Probability Theory

Lecture Notes

Licence L3 — 2025–2026

Yaë Ulrich Gaba

“Probability theory is nothing but common sense reduced to calculation.”
— Pierre-Simon Laplace

March 25, 2026

Contents

Preface

Aims of the Course

This **Probability Theory** course is aimed at second-year undergraduate students in mathematics, computer science, or the physical sciences. It provides a rigorous introduction to the mathematical theory of probability, grounded in the Kolmogorov axioms, while maintaining strong probabilistic intuition throughout.

Probability occupies a central place in modern mathematics and its applications. From statistical physics to machine learning, from quantitative finance to molecular biology, probabilistic modelling has become an indispensable tool for understanding and quantifying uncertainty.

Prerequisites

To approach this course effectively, students should be comfortable with:

- **Real analysis:** sequences, series, integrals (Riemann), uniform convergence, basic dominated convergence results.

- **Linear algebra:** vector spaces, matrices, eigenvalues (needed for the Markov chains chapter).

- **Set theory:** set operations, countability, cardinality.

- **Combinatorics:** permutations, combinations, multiplication principle.

Course Organisation

The course comprises twelve chapters, arranged in a logical progression:

Chapters 1–2: Foundations. We define the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, state the Kolmogorov axioms, and develop conditional probability and independence. These two chapters establish the axiomatic framework on which everything else rests.

Chapters 3–5: Random variables. We introduce discrete and then continuous random variables, with their distributions, cumulative distribution functions, and densities. Chapter 5 develops expectation, variance, and moments — the fundamental tools of probabilistic calculation.

Chapters 6–8: Distributions and transformations. Chapter 6 presents a catalogue of standard distributions (binomial, Poisson, normal, exponential, gamma, etc.). Chapter 7 studies functions of random variables, and Chapter 8 introduces joint distributions, covariance, and correlation.

Chapters 9–11: Limit theorems. This is the heart of the theory: the law of large numbers (weak and strong) in Chapter 9, the central limit theorem in Chapter 10, and characteristic functions in Chapter 11 — a powerful tool that yields an elegant proof of the CLT.

Chapter 12: Introduction to Markov chains. The final chapter opens the door to stochastic processes by studying Markov chains on finite or countable state spaces.

Methodology and Notation

Throughout the course we adopt the following conventions:

- **Definitions** are boxed and numbered. Every new concept is first defined rigorously before being illustrated by examples.
- **Theorems** are stated precisely and, wherever possible, proved in full. When a proof goes beyond the scope of the course, we provide a sketch or a reference.
- **Examples** are plentiful and detailed. They are an essential complement to the abstract theory.
- **Exercises** at the end of sections allow students to check their understanding and develop computational techniques.

Main notation.

Symbol	Meaning
Ω	Sample space (set of all outcomes)
$\mathcal{F}$	σ-algebra (collection of events)
$\mathbb{P}$	Probability measure
$\mathbb{E}[X]$	Expectation of the random variable X
$\mathrm{Var}(X)$	Variance of X
$\mathrm{Cov}(X, Y)$	Covariance of X and Y
F_X	Cumulative distribution function of X
f_X	Probability density function of X
$\varphi_X(t)$	Characteristic function of X
$X \sim \mathcal{L}$	X has distribution $\mathcal{L}$
$\xrightarrow{\text{a.s.}}$	Almost sure convergence
$\xrightarrow{\mathbb{P}}$	Convergence in probability
$\xrightarrow{\mathcal{L}}$	Convergence in distribution

Historical Perspective

Probability theory has a rich history that deserves brief mention. The earliest formal work dates to the correspondence between **Blaise Pascal** and **Pierre de Fermat** in 1654 concerning the problem of points. In the eighteenth century, **Jakob Bernoulli** proved the first version of the law of large numbers in his *Ars Conjectandi* (1713), while **Abraham de Moivre** obtained a primitive form of the central limit theorem.

In the nineteenth century, **Pierre-Simon de Laplace** synthesised the probabilistic knowledge of his era in his *Théorie analytique des probabilités* (1812). **Pafnuty Chebyshev**, **Andrey Markov**, and **Aleksandr Lyapunov** developed the analytic tools that would lead to the modern limit theorems.

The decisive revolution came in 1933, when **Andrey Kolmogorov** published his *Grundbegriffe der Wahrscheinlichkeitsrechnung*, founding probability theory on measure theory. This axiomatisation, which we adopt throughout this course, resolved numerous paradoxes and opened the way to the modern theory of stochastic processes.

Advice to Students

1. **Work regularly.** Probability cannot be learnt the night before the exam. Each chapter builds on the previous ones.

2. **Do the exercises.** Passive understanding of the theory is not enough. It is by solving problems that one develops probabilistic intuition.

3. **Draw pictures.** Sketch sets, plot densities, draw distribution functions. Visualisation is a precious ally.

4. **Check special cases.** When you obtain a general formula, test it on simple examples. This catches errors and strengthens understanding.

5. **Distinguish discrete from continuous.** Many errors arise from confusing sums with integrals, probability mass functions with densities. Always be aware of the setting in which you are working.

Main References

This course draws on many classical texts, including:

- J. Jacod and P. Protter, *Probability Essentials*, Springer.

- P. Billingsley, *Probability and Measure*, Wiley.

- W. Feller, *An Introduction to Probability Theory and Its Applications*, Vols. I and II, Wiley.

- G. Grimmett and D. Stirzaker, *Probability and Random Processes*, Oxford.

- R. Durrett, *Probability: Theory and Examples*, Cambridge.

- J. Rosenthal, *A First Look at Rigorous Probability Theory*, World Scientific.

- S. Ross, *A First Course in Probability*, Pearson.

Happy reading, and good luck with your studies.

Chapter 1

Probability Spaces

The calculus of probabilities was born in the gambling houses of the seventeenth century. Blaise Pascal and Pierre de Fermat, in their famous 1654 correspondence, sought to solve the "problem of points": how to fairly divide the stakes of a game that is interrupted. But it took nearly three centuries for probability theory to acquire rigorous mathematical foundations. It was Andrei Kolmogorov who, in 1933 in his *Grundbegriffe der Wahrscheinlichkeitsrechnung*, laid down the definitive axioms: a probability is nothing but a measure of total mass 1 on a measurable space.

This first chapter builds these foundations stone by stone: the sample space Ω of all possibilities, the σ-algebra $\mathcal{F}$ of observable events, and the probability measure $\mathbb{P}$ that assigns each event a degree of likelihood.

Intuition

This opening chapter lays the mathematical foundations of probability theory. We define the notion of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, present the Kolmogorov axioms, and establish the first properties of the probability measure.

