximise $f(x, y) = xy$ subject to $g(x, y) = x + y - 10 = 0$.

Step 1: Form the Lagrangian.

$$\mathcal{L}(x, y, \lambda) = xy - \lambda(x + y - 10)$$

Step 2: First-order conditions.

$$\frac{\partial \mathcal{L}}{\partial x} = y - \lambda = 0 \implies \lambda = y$$
$$\frac{\partial \mathcal{L}}{\partial y} = x - \lambda = 0 \implies \lambda = x$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = -(x + y - 10) = 0 \implies x + y = 10$$

Step 3: Solve. From the first two equations: $x = y$. Substituting into the constraint: $2x = 10$, so $x = 5$.

Solution: $(x^*, y^*) = (5, 5)$ with $f(5, 5) = 25$.

Step 4: Verify this is a maximum.

On the line $x+y = 10$, we can write $y = 10-x$, so $f = x(10-x) = 10x-x^2$. This is a downward parabola, confirming $(5, 5)$ is a maximum.

Example 10.5 (Minimising a Quadratic Subject to a Linear Constraint)**.** Minimise $f(x, y) = x^2 + 3y^2$ subject to $g(x, y) = x + 2y - 5 = 0$.

Step 1: Form the Lagrangian.

$$\mathcal{L}(x, y, \lambda) = x^2 + 3y^2 - \lambda(x + 2y - 5)$$

Step 2: First-order conditions.

$$\frac{\partial \mathcal{L}}{\partial x} = 2x - \lambda = 0 \implies x = \frac{\lambda}{2}$$
$$\frac{\partial \mathcal{L}}{\partial y} = 6y - 2\lambda = 0 \implies y = \frac{\lambda}{3}$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = -(x + 2y - 5) = 0 \implies x + 2y = 5$$

Step 3: Solve. Substituting $x = \frac{\lambda}{2}$ and $y = \frac{\lambda}{3}$ into the constraint:

$$\frac{\lambda}{2} + 2 \cdot \frac{\lambda}{3} = 5 \implies \frac{3\lambda + 4\lambda}{6} = 5 \implies \frac{7\lambda}{6} = 5 \implies \lambda = \frac{30}{7}$$

Therefore:

$$x^* = \frac{\lambda}{2} = \frac{15}{7}, \quad y^* = \frac{\lambda}{3} = \frac{10}{7}$$

Solution: $(x^*, y^*) = \left(\frac{15}{7}, \frac{10}{7}\right)$ with $f\left(\frac{15}{7}, \frac{10}{7}\right) = \frac{225}{49} + 3 \cdot \frac{100}{49} = \frac{225+300}{49} = \frac{525}{49} = \frac{75}{7}$.

Verification: Since $f(x, y) = x^2 + 3y^2$ is a convex function (sum of squares) and the constraint is linear, any critical point is a global minimum.

Example 10.6 (Finding Extrema on a Circle)**.** Find the maximum and minimum of $f(x, y) = x^2 + y^2$ subject to $g(x, y) = x^2 + y^2 - 4 = 0$.

Step 1: Form the Lagrangian.

$$\mathcal{L}(x, y, \lambda) = x^2 + y^2 - \lambda(x^2 + y^2 - 4)$$

Step 2: First-order conditions.

$$\frac{\partial \mathcal{L}}{\partial x} = 2x - 2\lambda x = 2x(1 - \lambda) = 0$$
$$\frac{\partial \mathcal{L}}{\partial y} = 2y - 2\lambda y = 2y(1 - \lambda) = 0$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = -(x^2 + y^2 - 4) = 0 \implies x^2 + y^2 = 4$$

Step 3: Solve. From the first two equations, either $(1 - \lambda) = 0$ or both $x = 0$ and $y = 0$.

If $x = y = 0$: this violates the constraint $x^2 + y^2 = 4$.

Therefore $\lambda = 1$, and any point on the circle $x^2 + y^2 = 4$ is a critical point.

Analysis: The objective function $f(x, y) = x^2 + y^2$ is constant on the constraint set (equal to 4 everywhere). This makes sense: we're measuring the distance from the origin, and every point on the circle is the same distance away!

Observation: This example shows that Lagrange multipliers give us *all* critical points, but when the objective is constant on the constraint, every point is both a maximum and a minimum.

10.6 Optimisation as Eigenvalue Problems

Some constrained optimisation problems reduce to finding eigenvalues and eigenvectors.

Example 10.7 (Maximising a Quadratic Form on the Unit Sphere)**.** Consider the problem:

$$\max_{\mathbf{x}\in\mathbb{R}^n} \mathbf{x}^T A\mathbf{x} \quad \text{subject to} \quad \|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = 1$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric.

Step 1: Form the Lagrangian.

$$\mathcal{L}(\mathbf{x}, \lambda) = \mathbf{x}^T A\mathbf{x} - \lambda(\mathbf{x}^T\mathbf{x} - 1)$$

Step 2: First-order condition. Taking the gradient with respect to $\mathbf{x}$:

$$\nabla_{\mathbf{x}}\mathcal{L} = 2A\mathbf{x} - 2\lambda\mathbf{x} = \mathbf{0}$$

This gives:

$$A\mathbf{x} = \lambda\mathbf{x}$$

This is the eigenvalue equation! The critical points are the eigenvectors of A, and the Lagrange multipliers are the eigenvalues.

Step 3: Find the optimum. If $\mathbf{x}$ is an eigenvector with eigenvalue λ, then:

$$\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T(\lambda\mathbf{x}) = \lambda\mathbf{x}^T\mathbf{x} = \lambda$$

(using $\|\mathbf{x}\| = 1$).

Conclusion: The maximum is $\lambda_{\max}$ (the largest eigenvalue), achieved at the corresponding eigenvector. The minimum is $\lambda_{\min}$.

Rayleigh Quotient

For a symmetric matrix A:

$$\lambda_{\min} \leq \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \leq \lambda_{\max}$$

for all non-zero $\mathbf{x}$. The bounds are achieved by the corresponding eigenvectors.

Why Eigenvalues?

Eigenvectors are the directions along which a matrix acts by pure scaling. The eigenvalues tell us how much. When optimising a quadratic form, we're asking: "In which direction does A stretch the most?" The answer is the eigenvector with the largest eigenvalue.

10.7 Inequality Constraints: KKT Conditions

In practice, constraints are often *inequalities* rather than equalities. For example, "budget cannot exceed \$1000" rather than "budget equals \$1000."

Definition 10.6 (Standard Constrained Optimisation Problem)**.** The standard form with inequality constraints is:

$$\begin{aligned} \text{minimise} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \\ & h_j(\mathbf{x}) = 0, \quad j = 1, \ldots, p \end{aligned}$$

Theorem 10.6 (Karush-Kuhn-Tucker (KKT) Conditions)**.** *For a local minimum $\mathbf{x}^*$ of the above problem (under suitable regularity conditions), there exist multipliers $\mu_i^* \geq 0$ and λ_j^* such that:*

1. ***Stationarity**: $\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \lambda_j^* \nabla h_j(\mathbf{x}^*) = \mathbf{0}$*
2. ***Primal feasibility**: $g_i(\mathbf{x}^*) \leq 0$ and $h_j(\mathbf{x}^*) = 0$*
3. ***Dual feasibility**: $\mu_i^* \geq 0$*
4. ***Complementary slackness**: $\mu_i^* g_i(\mathbf{x}^*) = 0$ for all i*

Complementary Slackness

The condition $\mu_i^* g_i(\mathbf{x}^*) = 0$ means: for each inequality constraint, either:

- The constraint is **active**: $g_i(\mathbf{x}^*) = 0$ (the constraint binds), and μ_i^* can be positive
- The constraint is **inactive**: $g_i(\mathbf{x}^*) < 0$ (the constraint has slack), and $\mu_i^* = 0$

Intuitively: we only "pay attention" to constraints that are actually binding at the optimum.

KKT for Convex Problems

When f is convex, all g_i are convex, and all h_j are affine (linear), the KKT conditions are not just necessary but also **sufficient** for global optimality. This makes convex optimisation particularly tractable.

10.8 Gradient Descent

When we cannot solve $\nabla f = \mathbf{0}$ analytically, we turn to iterative numerical methods. Gradient descent is the workhorse of modern machine learning.

10.8.1 The Algorithm

Definition 10.7 (Gradient Descent)**.** To minimise $f : \mathbb{R}^n \to \mathbb{R}$, gradient descent iterates:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)}) \tag{10.2}$$

where $\alpha > 0$ is the **learning rate** (or step size).

Gradient Descent Algorithm

Input: Starting point $\mathbf{x}^{(0)}$, learning rate α, tolerance ϵ
Repeat:

1. Compute gradient: $\mathbf{g} = \nabla f(\mathbf{x}^{(k)})$
2. Update: $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \mathbf{g}$
3. Check convergence: if $\|\mathbf{g}\| < \epsilon$, stop

Output: Approximate minimiser $\mathbf{x}^{(k)}$

Why Negative Gradient?

The gradient ∇f points "uphill" — in the direction of steepest increase. To minimise, we go "downhill" by following $-\nabla f$. Each step reduces the function value (for small enough α), gradually descending toward a minimum.

10.8.2 Choosing the Learning Rate

The learning rate α is critical:

Learning Rate Pitfalls

- **Too large**: Steps overshoot the minimum, causing oscillation or divergence
- **Too small**: Convergence is extremely slow; may get stuck in flat regions

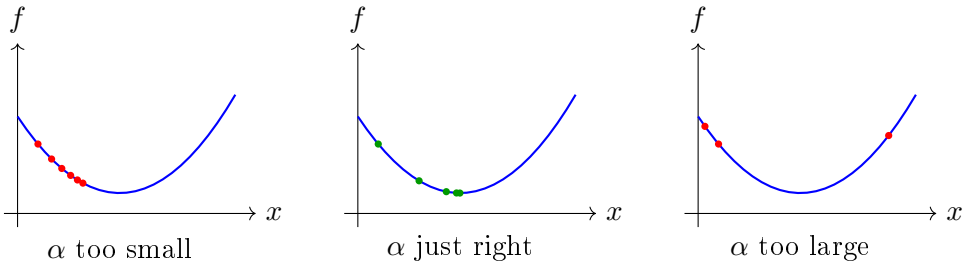


Figure 10.2: Effect of learning rate on gradient descent convergence.

Convergence Guarantee

For a function f with L-Lipschitz gradient (i.e., $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$), gradient descent with step size $\alpha \leq 1/L$ converges. If f is also convex, convergence is to a global minimum.

10.8.3 Variants of Gradient Descent

Modern machine learning uses several variants:

- **Stochastic Gradient Descent (SGD)**: Uses a random subset of data to estimate the gradient; much faster for large datasets

- **Momentum**: Accumulates past gradients to "smooth out" updates:

$$\mathbf{v}^{(k+1)} = \beta \mathbf{v}^{(k)} + \nabla f(\mathbf{x}^{(k)}), \quad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \mathbf{v}^{(k+1)}$$

- **Adam**: Adaptive learning rates per parameter, combining momentum with RMSprop
- **Newton's method**: Uses second-order information (Hessian) for faster convergence near minima

10.9 Application: Portfolio Optimisation

We now work through a complete example of constrained optimisation in finance.

10.9.1 Problem Setup

Consider a portfolio with two assets, A and B, with:

- Expected returns: r_A and r_B
- Portfolio weights: w_A and w_B (fractions invested in each asset)
- Variance-covariance matrix of returns:

$$\Sigma = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}$$

Objective: Maximise expected return while penalising variance (risk aversion).

Constraint: Weights must sum to 1 (fully invested): $w_A + w_B = 1$.

10.9.2 Formulating the Problem

The expected return is:

$$\mathbb{E}[\text{portfolio return}] = w_A r_A + w_B r_B$$

The portfolio variance is:

$$\begin{aligned} \operatorname{Var}(w_A r_A + w_B r_B) &= w_A^2 \operatorname{Var}(r_A) + w_B^2 \operatorname{Var}(r_B) + 2 w_A w_B \operatorname{Cov}(r_A, r_B) \\ &= w_A^2 \sigma_A^2 + w_B^2 \sigma_B^2 + 2 w_A w_B \sigma_{AB} \end{aligned}$$

With risk aversion parameter $\gamma \geq 0$, we maximise:

$$f(w_A, w_B) = w_A r_A + w_B r_B - \gamma \left(w_A^2 \sigma_A^2 + w_B^2 \sigma_B^2 + 2 w_A w_B \sigma_{AB} \right)$$

10.9.3 Worked Example with Specific Values

Example 10.8 (Portfolio Optimisation)**.** Suppose the variance-covariance matrix is:

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$$

So $\sigma_A^2 = 1$, $\sigma_B^2 = 2$, and $\sigma_{AB} = 0.5$.

Step 1: Form the Lagrangian.

$$\mathcal{L}(w_A, w_B, \lambda) = w_A r_A + w_B r_B - \gamma(w_A^2 + 2w_B^2 + w_A w_B) - \lambda(w_A + w_B - 1)$$

Step 2: First-order conditions.

$$\frac{\partial \mathcal{L}}{\partial w_A} = r_A - 2\gamma w_A - \gamma w_B - \lambda = 0$$
$$\frac{\partial \mathcal{L}}{\partial w_B} = r_B - 4\gamma w_B - \gamma w_A - \lambda = 0$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = -(w_A + w_B - 1) = 0$$

Step 3: Solve the system. From the first two equations:

$$r_A - 2\gamma w_A - \gamma w_B = \lambda$$
$$r_B - 4\gamma w_B - \gamma w_A = \lambda$$

Subtracting:

$$(r_A - r_B) - 2\gamma w_A + 4\gamma w_B - \gamma w_B + \gamma w_A = 0$$
$$(r_A - r_B) - \gamma w_A + 3\gamma w_B = 0$$

Combined with $w_A + w_B = 1$ (so $w_A = 1 - w_B$):

$$(r_A - r_B) - \gamma(1 - w_B) + 3\gamma w_B = 0$$
$$(r_A - r_B) - \gamma + \gamma w_B + 3\gamma w_B = 0$$
$$(r_A - r_B) - \gamma + 4\gamma w_B = 0$$
$$w_B^* = \frac{\gamma - (r_A - r_B)}{4\gamma} = \frac{1}{4} - \frac{r_A - r_B}{4\gamma}$$

And:

$$w_A^* = 1 - w_B^* = \frac{3}{4} + \frac{r_A - r_B}{4\gamma}$$

Interpretation:

- If $r_A = r_B$ (equal expected returns): $w_A^* = 3/4$, $w_B^* = 1/4$. We favour Asset A because it has lower variance ($\sigma_A^2 = 1 < 2 = \sigma_B^2$).
- If $r_A > r_B$: We allocate more to Asset A.
- As $\gamma \to \infty$ (extreme risk aversion): Weights converge to the minimum-variance portfolio.
- As $\gamma \to 0$ (risk-neutral): The penalty term vanishes; we invest entirely in the higher-return asset.

10.10 Connection to Maximum Likelihood Estimation

Maximum Likelihood Estimation (Chapter 4) is fundamentally an optimisation problem: we seek parameters θ that maximise the likelihood (or equivalently, the log-likelihood).

MLE as Optimisation

For i.i.d. data $x_1, \ldots, x_n$ from a distribution with parameter θ:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \ell(\theta) = \arg\max_{\theta} \sum_{i=1}^{n} \ln f(x_i; \theta)$$

The techniques from this chapter apply directly:

1. **First derivative (score function)**: Set $\frac{d\ell}{d\theta} = 0$ to find critical points
2. **Second derivative test**: Verify $\frac{d^2\ell}{d\theta^2} < 0$ to confirm a maximum
3. **Hessian for multiple parameters**: When estimating $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, the observed Fisher information is the negative Hessian:
$$\mathcal{I}(\boldsymbol{\theta}) = -\nabla^2 \ell(\boldsymbol{\theta})$$
4. **Gradient descent**: When the MLE has no closed form (e.g., logistic regression, neural networks), we use gradient ascent on $\ell(\theta)$.

Example 10.9 (Gradient Ascent for MLE)**.** For logistic regression with parameters $\boldsymbol{\beta}$, the log-likelihood has no closed-form maximiser. We use gradient ascent:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \alpha \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^{(k)})$$

Note the *plus* sign: we're ascending to maximise, not descending.

10.11 Summary

Key Concepts in Optimisation

Unconstrained Optimisation:

- Find critical points: $\nabla f = \mathbf{0}$
- Classify using Hessian eigenvalues: all positive $\Rightarrow$ min, all negative $\Rightarrow$ max, mixed $\Rightarrow$ saddle

Constrained Optimisation:

- Form Lagrangian: $\mathcal{L} = f - \lambda g$
- Solve: $\nabla \mathcal{L} = \mathbf{0}$
- KKT conditions extend to inequality constraints

Numerical Methods:

- Gradient descent: $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \nabla f$
- Learning rate selection is critical

10.12 Practice Exercises

Unconstrained Optimisation

1. Find and classify all critical points of $f(x) = x^4 - 8x^2 + 3$.
2. For $f(x, y) = x^2 + y^2 - 2x - 4y + 5$, find the critical point and determine whether it is a maximum, minimum, or saddle point.
3. Compute the Hessian of $f(x, y, z) = x^2 + y^2 + z^2 - xy$ and determine its definiteness.

Constrained Optimisation

4. Find the point on the line $2x + 3y = 6$ closest to the origin. (Hint: Minimise $x^2 + y^2$ subject to the constraint.)
5. A rectangular box with no top is to have a volume of 32 cubic metres. Find the dimensions that minimise the surface area.
6. Use Lagrange multipliers to find the maximum of $f(x, y) = x^2 y$ subject to $x^2 + y^2 = 3$.

Gradient Descent

7. Implement gradient descent to minimise $f(x) = (x - 3)^2 + 1$ starting from $x_0 = 0$ with learning rate $\alpha = 0.1$. How many iterations are needed to reach within 0.01 of the minimum?
8. For $f(x, y) = x^2 + 4y^2$, compute the gradient and write out the gradient descent update rule.

Part V

Appendices

Appendix A

Common Distributions

This appendix provides a comprehensive reference for probability distributions commonly encountered in statistics and data science. For each distribution, we present the definition, key functions (PMF/PDF, CDF, MGF), moments, and relationships to other distributions.

Part I: Discrete Distributions

Discrete distributions assign probabilities to countable outcomes. The probability mass function (PMF) gives $P(X = k)$ for each possible value k in the support.

A.1 Bernoulli Distribution

- **Definition and Parameters:**
  - Definition: Represents an experiment with exactly two possible outcomes, often referred to as "success" and "failure". It is the simplest discrete distribution.
  - Parameters:
    - $p \in [0, 1]$: Probability of success.
  - Notation: $X \sim \text{Bernoulli}(p)$ or $X \sim \text{Bern}(p)$
  - Support: $\{0, 1\}$
- **Probability Mass Function (PMF):**

$$P(X = k) = p^k(1-p)^{1-k} = \begin{cases} p & \text{if } k = 1 \\ 1-p & \text{if } k = 0 \end{cases}$$

- **Cumulative Distribution Function (CDF):**

$$F(k) = P(X \leq k) = \begin{cases} 0 & \text{for } k < 0 \\ 1-p & \text{for } 0 \leq k < 1 \\ 1 & \text{for } k \geq 1 \end{cases}$$

- **Moment Generating Function (MGF):**

$$M_X(t) = E[e^{tX}] = (1-p) + pe^t = 1 - p + pe^t$$

This follows directly from the definition: $M_X(t) = e^{t \cdot 0}(1-p) + e^{t \cdot 1}p$.

- **Expected Value and Variance:**

$$E(X) = p$$
$$\text{Var}(X) = p(1-p)$$

  Note that variance is maximised when $p = 0.5$, giving $\text{Var}(X) = 0.25$.