1.1 Random Experiments and Sample Spaces

Before any formalization, let us ask the naive question: what is a random experiment? It is an experiment whose outcome we cannot predict, but whose set of possible outcomes we can describe. A die roll, tomorrow's stock price, the lifetime of a transistor — all situations where uncertainty is irreducible, yet the structure of that uncertainty is accessible.

Definition 1.1 (Random experiment)**.** A **random experiment** is an experiment whose outcome cannot be predicted with certainty before it is performed, but for which the set of all possible outcomes can be described.

Definition 1.2 (Sample space)**.** The **sample space** is the set Ω of all possible outcomes of a random experiment. Each element $\omega \in \Omega$ is called a **sample point** or **outcome**.

Example 1.3.

1. **Roll of a die:** $\Omega = \{1, 2, 3, 4, 5, 6\}$.
2. **Coin toss:** $\Omega = \{\text{Heads}, \text{Tails}\}$.
3. **Lifetime of a component:** $\Omega = [0, +\infty)$.
4. **Infinite sequence of coin tosses:** $\Omega = \{0,1\}^{\mathbb{N}} = \{(\omega_1, \omega_2, \dots) : \omega_i \in \{0,1\}\}$.

Remark 1.4. The sample space Ω may be finite, countably infinite, or uncountable. The choice of Ω depends on the model; it is not unique for a given experiment.

1.2 σ-Algebras and Events

When Ω is finite or countable we can usually take every subset to be an event. For uncountable spaces such as $\mathbb{R}$, we must restrict to a well-behaved class of "measurable" sets.

Definition 1.5 (σ-algebra)**.** A **σ-algebra** (or σ-field) on Ω is a collection $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ satisfying:

(i) $\Omega \in \mathcal{F}$;

(ii) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ (closure under complementation);

(iii) if $(A_n)_{n\geq 1}$ is a sequence in $\mathcal{F}$, then $\displaystyle\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ (closure under countable unions).

Proposition 1.6 (Immediate properties)**.** Let $\mathcal{F}$ be a σ-algebra on Ω. Then:

(a) $\emptyset \in \mathcal{F}$;

(b) $\mathcal{F}$ is closed under countable intersections: if $A_n \in \mathcal{F}$ for all n, then $\bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$;

(c) $\mathcal{F}$ is closed under set differences: if $A, B \in \mathcal{F}$, then $A \setminus B \in \mathcal{F}$.

Proof. (a) $\emptyset = \Omega^c \in \mathcal{F}$ by (i) and (ii). (b) By De Morgan's laws, $\bigcap A_n = \big(\bigcup A_n^c\big)^c$, which belongs to $\mathcal{F}$ by (ii) and (iii). (c) $A \setminus B = A \cap B^c \in \mathcal{F}$ by (ii) and (b). □

Definition 1.7 (Event)**.** An **event** is an element of the σ-algebra $\mathcal{F}$, i.e. a subset of Ω to which a probability can be assigned. We say event A **occurs** if the outcome ω belongs to A.

Example 1.8 (Fundamental σ-algebras)**.** 1. **Trivial σ-algebra:** $\mathcal{F} = \{\emptyset, \Omega\}$. The smallest σ-algebra on Ω.

2. **Discrete σ-algebra:** $\mathcal{F} = \mathcal{P}(\Omega)$. The largest σ-algebra on Ω.

3. **Generated σ-algebra:** if $\mathcal{C}$ is a collection of subsets of Ω, then $\sigma(\mathcal{C})$ denotes the smallest σ-algebra containing $\mathcal{C}$.

Definition 1.9 (Borel σ-algebra)**.** The **Borel σ-algebra** on $\mathbb{R}$, denoted $\mathcal{B}(\mathbb{R})$, is the σ-algebra generated by the open subsets of $\mathbb{R}$:

$$\mathcal{B}(\mathbb{R}) = \sigma\big(\{O \subseteq \mathbb{R} : O \text{ is open}\}\big).$$

One can show that $\mathcal{B}(\mathbb{R}) = \sigma\big(\{(-\infty, a] : a \in \mathbb{R}\}\big) = \sigma\big(\{(a, b) : a < b\}\big)$.

1.3 The Kolmogorov Axioms

Definition 1.10 (Probability measure)**.** Let $(\Omega, \mathcal{F})$ be a measurable space. A **probability measure** is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying the **Kolmogorov axioms**:

(K1) **Non-negativity:** for every $A \in \mathcal{F}$, $\mathbb{P}(A) \geq 0$.

(K2) **Normalisation:** $\mathbb{P}(\Omega) = 1$.

(K3) **σ-additivity:** for every sequence $(A_n)_{n\geq 1}$ of *mutually disjoint* events (i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$):

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Definition 1.11 (Probability space)**.** The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**.

Finite vs. σ-additivity

Finite additivity ($\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ when $A \cap B = \emptyset$) follows from (K3). The σ-additivity axiom is strictly stronger: it guarantees consistency when passing to limits and is essential for the limit theorems of Chapters 9–11.

1.4 Properties of the Probability Measure

Theorem 1.12 (Fundamental properties)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For all $A, B \in \mathcal{F}$:*

(i) $\mathbb{P}(\emptyset) = 0$.

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

(iii) *If $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$ (monotonicity).*

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

(v) $\mathbb{P}(A) \leq 1$.

Proof. (i) Set $A_1 = \Omega$ and $A_n = \emptyset$ for $n \geq 2$. The sequence is disjoint and $\bigcup A_n = \Omega$. By (K3): $1 = \mathbb{P}(\Omega) = \mathbb{P}(\Omega) + \sum_{n\geq 2} \mathbb{P}(\emptyset)$, so $\mathbb{P}(\emptyset) = 0$.

(ii) $\Omega = A \cup A^c$ (disjoint), so $1 = \mathbb{P}(A) + \mathbb{P}(A^c)$.

(iii) $B = A \cup (B \setminus A)$ (disjoint), so $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$.

(iv) $A \cup B = A \cup (B \setminus A)$ and $B = (A \cap B) \cup (B \setminus A)$, giving $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$, and the result follows.

(v) From (ii): $\mathbb{P}(A) = 1 - \mathbb{P}(A^c) \leq 1$.