- **Worked Example:**
  - Problem Statement: A coin is biased with probability $p = 0.7$ of landing heads. What is the probability it lands tails?
  - Solution: Using the Bernoulli PMF, $P(X = 0) = 1 - 0.7 = 0.3$.

- **Relation to Other Distributions:**
  - A Bernoulli distribution is a special case of the Binomial distribution with $n = 1$.
  - The sum of n i.i.d. Bernoulli(p) random variables follows a Binomial(n, p) distribution.

- **Use Cases:**
  - Modelling the outcome of a single trial in any scenario with two possible outcomes: coin tosses, yes/no questions, pass/fail tests, click/no-click events.

- **Miscellaneous Notes:**
  - Named after Jacob Bernoulli, a Swiss mathematician.
  - The Bernoulli distribution is the building block for many other distributions.

A.2 Binomial Distribution

- **Definition and Parameters:**
  - Definition: Represents the number of successes in n independent Bernoulli trials, each with the same probability p of success.
  - Parameters:
    - $n \in \{1, 2, 3, \ldots\}$: Number of trials.
    - $p \in [0, 1]$: Probability of success on a single trial.
  - Notation: $X \sim \text{Binomial}(n, p)$ or $X \sim \text{Bin}(n, p)$
  - Support: $\{0, 1, 2, \ldots, n\}$

- **Probability Mass Function (PMF):**

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

  where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient.

- **Cumulative Distribution Function (CDF):**

$$F(k) = P(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$$

- **Moment Generating Function (MGF):**

$$M_X(t) = E[e^{tX}] = \left[(1-p) + pe^t\right]^n$$

This follows from the fact that the sum of n independent Bernoulli MGFs gives $(1-p+pe^t)^n$.

- **Expected Value and Variance:**

$$E(X) = np$$
$$\text{Var}(X) = np(1-p)$$

- **Worked Example:**
  - Problem Statement: A fair coin is tossed 5 times. What is the probability of getting exactly 3 heads?
  - Solution: Using the Binomial PMF with $n = 5$, $p = 0.5$, $k = 3$:

$$P(X = 3) = \binom{5}{3} \times 0.5^3 \times 0.5^2 = 10 \times 0.125 \times 0.25 = 0.3125$$

- **Relation to Other Distributions:**
  - Binomial$(1, p) \equiv$ Bernoulli(p).
  - As $n \to \infty$ with $np = \lambda$ held constant (i.e., $p \to 0$), the Binomial converges to Poisson(λ).
  - For large n, Binomial(n, p) is approximately Normal$(np, np(1-p))$ (by CLT).
- **Use Cases:**
  - Number of heads in coin tosses, number of defective items in a batch, number of correct answers on a multiple-choice test (if guessing randomly).
- **Miscellaneous Notes:**
  - One of the most widely used probability distributions in statistics.
  - Assumes trials are independent with constant success probability.

A.3 Multinomial Distribution

- **Definition and Parameters:**
  - Definition: An extension of the binomial distribution to experiments with more than two possible outcomes. It represents the outcomes of n independent trials, each of which can result in one of k possible categories.
  - Parameters:
    - n: Total number of trials.
    - k: Number of possible outcomes/categories.
    - $p_1, p_2, \ldots, p_k$: Probabilities of each outcome, where $\sum_{i=1}^{k} p_i = 1$.
  - Notation: $(X_1, \ldots, X_k) \sim$ Multinomial$(n, p_1, \ldots, p_k)$
  - Support: $\{(n_1, \ldots, n_k) : n_i \geq 0, \sum_i n_i = n\}$

- **Probability Mass Function (PMF):**

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

subject to $n_1 + n_2 + \cdots + n_k = n$.

- **Moment Generating Function (MGF):**

$$M_{\mathbf{X}}(t_1, \ldots, t_k) = \left(\sum_{i=1}^{k} p_i e^{t_i} \right)^n$$

- **Expected Value and Variance:**

$$\begin{aligned} E(X_i) &= np_i \\ \text{Var}(X_i) &= np_i(1 - p_i) \\ \text{Cov}(X_i, X_j) &= -np_i p_j \quad \text{for } i \neq j \end{aligned}$$

Note the negative covariance: if more outcomes fall in category i, fewer are available for category j.

- **Worked Example:**
  - Problem Statement: A die is rolled 10 times. What is the probability of getting exactly 2 ones, 3 twos, and 5 threes (and zero of the other faces)?
  - Solution: Using the Multinomial PMF:

$$P = \frac{10!}{2! \cdot 3! \cdot 5! \cdot 0! \cdot 0! \cdot 0!} \times \left(\frac{1}{6} \right)^{10} = \frac{2520}{6^{10}} \approx 4.17 \times 10^{-5}$$

- **Relation to Other Distributions:**
  - Multinomial with $k = 2$ reduces to the Binomial distribution.
  - Each marginal $X_i \sim \text{Binomial}(n, p_i)$.
- **Use Cases:**
  - Modelling outcomes with multiple categories: dice rolls, categorical survey responses, word counts in text analysis.

A.4 Geometric Distribution

- **Definition and Parameters:**
  - Definition: Describes the number of Bernoulli trials needed for the first success. It represents the probability that the first success occurs on the kth trial.
  - Parameters:
    - $p \in (0, 1]$: Probability of success on a single trial.
  - Notation: $X \sim \text{Geometric}(p)$ or $X \sim \text{Geom}(p)$
  - Support: $\{1, 2, 3, \ldots\}$ (trials until first success)
    Alternative parametrisation: Some texts define support as $\{0, 1, 2, \ldots\}$ (failures before first success). The formulas below use the "trials until success" convention.

- **Probability Mass Function (PMF):**

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

- **Cumulative Distribution Function (CDF):**

$$F(k) = P(X \leq k) = 1 - (1 - p)^k$$

- **Moment Generating Function (MGF):**

$$M_X(t) = \frac{pe^t}{1 - (1 - p)e^t} \quad \text{for } t < -\ln(1 - p)$$

- **Expected Value and Variance:**

$$E(X) = \frac{1}{p}$$
$$\text{Var}(X) = \frac{1 - p}{p^2}$$

- **Worked Example:**
  - Problem Statement: A fair coin is tossed repeatedly. What is the probability that the first head appears on the 3rd toss?
  - Solution: With $p = 0.5$:

$$P(X = 3) = (1 - 0.5)^{3-1} \times 0.5 = 0.25 \times 0.5 = 0.125$$

- **Memorylessness Property:** The Geometric distribution is the only discrete distribution with the memoryless property:

$$P(X > s + t \mid X > s) = P(X > t)$$

"Given that no success has occurred in the first s trials, the probability of waiting at least t more trials is the same as starting fresh."

- **Relation to Other Distributions:**
  - Geometric(p) is Negative Binomial($1, p$)—the special case of waiting for exactly one success.
  - The continuous analogue is the Exponential distribution.
- **Use Cases:**
  - Number of trials until first success: coin tosses until first head, sales calls until first sale, login attempts until success.

A.5 Negative Binomial Distribution

- **Definition and Parameters:**
  - Definition: Describes the number of Bernoulli trials needed to achieve r successes. It represents the probability that the rth success occurs on the kth trial.
  - Parameters:

    - $r \in \{1, 2, 3, \ldots\}$: Number of successes required.
    - $p \in (0, 1]$: Probability of success on a single trial.
  - Notation: $X \sim \text{NegBin}(r, p)$
  - Support: $\{r, r+1, r+2, \ldots\}$

- **Probability Mass Function (PMF):**

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r} \quad \text{for } k = r, r+1, r+2, \ldots$$

  The binomial coefficient $\binom{k-1}{r-1}$ counts the ways to arrange $r-1$ successes among the first $k-1$ trials.

- **Moment Generating Function (MGF):**

$$M_X(t) = \left(\frac{pe^t}{1-(1-p)e^t}\right)^r \quad \text{for } t < -\ln(1-p)$$

- **Expected Value and Variance:**

$$E(X) = \frac{r}{p}$$
$$\text{Var}(X) = \frac{r(1-p)}{p^2}$$

- **Worked Example:**
  - Problem Statement: A biased coin with $p = 0.4$ of landing heads is tossed repeatedly. What is the probability that the 5th head appears on the 8th toss?
  - Solution: We need exactly 4 heads in the first 7 tosses, then a head on toss 8:

$$P(X = 8) = \binom{7}{4} \times 0.4^5 \times 0.6^3 = 35 \times 0.01024 \times 0.216 \approx 0.0774$$

- **Relation to Other Distributions:**
  - $\text{NegBin}(1, p) \equiv \text{Geometric}(p)$.
  - Sum of r i.i.d. Geometric(p) random variables is $\text{NegBin}(r, p)$.
  - The continuous analogue is the Gamma distribution.

- **Use Cases:**
  - Number of trials to achieve a fixed number of successes, modelling overdispersed count data (where variance exceeds mean, unlike Poisson).

A.6 Hypergeometric Distribution

- **Definition and Parameters:**
  - Definition: Describes the probability of obtaining exactly k successes when drawing n items *without replacement* from a finite population of size N containing K success items.
  - Parameters:

    - N: Total population size.
    - K: Number of success items in the population.
    - n: Number of items drawn (sample size).
  - Notation: $X \sim \text{Hypergeometric}(N, K, n)$
  - Support: $\{\max(0, n - N + K), \ldots, \min(n, K)\}$

- **Probability Mass Function (PMF):**

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

  Interpretation: Choose k successes from K available, and $n - k$ failures from $N - K$ available, divided by total ways to choose n from N.

- **Moment Generating Function (MGF):** The MGF does not have a simple closed form. It can be expressed using hypergeometric functions:

$$M_X(t) = \frac{\binom{N-K}{n}}{\binom{N}{n}} \cdot {}_2F_1\left(-n, -K; N - K - n + 1; e^t\right)$$

  where ${}_2F_1$ is the hypergeometric function.

- **Expected Value and Variance:**

$$E(X) = n\frac{K}{N}$$
$$\text{Var}(X) = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$$

  The factor $\frac{N-n}{N-1}$ is the *finite population correction factor*. It reduces variance because sampling without replacement introduces negative correlation between draws.

- **Worked Example:**
  - Problem Statement: From a standard deck of 52 cards, 5 are drawn without replacement. What is the probability that exactly 3 are spades?
  - Solution: Here $N = 52$, $K = 13$ (spades), $n = 5$, $k = 3$:

$$P(X = 3) = \frac{\binom{13}{3}\binom{39}{2}}{\binom{52}{5}} = \frac{286 \times 741}{2598960} \approx 0.0815$$

- **Relation to Other Distributions:**
  - When $N \to \infty$ with $K/N = p$ fixed (i.e., sampling with replacement), $\text{Hypergeometric}(N, K, n)$ converges to $\text{Binomial}(n, p)$.
  - Fisher's exact test uses the hypergeometric distribution.

- **Use Cases:**
  - Sampling without replacement: drawing cards, quality control (selecting items from a batch), capture-recapture methods in ecology.

- **Miscellaneous Notes:**
  - The Hypergeometric provides exact probabilities, while Binomial approximates when sampling without replacement from large populations.
  - Common in biology/genetics for analysing enrichment of gene sets.

A.7 Poisson Distribution

- **Definition and Parameters:**
  - Definition: Describes the number of events occurring in a fixed interval of time or space, given that events occur independently at a constant average rate.
  - Parameters:
    - $\lambda > 0$: Average rate (expected number of events in the interval).
  - Notation: $X \sim \text{Poisson}(\lambda)$ or $X \sim \text{Pois}(\lambda)$
  - Support: $\{0, 1, 2, \ldots\}$
- **Probability Mass Function (PMF):**

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- **Cumulative Distribution Function (CDF):**

$$F(k) = P(X \leq k) = e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!} = \frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$$

  where $\Gamma(s, x)$ is the upper incomplete gamma function.
- **Moment Generating Function (MGF):**

$$M_X(t) = E[e^{tX}] = e^{\lambda(e^t - 1)}$$

  Derivation sketch: $M_X(t) = \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}$.
- **Expected Value and Variance:**

$$E(X) = \lambda$$
$$\text{Var}(X) = \lambda$$

  The equality of mean and variance is a defining characteristic of the Poisson distribution (*equidispersion*).
- **Worked Example:**
  - Problem Statement: A call centre receives an average of 5 calls per hour. What is the probability of receiving exactly 7 calls in a given hour?
  - Solution: With $\lambda = 5$:

$$P(X = 7) = \frac{5^7 e^{-5}}{7!} = \frac{78125 \times 0.00674}{5040} \approx 0.1044$$

- **Key Properties:**
  - **Additivity**: If $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$ are independent, then $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$.
  - **Conditioning**: Given $X + Y = n$, the conditional distribution of X is Binomial.
- **Relation to Other Distributions:**

  - Poisson is the limit of Binomial(n, p) as $n \to \infty$, $p \to 0$, with $np = \lambda$.
  - For large λ, Poisson$(\lambda) \approx$ Normal(λ, λ).
  - The inter-arrival times between Poisson events follow an Exponential distribution.

- **Use Cases:**
  - Count data: phone calls per hour, website hits per day, number of typos per page, radioactive decay events.

- **Miscellaneous Notes:**
  - Named after Siméon Denis Poisson.
  - Particularly known for modelling rare events.
  - If variance exceeds mean, consider Negative Binomial for overdispersed count data.

Part II: Continuous Distributions

Continuous distributions assign probabilities to intervals via probability density functions (PDFs). For any specific value x, $P(X = x) = 0$; probabilities are obtained by integrating the PDF over intervals.

A.8 Uniform Distribution (Continuous)

- **Definition and Parameters:**
  - Definition: All values in the interval $[a, b]$ are equally likely. The "maximum entropy" distribution for a bounded interval.
  - Parameters:
    - a: Lower bound.
    - $b > a$: Upper bound.
  - Notation: $X \sim \text{Uniform}(a, b)$ or $X \sim U(a, b)$
  - Support: $[a, b]$
- **Probability Density Function (PDF):**

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

- **Cumulative Distribution Function (CDF):**

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases}$$

- **Moment Generating Function (MGF):**

$$M_X(t) = \begin{cases} \frac{e^{tb} - e^{ta}}{t(b-a)} & \text{if } t \neq 0 \\ 1 & \text{if } t = 0 \end{cases}$$

- **Expected Value and Variance:**

$$E(X) = \frac{a+b}{2}$$
$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

- **Worked Example:**
  - Problem Statement: A random number generator produces values uniformly between 0 and 10. What is $P(3 \leq X \leq 7)$?
  - Solution: $P(3 \leq X \leq 7) = \frac{7-3}{10-0} = 0.4$.
- **Use Cases:**
  - Random number generation, modelling complete uncertainty over a bounded interval, prior distributions in Bayesian analysis.
- **Relation to Other Distributions:**
  - $U(0, 1)$ is the standard uniform; all other continuous distributions can be generated from it via inverse transform sampling.

A.9 Normal (Gaussian) Distribution

- **Definition and Parameters:**
  - Definition: The "bell curve" distribution, fundamental to statistics due to the Central Limit Theorem. Characterised by its symmetric, unimodal shape.
  - Parameters:
    - $\mu \in \mathbb{R}$: Mean (location parameter).
    - $\sigma^2 > 0$: Variance (scale parameter); σ is the standard deviation.
  - Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$ or $X \sim \text{Normal}(\mu, \sigma^2)$
  - Support: $(-\infty, \infty)$
- **Probability Density Function (PDF):**

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

- **Cumulative Distribution Function (CDF):**

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

  where Φ is the standard normal CDF and erf is the error function. No closed-form expression exists; values are typically obtained from tables or numerical computation.
- **Moment Generating Function (MGF):**

$$M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$$

- **Expected Value and Variance:**

$$E(X) = \mu$$
$$\text{Var}(X) = \sigma^2$$

- **Standard Normal Distribution:** When $\mu = 0$ and $\sigma^2 = 1$, we have the *standard normal* $Z \sim \mathcal{N}(0, 1)$:

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

  Standardisation: Any normal variable can be standardised:

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

- **Key Properties:**
  - **68-95-99.7 Rule**: Approximately 68%, 95%, and 99.7% of values lie within 1, 2, and 3 standard deviations of the mean.
  - **Linear combinations**: If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.
  - **Sum of normals**: If $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ are independent, then $\sum X_i \sim \mathcal{N}(\sum \mu_i, \sum \sigma_i^2)$.
- **Worked Example:**

  - Problem Statement: Heights are normally distributed with mean 170 cm and standard deviation 10 cm. What proportion of people are taller than 185 cm?
  - Solution: Standardise: $Z = \frac{185-170}{10} = 1.5$. Then $P(X > 185) = P(Z > 1.5) = 1 - \Phi(1.5) \approx 1 - 0.9332 = 0.0668$, or about 6.7%.

- **Relation to Other Distributions:**
  - **Central Limit Theorem**: The sum (or mean) of many i.i.d. random variables converges to Normal.
  - $Z^2 \sim \chi^2_1$ (chi-squared with 1 degree of freedom).
  - Ratio of independent standard normals follows a Cauchy distribution.
  - Log-normal: If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $e^X \sim \text{LogNormal}(\mu, \sigma^2)$.
- **Use Cases:**
  - Modelling measurement errors, natural phenomena (heights, test scores), financial returns (approximately), inference and hypothesis testing.

A.10 Exponential Distribution

- **Definition and Parameters:**
  - Definition: Models the time between events in a Poisson process, or equivalently, the waiting time until the first event. The continuous analogue of the Geometric distribution.
  - Parameters:
    - $\lambda > 0$: Rate parameter (events per unit time).
  - Notation: $X \sim \text{Exponential}(\lambda)$ or $X \sim \text{Exp}(\lambda)$
  - Support: $[0, \infty)$
    Alternative parametrisation: Some texts use $\beta = 1/\lambda$ (mean/scale parameter).
- **Probability Density Function (PDF):**

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$

- **Cumulative Distribution Function (CDF):**

$$F(x) = 1 - e^{-\lambda x} \quad \text{for } x \geq 0$$

- **Moment Generating Function (MGF):**

$$M_X(t) = \frac{\lambda}{\lambda - t} \quad \text{for } t < \lambda$$

- **Expected Value and Variance:**

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

- **Memorylessness Property:** The Exponential distribution is the only continuous distribution with the memoryless property:

$$P(X > s + t \mid X > s) = P(X > t)$$

"Given that no event has occurred by time s, the probability of waiting at least t more units is the same as starting fresh."

- **Worked Example:**
  - Problem Statement: Light bulbs fail at a rate of 0.1 per year (exponentially distributed). What is the probability a bulb lasts more than 15 years?
  - Solution: $P(X > 15) = e^{-0.1\times 15} = e^{-1.5} \approx 0.223$.
- **Relation to Other Distributions:**
  - Exponential(λ) is Gamma($1, \lambda$)—a gamma with shape parameter 1.
  - Sum of n i.i.d. Exponential(λ) is Gamma(n, λ).
  - Inter-arrival times in a Poisson(λ) process are Exponential(λ).
  - Continuous analogue of Geometric distribution.
- **Use Cases:**
  - Waiting times: time until next customer arrival, time until equipment failure (constant hazard rate), radioactive decay.

A.11 Gamma Distribution

- **Definition and Parameters:**
  - Definition: A flexible family of distributions for positive continuous variables. Models the time until α events occur in a Poisson process (generalisation of Exponential).
  - Parameters:
    - $\alpha > 0$: Shape parameter (number of events).
    - $\beta > 0$: Rate parameter (events per unit time).
  - Notation: $X \sim \text{Gamma}(\alpha, \beta)$
  - Support: $(0, \infty)$
    Alternative parametrisation: Some texts use scale $\theta = 1/\beta$ instead of rate β.
- **Probability Density Function (PDF):**

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad \text{for } x > 0$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$ is the gamma function. For positive integers, $\Gamma(n) = (n-1)!$.