□

Theorem 1.13 (Inclusion–exclusion formula)**.** *For events $A_1, \dots, A_n \in \mathcal{F}$:*

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

Corollary 1.14 (Boole's inequality / sub-additivity)**.** *For any sequence $(A_n)_{n\geq 1}$ of events:*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

1.5 Continuity of the Probability Measure

Theorem 1.15 (Monotone continuity)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.*

(i) ***Continuity from below:*** *if $A_1 \subseteq A_2 \subseteq \cdots$ (increasing), then*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

(ii) ***Continuity from above:*** *if $A_1 \supseteq A_2 \supseteq \cdots$ (decreasing), then*

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

Proof. (i) Define $B_1 = A_1$ and $B_n = A_n \setminus A_{n-1}$ for $n \geq 2$. The B_n are mutually disjoint and $\bigcup_{n=1}^{N} B_n = A_N$. By σ-additivity:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) = \lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{P}(B_n) = \lim_{N\to\infty} \mathbb{P}(A_N).$$

(ii) Apply (i) to the increasing sequence $A_1^c \subseteq A_2^c \subseteq \cdots$ and use $\mathbb{P}(A_n^c) = 1 - \mathbb{P}(A_n)$. □

1.6 Constructing Probability Measures

1.6.1 Uniform probability on a finite set

Definition 1.16 (Uniform probability)**.** If Ω is a finite set, the **uniform probability** (or equally likely model) is defined by:

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\text{number of favourable outcomes}}{\text{total number of outcomes}}.$$

When is the uniform model appropriate?

The assumption of equally likely outcomes is only legitimate when the symmetry of the problem justifies it (fair die, random draw, etc.). Never invoke it without verification!

Example 1.17 (The birthday problem)**.** In a group of n people, what is the probability that at least two share the same birthday? Assume 365 equally likely days.

The sample space is $\Omega = \{1, \ldots, 365\}^n$ with $|\Omega| = 365^n$. The complementary event "all birthdays are distinct" has cardinality $365 \times 364 \times \cdots \times (365 - n + 1)$. Hence:

$$\mathbb{P}(\text{at least one match}) = 1 - \frac{365!}{(365-n)!\,365^n}.$$

For $n = 23$ we obtain $\mathbb{P} \approx 0.507$: more likely than not!

1.6.2 Probability defined by a discrete mass function

Proposition 1.18. If Ω is countable and $(p_\omega)_{\omega\in\Omega}$ is a family of reals with $p_\omega \geq 0$ for all ω and $\sum_{\omega\in\Omega} p_\omega = 1$, then

$$\mathbb{P}(A) = \sum_{\omega\in A} p_\omega, \quad A \subseteq \Omega,$$

defines a probability measure on $(\Omega, \mathcal{P}(\Omega))$.

1.6.3 Construction on $\mathbb{R}$: continuous densities

Proposition 1.19. Let $f : \mathbb{R} \to [0, +\infty)$ be an integrable function with $\int_{-\infty}^{+\infty} f(x)\,dx = 1$. Then

$$\mathbb{P}(A) = \int_A f(x)\,dx, \quad A \in \mathcal{B}(\mathbb{R}),$$

defines a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The function f is called a **probability density function** (pdf).

1.7 The First Borel–Cantelli Lemma

Theorem 1.20 (First Borel–Cantelli lemma)**.** *Let $(A_n)_{n\geq 1}$ be a sequence of events. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then*

$$\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 0,$$

where $\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$ is the event "infinitely many A_n occur".

Proof. Set $B_n = \bigcup_{k=n}^{\infty} A_k$. The sequence (B_n) is decreasing and $\limsup A_n = \bigcap B_n$. By continuity from above:

$$\mathbb{P}(\limsup A_n) = \lim_{n\to\infty} \mathbb{P}(B_n) \leq \lim_{n\to\infty} \sum_{k=n}^{\infty} \mathbb{P}(A_k) = 0,$$

since the tail of a convergent series tends to zero. □

Remark 1.21. The **second** Borel–Cantelli lemma (a partial converse requiring independence) will be stated in Chapter 2.

1.8 Exercises

Exercise 1.1. Show that the intersection of two σ-algebras is a σ-algebra. Is the union of two σ-algebras always a σ-algebra? Give a counterexample.

Exercise 1.2. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ (fair die). Write down the σ-algebra generated by the partition $\big\{\{1, 2\}, \{3, 4\}, \{5, 6\}\big\}$ and count its elements.

Exercise 1.3. A fair die is rolled twice. Find the probability that:

(a) the sum equals 7;

(b) the sum is at least 10;

(c) both faces are the same.

Exercise 1.4. Three cards are drawn successively without replacement from a standard deck of 52 cards. Find the probability of getting exactly two aces.

Exercise 1.5. Let (A_n) be a decreasing sequence of events with $\mathbb{P}(A_n) = 1/n$. What is $\mathbb{P}\big(\bigcap_{n=1}^{\infty} A_n\big)$? Justify your answer.

Exercise 1.6 (Borel–Cantelli)**.** Let (A_n) be a sequence of events with $\mathbb{P}(A_n) = 1/n^2$. Show that $\mathbb{P}(\limsup A_n) = 0$. Now consider $\mathbb{P}(A_n) = 1/n$: can you still conclude?

Exercise 1.7. Prove the inclusion–exclusion formula (Theorem 1.13) by induction on n.

Exercise 1.8 (Derangements)**.** Place n letters into n envelopes at random (one letter per envelope). Let D_n be the number of letters in the correct envelope. Using inclusion–exclusion, show that

$$\mathbb{P}(D_n = 0) = \sum_{k=0}^{n} \frac{(-1)^k}{k!} \xrightarrow[n\to\infty]{} e^{-1} \approx 0.368.$$

Chapter 2

Conditional Probability and Independence

Imagine you are a doctor. A patient tests positive for a rare disease. The test is 99% accurate. Should you be worried? Most people—including many physicians—would say yes without hesitation. And yet, if the disease affects only one in ten thousand people, the patient is far more likely to be healthy than sick. This unsettling result, which has fooled experts for centuries, is a direct consequence of conditional probability.

The idea itself traces back to Thomas Bayes, a Presbyterian minister whose posthumous essay, published in 1763 by his friend Richard Price, laid the groundwork for reasoning under uncertainty. But it was Pierre-Simon Laplace who, independently and with far greater mathematical ambition, turned conditional probability into a systematic tool. The story of this chapter is the story of how a simple question—“what changes when we learn something new?”—reshapes the entire landscape of probability theory.

Intuition

Conditional probability allows us to update our information when an event is observed. Independence formalises the idea that two events have no influence on each other. These two concepts lie at the heart of all probability theory.

2.1 Conditional Probability

Definition 2.1 (Conditional probability)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $B \in \mathcal{F}$ an event with $\mathbb{P}(B) > 0$. The **conditional probability** of A given B is

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

Proposition 2.2. For fixed B with $\mathbb{P}(B) > 0$, the map $A \mapsto \mathbb{P}(A \mid B)$ is a probability measure on $(\Omega, \mathcal{F})$.

Proof. We verify the Kolmogorov axioms:

- $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\mathbb{P}(B) \geq 0$ for all A.

- $\mathbb{P}(\Omega \mid B) = \mathbb{P}(B)/\mathbb{P}(B) = 1$.

- If (A_n) are mutually disjoint, then so are $(A_n \cap B)$, and by σ-additivity of $\mathbb{P}$:

$$\mathbb{P}\left(\bigcup_n A_n \,\middle|\, B\right) = \frac{\mathbb{P}(\bigcup_n (A_n \cap B))}{\mathbb{P}(B)} = \frac{\sum_n \mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} = \sum_n \mathbb{P}(A_n \mid B). \qquad \square$$

Example 2.3. Two fair dice are rolled. Given that the sum is 8, what is the probability that the first die shows 3?

Let $A = \{\text{first die} = 3\}$ and $B = \{\text{sum} = 8\}$. The outcomes giving sum 8 are $(2,6), (3,5), (4,4), (5,3), (6,2)$, so $\mathbb{P}(B) = 5/36$. Since $A \cap B = \{(3,5)\}$, $\mathbb{P}(A \cap B) = 1/36$. Hence $\mathbb{P}(A \mid B) = (1/36)/(5/36) = 1/5$.

2.2 Chain Rule and Bayes' Theorem

Theorem 2.4 (Multiplication rule / chain rule)**.** *For events $A_1, \dots, A_n$ with $\mathbb{P}(A_1 \cap \cdots \cap A_{n-1}) > 0$:*

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

Definition 2.5 (Partition of the sample space)**.** A family $(B_i)_{i \in I}$ of events forms a **partition** of Ω if:

(i) $B_i \cap B_j = \emptyset$ for $i \neq j$ (disjoint);

(ii) $\bigcup_{i \in I} B_i = \Omega$ (exhaustive);

(iii) $\mathbb{P}(B_i) > 0$ for all i.

Theorem 2.6 (Law of total probability)**.** *Let $(B_i)_{i \in I}$ be a partition of Ω (with I finite or countable). For any $A \in \mathcal{F}$:*

$$\mathbb{P}(A) = \sum_{i \in I} \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i).$$

Proof. Since $A = A \cap \Omega = \bigcup_i (A \cap B_i)$ (disjoint union):

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i) = \sum_i \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i). \qquad \square$$

Theorem 2.7 (Bayes' theorem)**.** *Let $(B_i)_{i \in I}$ be a partition of Ω and A an event with $\mathbb{P}(A) > 0$. Then for every $j \in I$:*

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(A \mid B_j)\,\mathbb{P}(B_j)}{\sum_{i \in I} \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i)}.$$

Bayesian interpretation

The $\mathbb{P}(B_i)$ are the *prior* probabilities of the hypotheses B_i. After observing A, Bayes' theorem yields the *posterior* probabilities $\mathbb{P}(B_i \mid A)$, updated via the likelihood $\mathbb{P}(A \mid B_i)$.

Example 2.8 (Medical screening)**.** A test has sensitivity 99% (true positive) and specificity 95% (true negative). The disease prevalence is 0.1%. What is the probability of disease given a positive test?

Let D = "diseased" and T^+ = "test positive". $\mathbb{P}(D) = 0.001$, $\mathbb{P}(T^+ \mid D) = 0.99$, $\mathbb{P}(T^+ \mid D^c) = 0.05$.

$$\begin{aligned}\mathbb{P}(D \mid T^+) &= \frac{\mathbb{P}(T^+ \mid D)\,\mathbb{P}(D)}{\mathbb{P}(T^+ \mid D)\,\mathbb{P}(D) + \mathbb{P}(T^+ \mid D^c)\,\mathbb{P}(D^c)} \\ &= \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.05 \times 0.999} = \frac{0.00099}{0.05094} \approx 0.0194.\end{aligned}$$

Even with an excellent test, the probability of disease given a positive result is only about 2%, because prevalence is low.

2.3 Independence of Events

Definition 2.9 (Independence of two events)**.** Two events A and B are **independent**, written $A \perp\!\!\!\perp B$, if

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

Remark 2.10. If $\mathbb{P}(B) > 0$, independence is equivalent to $\mathbb{P}(A \mid B) = \mathbb{P}(A)$: knowing that B occurred does not change the probability of A.

Independence $\neq$ disjointness

Two disjoint events ($A \cap B = \emptyset$) with positive probabilities are *never* independent, since $\mathbb{P}(A \cap B) = 0 \neq \mathbb{P}(A)\mathbb{P}(B)$. Do not confuse these notions!

Definition 2.11 (Mutual independence)**.** Events $A_1, \dots, A_n$ are **mutually independent** if, for every subset $J \subseteq \{1, \dots, n\}$ with $|J| \geq 2$:

$$\mathbb{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbb{P}(A_j).$$

Pairwise $\neq$ mutual independence

Pairwise independence does not imply mutual independence. All sub-families must be checked, not just pairs.

Example 2.12 (Bernstein's counterexample)**.** Roll two fair dice. Define:

- A = "first die is even",
- B = "second die is even",
- C = "the sum is even".

One verifies $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = 1/2$ and every pair is independent: $\mathbb{P}(A \cap B) = 1/4 = \mathbb{P}(A)\mathbb{P}(B)$, etc. However, $\mathbb{P}(A \cap B \cap C) = 1/4 \neq 1/8 = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$. The three events are pairwise but not mutually independent.

Proposition 2.13. If A and B are independent, then so are (A, B^c), (A^c, B), and (A^c, B^c).

Proof. $\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A)(1 - \mathbb{P}(B)) = \mathbb{P}(A)\mathbb{P}(B^c)$. The other cases are analogous. □

2.4 The Second Borel–Cantelli Lemma

Theorem 2.14 (Second Borel–Cantelli lemma)**.** *Let $(A_n)_{n\geq 1}$ be a sequence of* ***mutually independent*** *events. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = +\infty$, then*

$$\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 1.$$

That is, almost surely infinitely many of the A_n occur.