- **Cumulative Distribution Function (CDF):**

$$F(x) = \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}$$

where $\gamma(\alpha, x)$ is the lower incomplete gamma function.

- **Moment Generating Function (MGF):**

$$M_X(t) = \left(\frac{\beta}{\beta - t}\right)^{\alpha} \quad \text{for } t < \beta$$

- **Expected Value and Variance:**

$$E(X) = \frac{\alpha}{\beta}$$
$$\text{Var}(X) = \frac{\alpha}{\beta^2}$$

- **Special Cases:**
  - Gamma$(1, \beta)$ = Exponential(β)
  - Gamma$(n/2, 1/2) = \chi^2_n$ (chi-squared with n degrees of freedom)
  - Gamma(α, α) has mean 1 for any α
- **Worked Example:**
  - Problem Statement: Calls arrive at a rate of 2 per minute. What is the expected time until the 5th call, and its variance?
  - Solution: Time until 5th call is Gamma$(5, 2)$. Mean $= 5/2 = 2.5$ minutes, Variance $= 5/4 = 1.25$ minutes^2.
- **Relation to Other Distributions:**
  - Sum of n i.i.d. Exponential(β) is Gamma(n, β).
  - Continuous analogue of Negative Binomial.
  - Chi-squared is a special case.
  - Gamma is conjugate prior for Poisson rate parameter in Bayesian inference.
- **Use Cases:**
  - Waiting times for multiple events, modelling positive skewed data (insurance claims, rainfall), Bayesian inference.

A.12 Beta Distribution

- **Definition and Parameters:**
  - Definition: A flexible family of distributions on the interval $[0, 1]$. Commonly used to model probabilities, proportions, and percentages.
  - Parameters:
    - $\alpha > 0$: Shape parameter (can be interpreted as "pseudo-counts" of successes + 1).
    - $\beta > 0$: Shape parameter (can be interpreted as "pseudo-counts" of failures + 1).
  - Notation: $X \sim \text{Beta}(\alpha, \beta)$
  - Support: $(0, 1)$ or $[0, 1]$ depending on parameter values

- **Probability Density Function (PDF):**

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \quad \text{for } 0 < x < 1$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function.

- **Cumulative Distribution Function (CDF):**

$$F(x) = I_x(\alpha,\beta) = \frac{B(x;\alpha,\beta)}{B(\alpha,\beta)}$$

where I_x is the regularised incomplete beta function.

- **Moment Generating Function (MGF):**

$$M_X(t) = 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{t^k}{k!}$$

The MGF does not have a simple closed form; it can be expressed in terms of confluent hypergeometric functions.

- **Expected Value and Variance:**

$$E(X) = \frac{\alpha}{\alpha+\beta}$$
$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- **Special Cases and Shape Behaviour:**
  - $\text{Beta}(1,1) = \text{Uniform}(0,1)$
  - $\text{Beta}(\alpha,\alpha)$ is symmetric around 0.5
  - $\alpha < 1, \beta < 1$: U-shaped (bimodal at boundaries)
  - $\alpha > 1, \beta > 1$: Unimodal
  - $\alpha = \beta = 0.5$: Arcsine distribution

- **Worked Example:**
  - Problem Statement: In Bayesian inference, you start with a $\text{Beta}(2,2)$ prior on a coin's probability of heads. After observing 7 heads and 3 tails, what is the posterior distribution?
  - Solution: The posterior is $\text{Beta}(2+7, 2+3) = \text{Beta}(9,5)$, with posterior mean $9/14 \approx 0.643$.

- **Relation to Other Distributions:**
  - Conjugate prior for the Binomial/Bernoulli probability parameter.
  - If $X \sim \text{Gamma}(\alpha,1)$ and $Y \sim \text{Gamma}(\beta,1)$ are independent, then $\frac{X}{X+Y} \sim \text{Beta}(\alpha,\beta)$.
  - Order statistics of $\text{Uniform}(0,1)$ follow Beta distributions.

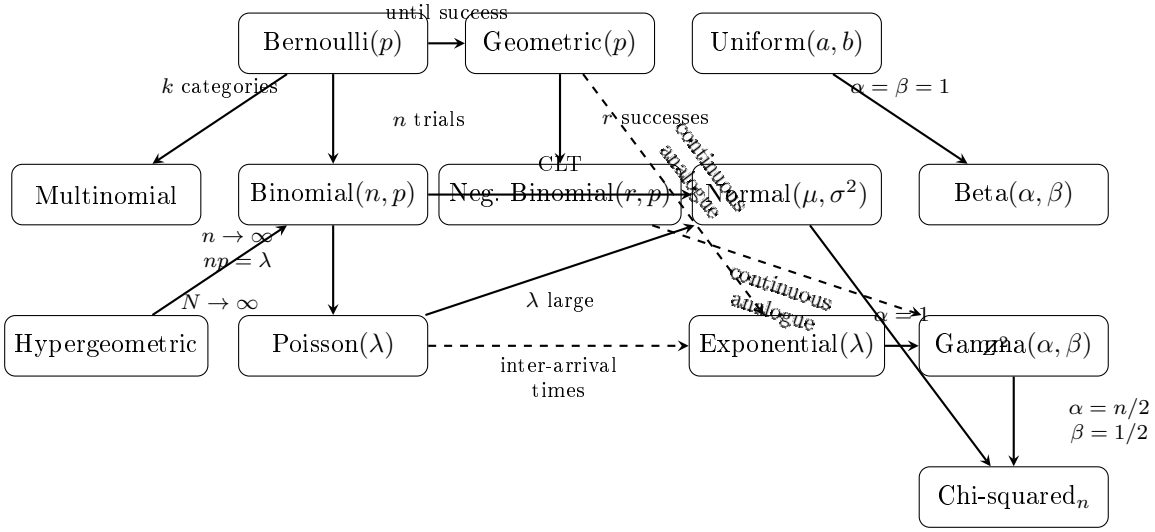
- **Use Cases:**
  - Modelling probabilities and proportions, Bayesian inference (prior for binomial p), A/B testing, project completion percentages.

Part III: Distribution Relationships and Summary

Understanding how distributions relate to each other provides insight into when approximations are valid and reveals the underlying structure of probability theory.

Relationship Diagram

The following diagram illustrates key relationships between distributions:



Solid arrows indicate limiting relationships or special cases. *Dashed arrows* indicate analogous relationships between discrete and continuous distributions.

Summary Table: Discrete Distributions

Distribution	PMF	Mean	Variance	MGF
Bernoulli(p)	$p^k(1-p)^{1-k}$	p	$p(1-p)$	$1-p+pe^t$
Binomial(n, p)	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$	$(1-p+pe^t)^n$
Geometric(p)	$(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$
Neg. Binomial(r, p)	$\binom{k-1}{r-1}p^r(1-p)^{k-r}$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$\left(\frac{pe^t}{1-(1-p)e^t}\right)^r$
Poisson(λ)	$\frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ	$e^{\lambda(e^t-1)}$

Summary Table: Continuous Distributions

Distribution	PDF	Mean	Variance	MGF
Uniform(a, b)	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$
Normal(μ, σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	$e^{\mu t + \frac{\sigma^2 t^2}{2}}$
Exponential(λ)	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda - t}$
Gamma(α, β)	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\left(\frac{\beta}{\beta - t}\right)^\alpha$
Beta(α, β)	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	(complex)

Key Relationships and Approximations

1. **Binomial to Poisson**: When n is large, p is small, and $\lambda = np$ is moderate:

$$\text{Binomial}(n, p) \approx \text{Poisson}(np)$$

   Rule of thumb: $n \geq 20$ and $p \leq 0.05$.

2. **Binomial to Normal** (CLT): When both np and $n(1-p)$ are large (typically ≥ 10):

$$\text{Binomial}(n, p) \approx \mathcal{N}(np, np(1-p))$$

3. **Poisson to Normal**: When λ is large (typically ≥ 30):

$$\text{Poisson}(\lambda) \approx \mathcal{N}(\lambda, \lambda)$$

4. **Hypergeometric to Binomial**: When N is much larger than n (finite population correction negligible):

$$\text{Hypergeometric}(N, K, n) \approx \text{Binomial}(n, K/N)$$

5. **Conjugate Pairs** (Bayesian inference):
   - Beta prior + Binomial likelihood $\to$ Beta posterior
   - Gamma prior + Poisson likelihood $\to$ Gamma posterior
   - Normal prior + Normal likelihood $\to$ Normal posterior

Memoryless Distributions

Only two distributions possess the memoryless property $P(X > s + t \mid X > s) = P(X > t)$:

- **Geometric** (discrete): Number of trials until first success
- **Exponential** (continuous): Waiting time until first event

This property implies a "constant hazard rate"—the probability of an event occurring in the next instant is independent of how long you have already waited.