Proof. It suffices to show $\mathbb{P}\big(\bigcup_{k=n}^{\infty} A_k\big) = 1$ for all n. Taking complements:

$$\mathbb{P}\left(\bigcap_{k=n}^{N} A_k^c\right) = \prod_{k=n}^{N} \mathbb{P}(A_k^c) = \prod_{k=n}^{N}(1-\mathbb{P}(A_k)) \leq \prod_{k=n}^{N} e^{-\mathbb{P}(A_k)} = \exp\left(-\sum_{k=n}^{N} \mathbb{P}(A_k)\right).$$

As $N \to \infty$, the sum diverges, so the product tends to 0. By continuity from above: $\mathbb{P}\big(\bigcap_{k=n}^{\infty} A_k^c\big) = 0$, hence $\mathbb{P}\big(\bigcup_{k=n}^{\infty} A_k\big) = 1$. □

Summary: Borel–Cantelli

- $\sum \mathbb{P}(A_n) < \infty \implies \mathbb{P}(\limsup A_n) = 0$ (no hypothesis needed).
- $\sum \mathbb{P}(A_n) = \infty$ **and** independence $\implies \mathbb{P}(\limsup A_n) = 1$.

2.5 Independence of σ-Algebras and Random Variables

Definition 2.15 (Independent σ-algebras)**.** Two sub-σ-algebras $\mathcal{G}_1, \mathcal{G}_2 \subseteq \mathcal{F}$ are **independent** if, for all $A \in \mathcal{G}_1$ and $B \in \mathcal{G}_2$:

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

Definition 2.16 (Independent random variables — preview)**.** Two random variables X and Y are **independent** if the σ-algebras they generate, $\sigma(X)$ and $\sigma(Y)$, are independent. In terms of distributions (see Chapter 3 for full definitions), this amounts to:

$$\mathbb{P}(X \in A,\ Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B)$$

for all Borel sets $A, B \in \mathcal{B}(\mathbb{R})$.

2.6 Classical Applications

Example 2.17 (The Monty Hall problem)**.** A contestant chooses one of three doors. Behind one is a car; behind the other two are goats. The host, who knows where the car is, opens a door revealing a goat, then offers the contestant the chance to switch.

Let C_i = "the car is behind door i" ($i = 1, 2, 3$), and suppose the contestant picks door 1. By Bayes' theorem, the probability of winning by switching is $2/3$, while staying gives $1/3$. Switching is therefore advantageous.

Example 2.18 (Gambler's ruin)**.** A gambler starts with capital $k \in \{1, \dots, N-1\}$. At each round, they win \$1 with probability p or lose \$1 with probability $q = 1 - p$, independently. Play stops when the capital reaches 0 (ruin) or N (target).

Using conditional probability and a recurrence argument, the ruin probability is:

$$\mathbb{P}(\text{ruin} \mid X_0 = k) = \begin{cases} \dfrac{(q/p)^k - (q/p)^N}{1 - (q/p)^N} & \text{if } p \neq q, \\ 1 - \dfrac{k}{N} & \text{if } p = q = \frac{1}{2}. \end{cases}$$

2.7 Exercises

Exercise 2.1. A batch contains 10 items, 3 of which are defective. Two items are drawn successively without replacement. Find the probability that the second is defective given that the first is defective.

Exercise 2.2. Three machines M_1, M_2, M_3 produce 50%, 30%, and 20% of a factory's output respectively. Their defect rates are 2%, 3%, and 5%. A randomly chosen item is defective. What is the probability it came from M_2?

Exercise 2.3. Show that if A, B, C are mutually independent, then $A \cup B$ and C are independent.

Exercise 2.4. A fair coin is tossed n times. Let A_k be the event "the k-th toss is heads". Show that the A_k are mutually independent.

Exercise 2.5 (Iterated Bayes)**.** An urn contains 1 red and 1 blue ball. A ball is drawn at random, replaced, and a ball of the same colour is added. After n draws, find the probability that all drawn balls were red.

Exercise 2.6. Let $(A_n)_{n \geq 1}$ be independent events with $\mathbb{P}(A_n) = 1/(n+1)$. Show that $\mathbb{P}(\limsup A_n) = 1$. Interpret the result.

Exercise 2.7 (Simpson's paradox)**.** Give a numerical example where $\mathbb{P}(A \mid B) > \mathbb{P}(A \mid B^c)$ but $\mathbb{P}(A \mid B \cap C) < \mathbb{P}(A \mid B^c \cap C)$ and $\mathbb{P}(A \mid B \cap C^c) < \mathbb{P}(A \mid B^c \cap C^c)$. Explain the phenomenon.

Exercise 2.8. Let $p \in (0,1)$ and $(X_n)_{n \geq 1}$ be independent Bernoulli trials with parameter p. Show that $\mathbb{P}(\exists\, n : X_n = 1) = 1$.

Chapter 3

Discrete Random Variables

Rolling a die, counting customers in a queue, observing radioactive decays per second: in each of these situations, chance produces an outcome that can be enumerated. Discrete random variables—those taking only finitely or countably many values—are the first we encounter, and the most intuitive. Yet they hold surprises: the Poisson distribution, discovered by Siméon Denis Poisson in 1837 to model judicial errors, reappears everywhere, from call centres to V-2 bomb impacts on London. The geometric distribution captures the patience of the stubborn gambler, while the binomial distribution, direct heir to Jacob Bernoulli's work, remains the reference model for repeated experiments.

This chapter builds the formal framework for discrete random variables and develops the classical distributions that populate all of probability theory.

Intuition

A random variable is a function that assigns a numerical value to the outcome of a random experiment. This chapter treats the discrete case: the set of values taken is finite or countable.

3.1 Definition and Distribution

Definition 3.1 (Random variable)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A **real-valued random variable** is a measurable function $X : \Omega \to \mathbb{R}$, i.e. for every $B \in \mathcal{B}(\mathbb{R})$:

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}.$$

Definition 3.2 (Discrete random variable)**.** A random variable X is **discrete** if it takes values in a finite or countable set $\{x_1, x_2, \ldots\} \subseteq \mathbb{R}$.

Definition 3.3 (Probability mass function)**.** The **distribution** of X is the probability measure P_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by $P_X(B) = \mathbb{P}(X \in B)$. For a discrete variable, the distribution is completely determined by the **probability mass function** (pmf):

$$p_X(x_k) = \mathbb{P}(X = x_k), \quad k = 1, 2, \ldots$$

with $p_X(x_k) \geq 0$ and $\sum_k p_X(x_k) = 1$.

Definition 3.4 (Cumulative distribution function)**.** The **cumulative distribution function** (cdf) of X is:

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_k \leq x} p_X(x_k), \quad x \in \mathbb{R}.$$

Proposition 3.5 (Properties of F_X)**.** (i) F_X is non-decreasing.

(ii) $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to+\infty} F_X(x) = 1$.

(iii) F_X is right-continuous.

(iv) $\mathbb{P}(X = a) = F_X(a) - F_X(a^-)$ (the jump at a).

(v) For a discrete variable, F_X is a step function.

3.2 Classical Discrete Distributions

3.2.1 Bernoulli distribution

Definition 3.6 (Bernoulli)**.** $X \sim \text{Bernoulli}(p)$ with $p \in [0, 1]$:

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p = q.$$

$\mathbb{E}[X] = p$, $\text{Var}(X) = pq$.

3.2.2 Binomial distribution

Definition 3.7 (Binomial)**.** $X \sim \text{Bin}(n, p)$: the number of successes in n independent Bernoulli trials.

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

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

Example 3.8. A fair coin is tossed 10 times. The probability of exactly 6 heads is:

$$\mathbb{P}(X = 6) = \binom{10}{6} \left(\frac{1}{2}\right)^{10} = \frac{210}{1024} \approx 0.205.$$

3.2.3 Geometric distribution

Definition 3.9 (Geometric)**.** $X \sim \text{Geom}(p)$: the number of trials until the first success.

$$\mathbb{P}(X = k) = (1 - p)^{k-1} p, \quad k = 1, 2, 3, \ldots$$

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

Proposition 3.10 (Memorylessness)**.** The geometric distribution is the only discrete distribution on $\mathbb{N}^*$ satisfying the **memoryless property**:

$$\mathbb{P}(X > m + n \mid X > m) = \mathbb{P}(X > n) \quad \forall\, m, n \geq 0.$$

Proof. $\mathbb{P}(X > k) = (1 - p)^k$. Hence:

$$\mathbb{P}(X > m + n \mid X > m) = \frac{(1 - p)^{m+n}}{(1 - p)^m} = (1 - p)^n = \mathbb{P}(X > n). \qquad \square$$

3.2.4 Poisson distribution

Definition 3.11 (Poisson)**.** $X \sim \text{Poisson}(\lambda)$ with $\lambda > 0$:

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

$\mathbb{E}[X] = \lambda$, $\text{Var}(X) = \lambda$.

Theorem 3.12 (Poisson approximation)**.** *If $X_n \sim \text{Bin}(n, p_n)$ with $np_n \to \lambda$ as $n \to \infty$, then for every $k \in \mathbb{N}$:*

$$\mathbb{P}(X_n = k) \xrightarrow[n\to\infty]{} e^{-\lambda}\frac{\lambda^k}{k!}.$$

Proof.

$$\begin{aligned}\mathbb{P}(X_n = k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\ &= \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \cdot \underbrace{\frac{n(n-1)\cdots(n-k+1)}{n^k}}_{\to 1} \cdot \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \cdot \underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\to 1} \xrightarrow[n\to\infty]{} e^{-\lambda}\frac{\lambda^k}{k!}.\end{aligned}$$

□

When to use Poisson?

The Poisson distribution models the number of rare events in a given interval (phone calls, radioactive particles, failures, etc.). Rule of thumb: if $n \geq 30$ and $p \leq 0.1$, approximate $\text{Bin}(n, p)$ by $\text{Poisson}(np)$.

3.2.5 Hypergeometric distribution

Definition 3.13 (Hypergeometric)**.** Draw n items without replacement from a population of N items, K of which are marked. X = number of marked items among the n drawn:

$$\mathbb{P}(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \quad k = \max(0, n + K - N), \dots, \min(n, K).$$

$\mathbb{E}[X] = nK/N$.

3.3 Summary Table

Fundamental discrete distributions

Distribution	**Parameters**	$\mathbb{E}[X]$	$\mathrm{Var}(X)$
Bernoulli(p)	$p \in [0,1]$	p	$p(1-p)$
Binomial(n,p)	$n \in \mathbb{N}^*,\ p \in [0,1]$	np	$np(1-p)$
Geometric(p)	$p \in (0,1]$	$1/p$	$(1-p)/p^2$
Poisson(λ)	$\lambda > 0$	λ	λ
Uniform$\{1,\dots,n\}$	$n \in \mathbb{N}^*$	$(n+1)/2$	$(n^2-1)/12$

3.4 Probability Generating Functions

Definition 3.14 (Probability generating function)**.** Let X be a random variable taking values in $\mathbb{N}$. The **probability generating function** (pgf) of X is:

$$G_X(s) = \mathbb{E}[s^X] = \sum_{k=0}^{\infty} \mathbb{P}(X=k)\, s^k, \quad |s| \le 1.$$

Proposition 3.15 (Properties of G_X)**.** (i) $G_X(1) = 1$.

(ii) $\mathbb{E}[X] = G'_X(1)$.

(iii) $\mathbb{E}[X(X-1)] = G''_X(1)$, whence $\mathrm{Var}(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$.

(iv) $\mathbb{P}(X=k) = G_X^{(k)}(0)/k!$.

(v) If X and Y are independent, $G_{X+Y}(s) = G_X(s) \cdot G_Y(s)$.

Example 3.16 (Pgf of the Poisson distribution)**.** If $X \sim \mathrm{Poisson}(\lambda)$:

$$G_X(s) = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} s^k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda s)^k}{k!} = e^{-\lambda} \cdot e^{\lambda s} = e^{\lambda(s-1)}.$$

We recover $G'_X(1) = \lambda = \mathbb{E}[X]$ and $G''_X(1) = \lambda^2$, giving $\mathrm{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$.

3.5 Exercises

Exercise 3.1. Let $X \sim \mathrm{Bin}(n,p)$. Show directly (by calculation) that $\mathbb{E}[X] = np$ and $\mathrm{Var}(X) = np(1-p)$.

Exercise 3.2. Let $X \sim \mathrm{Geom}(p)$. Compute $\mathbb{E}[X]$ using the probability generating function.

Exercise 3.3. Let $X \sim \mathrm{Poisson}(\lambda)$ and $Y \sim \mathrm{Poisson}(\mu)$, independent. Show that $X + Y \sim \mathrm{Poisson}(\lambda + \mu)$ using probability generating functions.

Exercise 3.4. A telephone exchange receives an average of 4 calls per minute, modelled by a Poisson distribution. Find the probability of:

(a) exactly 3 calls in one minute;

(b) at least 1 call in 30 seconds;

(c) more than 10 calls in 2 minutes.

Exercise 3.5. A bag contains 5 red and 15 blue balls. Four balls are drawn without replacement. Find the probability of getting exactly 2 red balls.

Exercise 3.6. A fair die is rolled until a 6 appears. Let X be the number of rolls.

(a) What is the distribution of X?

(b) Compute $\mathbb{P}(X > 10)$.

(c) Compute $\mathbb{P}(X > 15 \mid X > 5)$. Verify the memoryless property.