Appendix B

Cheat Sheet I

B.1 Session 1: Probability Theory

De Morgan's Law for Unions & Complements:

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

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

Multiplication rule: think in terms of trees $\rightarrow n^k$
NB chronological order doesn't actually matter here — counter intuitive

Combinations = when order does not matter
Permutations = when order/position matters $n!$

	Order Matters	**Order Doesn't Matter**
With Replacement	n^k	$\binom{n+k-1}{k}$
Without Replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$

**NB: order matters, sampling w/o replacement also written as* $n \cdot (n-1) \cdot (n-2) \cdots (n-k+1)$

Birthday Problem — counting complement
What's the probability of no matching birthdays?
This amounts to sampling the days of the year without replacement:

$$\begin{aligned} P(\text{no birthday match}) &= \frac{\text{number of ways to not repeat birthdays}}{\text{number of total possibilities}} \\ &= \frac{365 \times 364 \times \cdots \times (365-k+1)}{365^k} \\ P(\text{birthday match}) &= 1 - \frac{365 \times 364 \times \cdots \times (365-k+1)}{365^k} \end{aligned}$$

Factorial Overcounting:
when arranging n distinct items: $n!$ ways to do so...
...BUT if k items are identical $\rightarrow$ divide by $k!$

- when assigning to multiple groups: we are overcounting by the number of groups factorial $\rightarrow$ divide by groups factorial

- STATISTICS : overcounts Ss, Ts, Is, there are 3 Ss → divide 3!3!2!
- same with the multinomial: you divide by the repeats factorial: eg number of ways to sort 10 ppl into a group: $\frac{10!}{3!3!4!}$
- problem of non-repeat sampling (ie bday problem)
  - numerator = successes = order matters, w/o replacement $(n \times (n-1) \times (n-2) \times \cdots \times (n-k+1))$
  - Denominator = total = order matters, w/ replacement (n^k)

When working with multiple events (eg 10 heads); often easier to say what is the probability of that NEVER happening: ie 1 single event...
... if something seems tedious: check its complement

Any probability function P must satisfy the following two axioms:

- $P(\emptyset) = 0, P(S) = 1.$
- If $A_1, A_2, \ldots$ are disjoint events, then

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j)$$

Properties of Probabilities

- $P(A^c) = 1 - P(A)$
- If $A \subseteq B$, then $P(A) \leq P(B)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

B.2 Session 2: Conditional Probability & Random Variables

B.2.1 Conditional Probability

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

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

Intuitively: Venn diagram overlap, renormalised for $P(B)$

Bayes Rule

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

LOTP

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

Putting them together:
With Bayes: when applying LOTP to bottom: ensure you multiply by prob of the conditional.
- Eg: $P(T_c|D) \times P(D) + P(T_c|D_c) \times P(D_c)$
- Eg: $P(\text{ObservedData}|\text{Coin}_1) \times P(\text{Coin}_1) + P(\text{OD}|\text{Coin}_2) \times P(\text{Coin}_2) + P(\text{OD}|\text{Coin}_3) \times P(\text{Coin}_3)$

Bayes' Rule w/ Extra Conditioning:

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

LOTP w/ Extra Conditioning:

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

Independence of Events

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

NB:

- in a Venn diag, if the overlap is equal to the product of $P(A)$ and $P(B)$.
- in a Contingency table, this means cells equal to marginals.

Equivalent to

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

- Independence is symmetric
- If A and B are independent:
  - A and B^c are independent,
  - A^c and B are independent,
  - A^c and B^c are independent

This doesn't carry through for conditional independence.

Independence of 3 events:
Needs to be more than pairwise independence (conditions 1–3)

$$P(A \cap B) = P(A)P(B) \tag{B.1}$$
$$P(A \cap C) = P(A)P(C) \tag{B.2}$$
$$P(B \cap C) = P(B)P(C) \tag{B.3}$$
$$P(A \cap B \cap C) = P(A)P(B)P(C) \tag{B.4}$$

Conditional Independence

- Conditional independence given E does not imply conditional independence given E^c
- Conditional independence does not imply independence
- Independence does not imply conditional independence

B.2.2 Random Variables

- **r.v.** is a function from the sample space S to the real number line $\mathbb{R}$; assigns a numerical value $X(s)$ to each possible outcome s of the experiment.
- **Support of X** is defined as all the values x such that $P(X = x) > 0$.
- **PMF** of X is the function $p_X(x) = P(X = x)$. This is positive if x is in the support of X, and 0 otherwise.

Building PMF:

1. Immediately write $P(X = k) = \dots$
2. Enumerate all possible outcomes $(X = 1, X = 2, X = 3 \dots)$.
   consider what support of X could be
3. calculate probabilities for each outcome. (*Example: P(X=0) = P(TT) = 1/4)*
   is there a functional form you can generalise to?
   NB: if it is NOT a binary outcome, might have to permute

PMF:

- $P(X = 0) = \frac{1}{4}$
- $P(X = 1) = 1/2$
- $P(X = 2) = 1/4$
- and $p_X(x) = 0$ for all other values of x.

PMFs must (1) be non negative, and (2) sum to 1.

Bernoulli:

$$P(X = k) = \begin{cases} 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$$

Binomial:

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

Discrete Uniform:

$$P(X \in A) = \frac{|A|}{|C|}$$

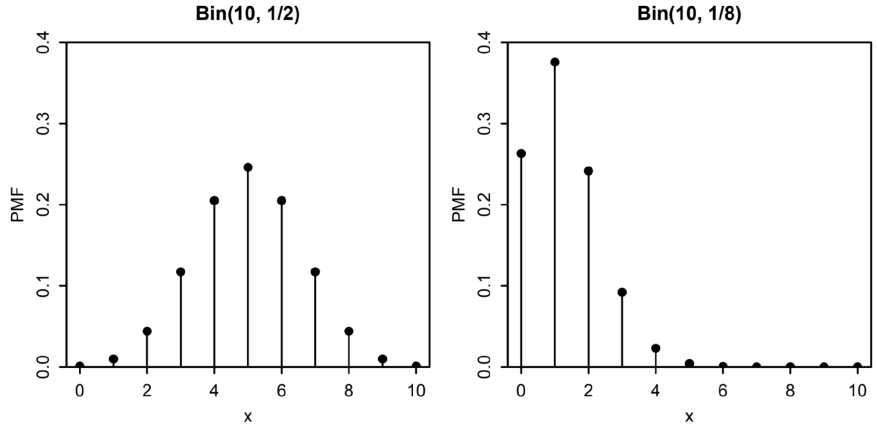


Figure B.1: Binomial PDFs

CDFs CDF of X, is the function $F_X(x) = P(X \leq x)$.

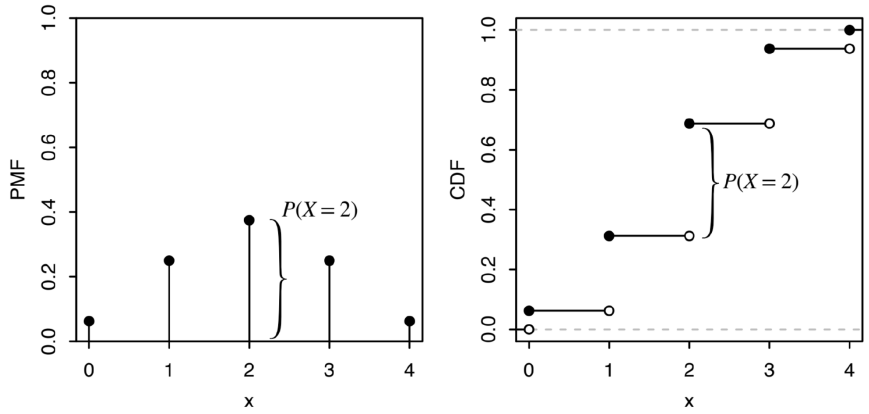


Figure B.2: PMF, CDF of $X \sim \text{Bin}(4, 1/2)$

B.3 Session 3: Joint r.v.s

- **Joint probability:** $P(A \cap B)$ or $P(A, B)$
- **Marginal (unconditional) probability:** $P(A)$
- **Conditional probability:** $P(A|B) = P(A, B)/P(B)$
- **Intersections via conditioning:** $P(A, B) = P(A)P(A|B)$
- **Unions via inclusion-exclusion:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

B.3.1 Independence of joint r.v.s

Continuous r.v.s

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

Discrete r.v.s

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Conditional Independence

$$P(X \leq x, Y \leq y|Z = z) = P(X \leq x|Z = z)P(Y \leq y|Z = z)$$

B.3.2 Expectation

= weighted avg of possible values X can take:

$$E(X) = \sum_x x \cdot \underbrace{P(X = x)}_{\text{PMF at } x}$$

Eg 2x coin flip (heads)

$$E(X) = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$$

Eg Bernoulli: $X \sim \text{Bern}(p)$:

$$E(X) = 1p + 0(1 - p) = p$$

Linearity of Expectation:

1. $E(cX) = cE(X)$
2. $E(X + Y) = E(X) + E(Y)$

B.3.3 Variance

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

Variance facts:

- $\text{Var}(c) = 0$ for any constant c
- $\text{Var}(X + c) = \text{Var}(X)$ for any constant c
- $\text{Var}(cX) = c^2 \text{Var}(X)$ for any constant c $\leftarrow$ **NB**
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ only if X and Y are independent.
  Caution: unlike expectation, variance is not linear
  - $\text{Var}(cX) \neq c\,\text{Var}(X)$
  - $\text{Var}(X + Y) \neq \text{Var}(X) + \text{Var}(Y)$ (in general)

  Except when the two r.v.s are independent! then $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$

B.3.4 Marginal & Conditional Joint PMFs

Marginal PMF of X Sum over all y; marginalise out Y

$$P(X = x) = \sum_y P(X = x, Y = y)$$

if interested in $(Y = 1) \rightarrow$ sum over all Xs, that $(Y = 1, X = x)$

Conditional PMF of Y Joint divided by marginal.

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

Another way to think of it (same as above): Conditional for joint r.v.s is when

$$\begin{aligned} &P(Y = 1|X = 1) \\ &\quad = P(Y = 1 \cap X = 1)/P(X = 1) \\ &\quad = (1/30)/(5/30) \end{aligned}$$

$$\begin{aligned} &P(X = 1|Y = 2) \\ &\quad = P(X = 1 \cap Y = 2)/P(Y = 2) \\ &\quad = (4/30)/(24/30) \\ &\quad = 1/6 \end{aligned}$$

NB: where the Marginal X vs Marginal Y is... need to know which to make denominator for conditionals

	$Y = 1$	$Y = 0$	Marginal X
$X = 1$	$\frac{5}{100}$	$\frac{20}{100}$	$\frac{25}{100}$
$X = 0$	$\frac{3}{100}$	$\frac{72}{100}$	$\frac{75}{100}$
Marginal Y	$\frac{8}{100}$	$\frac{92}{100}$	1

Table B.1: Contingency table for X and Y with Marginal Distributions

Contingency table is just another way to express PMF for joint variables; it expresses how they move together
marginal is summing over the other thing
conditional is fixing the other thing at some value
joint is how they move together
Test for Independence:

$$\text{If independent: } P(X = x, Y = y) = P(X = x)P(Y = y)$$

so if the **cell value, is the product of the marginals**
for independence: every cell needs to be the product of the marginals.

B.4 Session 4: Calculus

1. **Rule 1: Powers**: $\frac{d}{dx}x^n = nx^{n-1}$
2. **Rule 2: Sum/Differences**:
$$\frac{d}{dx}(f(x) \pm g(x)) = \frac{d}{dx}f(x) \pm \frac{d}{dx}g(x)$$
3. **Rule 3: Constant Multiples**
$$\frac{d}{dx}[kf(x)] = k\frac{d}{dx}f(x)$$
4. **Rule 4: Products**
$$\frac{d}{dx}[g(x)f(x)] = g'(x) \cdot f(x) + g(x) \cdot f'(x)$$
5. **Rule 5: Quotients**
$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{g(x) \cdot f'(x) - f(x) \cdot g'(x)}{g(x)^2}$$
6. **Rule 6: Chain**
If y is a function of u, and u is a function of x, then:
$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$
ALWAYS REMEMBER TO PLUG THE u VALUE BACK IN
7. **Rule 7: Natural Exponential**
$$y(x) = e^x \rightarrow \frac{dy}{dx} = e^x$$
8. **Rule 8: Natural Logarithms**
$$y = \ln(x) \rightarrow \frac{dy}{dx} = \frac{1}{x}$$
9. **General: Exponential Functions**
Power rule is for when the exponent is a **constant**.
Exponential functions take x as the exponent itself.
$$\frac{d}{dx}a^x = \ln(a) \times a^x$$
$$\frac{d}{dx}a^{bx} = b \times \ln(a) \times a^{bx}$$
Example: a^x
10^x becomes $\ln(10) \times 10^x$:
$$f(x) = \frac{10^x}{\ln(10)}$$
$$f'(x) = \ln(10) \times 10^x \times \frac{1}{\ln(10)} = 10^x$$

Example: a^{bx}

$$f(x) = 2^{4x} + 4x^2$$
$$f'(x) = 4\ln(2) \times 2^{4x} + 8x$$

B.5 MLE

MLE steps:

1. (Identify the distribution: write out the PMF)
2. **Write the likelihood as a function of the data:**

   $$L(\theta) = P(x_1, x_2 \ldots . x_i)$$

   becomes:

   $$L(x_1, x_2 \ldots . x_i; \theta) = \prod_{j=1}^{n} \text{the PMF with } \theta \text{ substituted in for parameter of interest / } p???$$

3. **Expand as products**: we assume each event is i.i.d, so the likelihood is the product of each
   - First, write as a series of products for each r.v.:

     $$L(\lambda) = P(X = 1) \times P(X = 3) \times P(X = 1) \times \ldots$$

   - Then expand each using the relevant distribution with the data for the r.v plugged in:

     $$\frac{e^{-\lambda}\lambda^1}{1!} \times \frac{e^{-\lambda}\lambda^3}{3!} \times \frac{e^{-\lambda}\lambda^1}{1!} \times \ldots$$

   - collect terms and simplify as much as possible.
4. **Take the log-likelihood:**

   $$\ell(\lambda) = \log(L(\lambda))$$

   Use the properties of logs to break it up into its components parts to reshape it into nice $+/-$ equation (ie without products or quotients... use these rules of logs!!)
5. **Derive the log with respect to parameter of interest:**
   - constants (ie not dependent on parameter of interest) drop out.
   - parameter terms differentiate as usual.
   - derivative of log =

     $$\frac{d}{dx}\ln(x) = \frac{1}{x}$$
     $$\frac{d}{dx}\ln(2x) = \frac{1}{2x} \times 2 = \frac{1}{x}$$

     NB chain rule here

     $$\frac{d}{d\lambda}7\log(\lambda) \ = 7 \times \frac{1}{\lambda} = \frac{7}{\lambda}$$

6. **Set to 0**
7. Solve for θ

Log rules:

$$\log(ab) = \log(a) + \log(b)$$
$$\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$$
$$\log\left(\prod_{j=1}^{n} x_j\right) = \sum_{j=1}^{n} \log(x_j)$$
$$\log(a^b) = b\log(a)$$
$$\log(1) = 0$$
$$\log(e^x) = x$$
$$\frac{d}{dx}\ln(x) = \frac{1}{x}$$

Exponent rules:

$$a^m \times a^n = a^{m+n}$$
$$\frac{a^m}{a^n} = a^{m-n}$$
$$(a^m)^n = a^{m\times n}$$
$$(ab)^n = a^n \times b^n$$
$$\left(\frac{a}{b}\right)^n = \frac{a^n}{b^n}$$
$$a^0 = 1 \quad (\text{where } a \neq 0)$$
$$a^{-n} = \frac{1}{a^n} \quad (\text{where } a \neq 0)$$
$$a^1 = a$$

B.6 Taylor Series Approximation

1. Find derivatives (n many, depending on polynomial degree specified)
2. Evaluate them ($a = \{x\}$)
3. Insert each of these into the Taylor formula...
4. ...Simultaneously: plug in a values with the given centring coordinate. This will leave various x values: this is your line formula.

Appendix C

Cheat Sheet II

C.1 Week 6: Continuous R.V.s Meets Probability

C.1.1 Continuous r.v.s

- r.v. has continuous distribution if its CDF is differentiable
- PDF of X is derivative of CDF: $f(x) = F'(x)$
- CDF of X is integral of PDF (ie the area under the curve)
- unlike discrete r.v.s, for continuous: $P(X = x) = 0$ for all x
  - PDF of X gives probability density
  - BUT, CDF remains interpretable as $P(X \leq x)$ (as it represents AUC under PDF)
  - The probability that a continuous random variable falls within a particular interval is given by area under the PDF curve over that interval.
- valid PDF conditions: 1) non-negative, 2) integrates to 1

C.1.2 Expectation of continuous r.v.

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

= centre of mass / balance point
"To find the expected value of a continuous random variable, take every possible value that variable can have, multiply each by the probability of that value occurring, and then sum all these products together."

C.1.3 Uniform, continuous

PDF

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

Proof it is a valid PDF: it sums (integrates) to 1: rectangle $(b-a) \times (1/(b-a)) = 1$

CDF

$$F(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x-a}{b-a} & \text{if } a < x < b \\ 1 & \text{if } x \geq b \end{cases}$$

C.1.4 Normal

μ = location; σ^2 = scale (variance)

C.1.5 Standardisation

Can transform any Normal-distributed r.v. into a Standard Normal (this is a z-score):

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

NB: r.v. itself does NOT have to be Normal distribution — can be ANY distribution; so long as it is i.i.d → you can take the sample mean, standardise it, and that will be normally distributed.

C.1.6 Exponential

- time to wait before first success
  - = continuous analog of the Geometric (number of failures until first success in a sequence of Bernoulli trials)
  - λ = rate of success for some unit of time
  - memorylessness

C.1.7 Joint Distributions of Continuous r.v.s

- Joint CDF: $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$
- Joint PDF: $f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$ — first take the partial derivative of $F_{X,Y}(x, y)$ with respect to x, and then take the partial derivative of the result with respect to y
  - e.g. CDF: $F(x, y) = \frac{1}{2}x^2y^3 \rightarrow$ PDF: $f(x, y) = 3xy^2$
- Marginal Distribution of X from Joint PDF, integrate over all values of Y:
  - $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$
- Conditional PDF (of Y, given $X = x$):
  - $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$
  - Conditional = Joint PDF / Marginal

C.1.8 Bayes Rule and LOTP for continuous r.v.s

- Bayes rule: $f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y) f_Y(y)}{f_X(x)}$
- LOTP: $f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y)\, dy$

C.2 Week 7: Continuous R.V.s II

C.2.1 Covariance

- move together +; move opp -; independent 0
- "expectation of the product, minus the product of the expectations"
- $\text{Cov}(X, Y) = E((X - E[X])(Y - E[Y]))$
- $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$
- independence: covariance is 0: $E[XY] = E[X]E[Y]$

Some Covariance rules

1. $\text{Cov}(X, X) = \text{Var}(X)$
2. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
3. $\text{Cov}(X, c) = 0$ for any constant c
4. $\text{Cov}(aX, Y) = a \cdot \text{Cov}(X, Y)$ for any constant a
5. $\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$
6. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \cdot \text{Cov}(X, Y)$
7. **additional:** $E[E[X]]$: expectation of a constant is just a constant.

C.2.2 Correlation

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$

Correlation coefficient: $\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$

corr between -1 and 1

Due to non-linearity in the data (think of parabola, $y = x^2$ — perfectly dependent, but uncorrelated):

- Independence $\rightarrow$ uncorrelated
- Uncorrelated $\not\rightarrow$ Independence

Independence defined as: $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$
For Independent variables: $E[XY] = E[X]E[Y]$
For independent variables ($= 0$ covariance), their correlation is 0. But not necessarily vice versa.