Exercise 3.7. Prove that the geometric distribution is the only discrete distribution on $\mathbb{N}^*$ with the memoryless property. *Hint:* set $g(n) = \mathbb{P}(X > n)$ and use the functional equation $g(m + n) = g(m)g(n)$.

Exercise 3.8. Let $N \sim \text{Poisson}(\lambda)$ and $(X_i)_{i\geq 1}$ be i.i.d. Bernoulli(p), independent of N. Set $S = \sum_{i=1}^{N} X_i$. Show that $S \sim \text{Poisson}(\lambda p)$.

Chapter 4

Continuous Random Variables

With discrete variables, we counted. With continuous variables, we measure. The transition is more than a change of vocabulary: it requires replacing sums with integrals, point masses with densities, and coming to terms with the idea that in the continuous case, the probability of any *exact* value is always zero. It was Abraham de Moivre who, around 1733, opened this path by approximating the binomial distribution with the bell curve—what we now call the normal distribution, and which Gauss would immortalise in his astronomical work.

Intuition

A continuous random variable takes values in an interval (or all of $\mathbb{R}$) and its distribution is described by a **probability density function**. This chapter develops the tools specific to the continuous case: densities, cumulative distribution functions, and quantiles.

4.1 Probability Density Functions

Definition 4.1 (Absolutely continuous random variable)**.** A random variable X is **continuous** (or absolutely continuous) if there exists an integrable function $f_X : \mathbb{R} \to [0, +\infty)$ such that for all $a \leq b$:

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

The function f_X is called the **probability density function** (pdf) of X.

Proposition 4.2 (Properties of the pdf)**.** (i) $f_X(x) \geq 0$ for all $x \in \mathbb{R}$.

(ii) $\int_{-\infty}^{+\infty} f_X(x)\,dx = 1$.

(iii) $\mathbb{P}(X = a) = 0$ for every $a \in \mathbb{R}$.

(iv) $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b)$.

A density is not a probability

It is possible for $f_X(x) > 1$ at some values of x. The density is a “concentration of probability”, not a probability itself. Only its integral over an interval gives a probability.

CHAPTER 4. CONTINUOUS RANDOM VARIABLES

4.2 Cumulative Distribution Function

Definition 4.3 (CDF).

$$F_X(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f_X(t)\,dt, \quad x \in \mathbb{R}.$$

Theorem 4.4 (Density–CDF relationship). *If F_X is differentiable at x, then $f_X(x) = F_X'(x)$. More generally, f_X equals the derivative of F_X almost everywhere.*

Proposition 4.5 (Properties of F_X). (i) F_X is non-decreasing and continuous (in the absolutely continuous case).

(ii) $\lim_{x\to-\infty} F_X(x) = 0$, $\lim_{x\to+\infty} F_X(x) = 1$.

(iii) $\mathbb{P}(a < X \le b) = F_X(b) - F_X(a)$.

4.3 Classical Continuous Distributions

4.3.1 Continuous uniform distribution

Definition 4.6 (Uniform on $[a,b]$). $X \sim \mathcal{U}([a,b])$ with $a < b$:

$$f_X(x) = \frac{1}{b-a}\,\mathbf{1}_{[a,b]}(x), \quad F_X(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} & \text{if } a \le x \le b, \\ 1 & \text{if } x > b. \end{cases}$$

$\mathbb{E}[X] = (a+b)/2$, $\mathrm{Var}(X) = (b-a)^2/12$.

4.3.2 Exponential distribution

Definition 4.7 (Exponential). $X \sim \mathrm{Exp}(\lambda)$ with $\lambda > 0$:

$$f_X(x) = \lambda e^{-\lambda x}\,\mathbf{1}_{[0,+\infty)}(x), \quad F_X(x) = (1 - e^{-\lambda x})\,\mathbf{1}_{[0,+\infty)}(x).$$

$\mathbb{E}[X] = 1/\lambda$, $\mathrm{Var}(X) = 1/\lambda^2$.

Theorem 4.8 (Memorylessness). *The exponential distribution is the only continuous distribution on $[0,+\infty)$ satisfying:*

$$\mathbb{P}(X > s+t \mid X > s) = \mathbb{P}(X > t) \quad \forall\, s,t \ge 0.$$

Proof. Let $g(t) = \mathbb{P}(X > t) = e^{-\lambda t}$. Then:

$$\mathbb{P}(X > s+t \mid X > s) = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = g(t).$$

Conversely, if $g(s+t) = g(s)g(t)$ with g continuous, decreasing, and $g(0) = 1$, then $g(t) = e^{-\lambda t}$ for some $\lambda > 0$, which characterises the exponential distribution. □

Exponential and Poisson

If arrivals follow a Poisson process of rate λ, then the time between consecutive arrivals follows an $\mathrm{Exp}(\lambda)$ distribution.

4.3.3 Normal (Gaussian) distribution

Definition 4.9 (Normal distribution)**.** $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma > 0$:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

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

Definition 4.10 (Standard normal)**.** If $Z \sim \mathcal{N}(0,1)$, we write $\Phi(x) = \mathbb{P}(Z \le x)$ for its CDF. The standardisation relationship is:

$$\text{if } X \sim \mathcal{N}(\mu, \sigma^2), \quad \mathbb{P}(X \le x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Key values of Φ

z	1.645	1.96	2.326	2.576	3.291
$\Phi(z)$	0.95	0.975	0.99	0.995	0.9995

By symmetry: $\Phi(-z) = 1 - \Phi(z)$.

Proposition 4.11 (The $k\sigma$ rule)**.** If $X \sim \mathcal{N}(\mu, \sigma^2)$:

- $\mathbb{P}(|X-\mu| \le \sigma) \approx 0.6827 \quad (1\sigma \text{ rule})$,
- $\mathbb{P}(|X-\mu| \le 2\sigma) \approx 0.9545 \quad (2\sigma \text{ rule})$,
- $\mathbb{P}(|X-\mu| \le 3\sigma) \approx 0.9973 \quad (3\sigma \text{ rule})$.

4.3.4 Gamma distribution

Definition 4.12 (Gamma)**.** $X \sim \Gamma(\alpha, \lambda)$ with $\alpha, \lambda > 0$:

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \mathbf{1}_{(0,+\infty)}(x),$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$. We have $\mathbb{E}[X] = \alpha/\lambda$ and $\mathrm{Var}(X) = \alpha/\lambda^2$.

Remark 4.13. $\Gamma(1,\lambda) = \mathrm{Exp}(\lambda)$ and $\Gamma(n/2, 1/2) = \chi^2(n)$ (the chi-squared distribution with n degrees of freedom).