Example of proving independence:

$$E(X) = \frac{1}{2}(1) + \frac{1}{2}(2) = 1.5$$
$$E(Y) = \frac{1}{2}(3) + \frac{1}{2}(4) = 3.5$$

$E[XY]$ can be calculated by considering all combinations of X and Y
We have four combinations: (1, 3), (1, 4), (2, 3), and (2, 4). Each combination occurs with a probability of 1/4, since the probabilities of X and Y are each 1/2.

$$E[XY] = \frac{1}{4}(1 \times 3) + \frac{1}{4}(1 \times 4) + \frac{1}{4}(2 \times 3) + \frac{1}{4}(2 \times 4) = 5.25$$
$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = 5.25 - (1.5 \times 3.5) = 0$$

Since the covariance is 0, it suggests that X and Y are uncorrelated

C.2.3 Law of Large Numbers

As n grows large, the sample mean $\bar{X}$ converges to the true mean μ:

- Sample mean (if i.i.d): $\bar{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}$
- sample mean is itself r.v.:
  - Expectation = μ (i.e. the population mean — the sample mean and the population mean converge as n grows)
  - Variance = $\frac{\sigma^2}{n}$
  - Standard Deviation = $\frac{\sigma}{\sqrt{n}}$

C.2.4 Central Limit Theorem

The standardised sample mean (standardised $\bar{X}$) converges in distribution to the standard Normal as $n \to \infty$:

1. subtract expectation μ
2. dividing by standard deviation = $\frac{\sigma}{\sqrt{n}}$

So, regardless of underlying distribution: $\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0, 1)$

Also sum:

$$\sum_{i=1}^{n} X_i \xrightarrow{d} \mathcal{N}(n\mu, n\sigma^2) \quad \text{as} \quad n \to \infty$$

C.2.5 Example: Normal Approximation to the Binomial

Recalling that the Binomial(n, p) is the sum of n Bernoullis with probability p, we can even use the Normal distribution to approximate the Binomial.

$$\sum_{i=1}^{n} X_i \approx \mathcal{N}(n\mu, n\sigma^2)$$

and recalling that the mean of a Bernoulli is p and its variance is $p(1-p)$, we can use the CLT to say:

$$\sum_{i=1}^{n} X_i \approx \mathcal{N}(np, np(1-p))$$

NB:
Variance:

$$E[(X-\mu)^2] = \sum_i (x_i - \mu)^2 \cdot P(x_i)$$

for uniform:

$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

C.3 Lab 7: EM Algorithm

1. Initialise parameters
   - μ_1, μ_2 = means
   - σ_1, σ_2 = standard deviations
   - π_1, π_2 = mixing proportions (the initial prob of being in one distribution)
2. E-step: Expectation — compute responsibilities of each data point (= calculate the γ of each data point: the prob that each data point belongs to each component given current parameter values)
3. M-step: Maximisation — update the parameters based on the responsibilities (= MLE of each parameter given the γ values for each data point (the parameters defined as some function involving sums of gamma-data point, so will give a single value).
4. Evaluate the new log-likelihood with new (i) parameter, (ii) responsibilities.
5. Check for convergence

NB: how Bayes rule used to obtain gammas:

$$\begin{aligned}\Pr(z_{1i} = 1|x_i) &= \frac{f(x_i|z_{1i}=1)\Pr(z_{1i}=1)}{f(x_i)} \\ &= \frac{\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1)}{\pi_1 \mathcal{N}(x_i|\mu_1, \sigma_1) + \pi_2 \mathcal{N}(x_i|\mu_2, \sigma_2)}\end{aligned}$$

C.4 Week 8: Matrix Algebra

- Matrix dimensions $m \times n$ = m rows, n columns
- row values $\rightarrow$ column, column values $\rightarrow$ row
- to add 2 matrices: need same dimensions
- $\mathbf{a}^T\mathbf{b} = [a_1, a_2, a_3] \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = a_1b_1 + a_2b_2 + a_3b_3 \ \ldots$ NB: output is scalar
- Matrix multiplication
  - conformable: $m \times n$ and $n \times p$ (LHS columns = RHS rows) $\rightarrow$ output $m \times p$ — the outer values
  - for each element a_{ij}: sum of the products of the elements of the corresponding **row of A** and the corresponding **column of B**.
  - so for item a_{12}:
    - all the elements of row 1 from A
    - all the elements of column 2 from B
    - multiplied by each other
    - summed
- Transpose Facts
  - $(A^T)^T = A$
  - $(A + B)^T = A^T + B^T$
  - $(AB)^T = B^TA^T$
  - $\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\mathbf{a}$
- Identity matrix
  - $I_n\mathbf{x}_n = \mathbf{x}_n$
  - $I_mA_{m\times n} = A_{m\times n}$ and $A_{m\times n}I_n = A_{m\times n}$
  - $A^{-1}A = I_n$
- Vector Norms (measure of magnitude)
  - L1: $\|\mathbf{x}\|_1 = \sum_i |x_i|$ = take absolute values of elements before summing
  - L2: $\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$ = square root of the sum of the squares of the vector's elements

C.5 Lab 8: Regression

C.5.1 Linear Regression

The objective function for least squares regression is:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip})^2$$

Representing the squared error minimisation problem (the cost function) in Matrix form:

$$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

This expanded **cost function** (expressed in matrix notation) is what we will be minimising (by taking first derivative, set to zero)

(1) Setting to 0 → (2) taking first derivative with respect to $\boldsymbol{\beta}$ → (3) simplifying:

$$-2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0$$

Solving for beta:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

C.5.2 Penalised regression

The objective function in Lasso Regression becomes:

$$\text{Minimize} \left(\text{Residual Sum of Squares} + \lambda \sum_{i=1}^{p} |\beta_i| \right)$$

The objective function in Ridge Regression is:

$$\text{Minimize} \left(\text{Residual Sum of Squares} + \lambda \sum_{i=1}^{p} \beta_i^2 \right)$$

C.6 Week 9: Linear Algebra II

- Linear Dependence: $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \mathbf{0}$
  - 1) set up the equation $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \mathbf{0}$
  - 2) solve the system of equations: take one $c_n\mathbf{v}_n$ over to the other side so you can write one of the $c_n\mathbf{v}_n$ as a function of the others, then plug back in, etc.
    I STILL DONT GET HOW TO SOLVE THESE SYSTEMS?
- Span = all linear combinations of the vectors in the set.
  - S is a spanning set for V if all the dimensions of V can be represented by linear combinations of S. A Spanning Set S must contain at least as many elements as the linearly independent vectors from V.
  - There are exactly n orthogonal / linearly independent vectors in $\mathbb{R}^n$ — these are the basis vectors (another way to phrase: number of linearly independent vectors in V define dimension of its vector space)
- Determinant:
  - **non-zero (for square matrix) → linear independence → invertible**
  - determinant of a matrix reflects how the transformation changes the dimensions of the space

  - characteristic polynomial of a matrix, used to find its eigenvalues, is derived from its determinant
  - 2x2 Matrix: $\det(A) = ad - bc$
  - 3x3 matrix: $\det(A) = a(ei - fh) - b(di - fg) + c(dh - eg)$
- inverse conditions: $(AA^{-1} = A^{-1}A = I_n)$
  1. matrix is square
  2. columns are linearly independent (full rank)
  3. non-zero determinant
- eigenvalue-vector pairings:
  - $A\mathbf{v} = \lambda\mathbf{v}$
  - eigenvalue equation: $(\lambda I_n - A)\mathbf{v} = \mathbf{0}$
  - characteristic equation: $\det(A - \lambda I) = 0$ or $\det(\lambda I_n - A) = 0$ finds eigenvalues
  - existence conditions: 1) square, 2) a lambda scalar that fulfils characteristic equation (i.e. linearly independent columns)
  - eigenspace of λ, E_λ, is all vectors $\mathbf{v}$ that satisfy the condition $A\mathbf{v} = \lambda\mathbf{v}$
- Eigendecomposition: $A = Q\Lambda Q^{-1}$
  - A = square matrix $n \times n$
  - Λ = diagonal matrix with eigenvalues of A, $n \times n$
  - Q = matrix whose columns are eigenvectors of A, corresponding to eigenvalues in Λ
- SVD = $A = U\Sigma V^T$
  - $A \in \mathbb{R}^{m \times n}$
  - U is a matrix containing left singular vectors of A (contains the m eigenvectors of the square matrix AA^T) $(m \times m)$
  - V^T is a matrix containing the right singular vectors of A (contains the n eigenvectors of the square matrix A^TA) $(n \times n)$
  - Σ is a matrix where the diagonal entries are the singular values of A (square roots of the eigenvalues of A^TA (or AA^T) on the diagonal) $(m \times n)$
  - ...
  - U & V^T: how much matrix A rotates an object
  - Σ: associated scaling
    - number of non-zero singular values gives rank of the matrix (number of linearly independent columns) — full rank: nothing redundant.

Dimensions:

- $A = m \times n$
- $U = m \times m$
- $\Sigma = m \times n$
- $V^T = n \times n$
- + NUMBER OF LINEARLY INDEPENDENT COLUMNS OR ROWS NEEDED TO FORM BASIS FOR COLUMN OR ROW SPACE OF A = RANK OF THE MATRIX A = THE NUMBER OF NON ZERO SINGULAR VALUES IN Σ

C.7 Lab 9: PCA

C.7.1 As Variance Maximisation

$$\mathbf{z}_1 = \phi_{11}x_1 + \phi_{21}x_2 + \ldots + \phi_{p1}x_p$$

Maximising variance of the vector of linear combinations $\mathbf{z}_1$ (the first principal component), i.e. maximising the variance across its elements (which represent linear combination associated with each observation, consisting of scaled variables, summed)

$$\begin{aligned}\max_{\phi_{11},\phi_{21},\ldots,\phi_{p1}} \text{Var}(\mathbf{z}_1) &= \max_{\phi_{11},\phi_{21},\ldots,\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n}(z_{i1} - \bar{\mathbf{z}}_1)^2 \\ &= \max_{\phi_{11},\phi_{21},\ldots,\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n} z_{i1}^2\end{aligned}$$

Constrain associated loadings:

$$\sum_{j=1}^{p}\phi_{j1}^2 = 1$$

C.7.2 As Eigendecomposition

- derive a Variance-covariance matrix
- Eigendecompose that: $S\boldsymbol{\phi}_1 = \lambda_1\boldsymbol{\phi}_1$
- λ_1 = eigenvalue for first principal component
- $\boldsymbol{\phi}_1$ = loadings for first principal component
- i think you would then multiply the original data matrix by the loadings vector to get the first principal component.

C.8 Week 10: Optimisation

- 2nd deriv test: pos = min; neg = max
- Lagrangian: $\mathcal{L}(x, y, \lambda) = f(x, y) + \lambda g(x, y)$
- constrained opti: 1) set up Lagrangian, 2) Take the first derivatives $\frac{\partial\mathcal{L}}{\partial x}$, $\frac{\partial\mathcal{L}}{\partial y}$, and $\frac{\partial\mathcal{L}}{\partial\lambda}$ and set them equal to 0, 3) solve
  PRACTICE SOLVING
- Hessian: each element a 2nd order partial derivative of a function. Has same dimensions as number of variables in the function:
  - all eigenvalues pos: local min
  - all eigenvalues neg: local max
  - mixed: saddle point
- Gradient (of a function)