4.3.5 Beta distribution

Definition 4.14 (Beta)**.** $X \sim \mathrm{Beta}(\alpha, \beta)$ with $\alpha, \beta > 0$:

$$f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \mathbf{1}_{(0,1)}(x),$$

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. $\mathbb{E}[X] = \alpha/(\alpha+\beta)$.

4.4 Quantiles and Median

Definition 4.15 (Quantile)**.** The **quantile of order** p $(0 < p < 1)$ of X is:

$$q_p = \inf\{x \in \mathbb{R} : F_X(x) \geq p\} = F_X^{-1}(p).$$

Definition 4.16 (Median)**.** The **median** of X is the quantile of order $1/2$: $\mathrm{Med}(X) = q_{1/2}$.

Example 4.17. If $X \sim \mathrm{Exp}(\lambda)$, then $F_X(x) = 1 - e^{-\lambda x}$ for $x \geq 0$. The quantile of order p is:

$$q_p = -\frac{\ln(1-p)}{\lambda}.$$

In particular, $\mathrm{Med}(X) = \ln 2/\lambda$.

4.5 Simulation by Inversion

Theorem 4.18 (Inverse transform method)**.** *If $U \sim \mathcal{U}([0,1])$ and F is a continuous, strictly increasing CDF, then $X = F^{-1}(U)$ has CDF F.*

Proof. $\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$. □

Example 4.19. To simulate $X \sim \mathrm{Exp}(\lambda)$, generate $U \sim \mathcal{U}([0,1])$ and set $X = -\ln(1-U)/\lambda = -\ln(U)/\lambda$ (since $1-U$ has the same distribution as U).

4.6 Exercises

Exercise 4.1. Let X have density $f(x) = cx^2\,\mathbf{1}_{[0,1]}(x)$. Find c, $\mathbb{E}[X]$, $\mathrm{Var}(X)$, and F_X.

Exercise 4.2. Let $X \sim \mathcal{N}(100, 25)$. Compute:

(a) $\mathbb{P}(X > 105)$;

(b) $\mathbb{P}(90 < X < 110)$;

(c) the quantile $q_{0.9}$.

Exercise 4.3. Show that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = (X-\mu)/\sigma \sim \mathcal{N}(0,1)$.

Exercise 4.4. Let $X \sim \mathrm{Exp}(\lambda)$. Find the density and expectation of $Y = X^2$.

Exercise 4.5. Show that the Gaussian integral equals $\int_{-\infty}^{+\infty} e^{-x^2/2}\,dx = \sqrt{2\pi}$. *Hint:* compute I^2 using polar coordinates.

Exercise 4.6. Let $X \sim \Gamma(\alpha, \lambda)$. Show that $Y = 2\lambda X \sim \chi^2(2\alpha)$ when $\alpha \in \mathbb{N}^*/2$.

Exercise 4.7 (Simulation)**.** Describe a method to simulate a random variable with density $f(x) = 2x\,\mathbf{1}_{[0,1]}(x)$ using the inverse transform method.

Exercise 4.8. Let X have density $f(x) = \frac{2}{\pi(1+x^2)}\,\mathbf{1}_{(0,+\infty)}(x)$. Verify that f is a valid density and compute F_X. Does $\mathbb{E}[X]$ exist?

Chapter 5

Expectation, Variance, Moments

If you play a game of chance a large number of times, your average gain per game will converge to a precise number: the *expectation*. This result, anticipated by Pascal and Fermat as early as 1654, is the law of large numbers. Expectation summarises the "location" of a random variable; *variance*, introduced by Ronald Fisher in 1918, measures its "spread." Together, these two parameters capture the essence of a distribution's behaviour, and higher-order moments—skewness, kurtosis—refine the portrait.

Intuition

Expectation and variance are the two fundamental summary parameters of a random variable: location and spread. This chapter defines them rigorously, establishes their properties, and introduces higher-order moments.

5.1 Expectation

Definition 5.1 (Expectation — discrete case)**.** Let X be a discrete random variable taking values $x_1, x_2, \ldots$ The **expectation** (or **mean**) of X is:

$$\mathbb{E}[X] = \sum_k x_k \, \mathbb{P}(X = x_k),$$

provided $\sum_k |x_k| \, \mathbb{P}(X = x_k) < \infty$ (absolute convergence).

Definition 5.2 (Expectation — continuous case)**.** If X has density f_X, the **expectation** is:

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

provided $\int |x| \, f_X(x) \, dx < \infty$.

Theorem 5.3 (Law of the unconscious statistician (LOTUS))**.** *If $g : \mathbb{R} \to \mathbb{R}$ is measurable:*

- ***Discrete:*** $\mathbb{E}[g(X)] = \sum_k g(x_k) \, \mathbb{P}(X = x_k)$.
- ***Continuous:*** $\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) \, f_X(x) \, dx$.

(Provided the sums/integrals converge absolutely.)

CHAPTER 5. EXPECTATION, VARIANCE, MOMENTS

Why LOTUS is so useful

The transfer formula lets us compute $\mathbb{E}[g(X)]$ directly from the distribution of X, *without* having to find the distribution of $Y = g(X)$.

Theorem 5.4 (Properties of expectation). *(i)* ***Linearity:*** $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ *for all* $a, b \in \mathbb{R}$.

(ii) ***Positivity:*** *if* $X \geq 0$ *a.s., then* $\mathbb{E}[X] \geq 0$.

(iii) ***Monotonicity:*** *if* $X \leq Y$ *a.s., then* $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

(iv) $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$.

(v) ***Product for independent variables:*** *if* $X \perp\!\!\!\perp Y$ *and* $\mathbb{E}[|X|], \mathbb{E}[|Y|] < \infty$*, then* $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.

Example 5.5. Let $X \sim \text{Exp}(\lambda)$. Then:

$$\mathbb{E}[X] = \int_0^\infty x\,\lambda e^{-\lambda x}\,dx = \left[-x\,e^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x}\,dx = 0 + \frac{1}{\lambda} = \frac{1}{\lambda}.$$

5.2 Variance

Definition 5.6 (Variance and standard deviation). The **variance** of X (when $\mathbb{E}[X^2] < \infty$) is:

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

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

Theorem 5.7 (König–Huygens formula).

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

Proof.

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

Proposition 5.8 (Properties of variance). (i) $\text{Var}(X) \geq 0$, with equality iff X is constant a.s.

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

(iii) If $X \perp\!\!\!\perp Y$, then $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.

(iv) In general: $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$.

Summary: expectation and variance

$$\mathbb{E}[aX+b] = a\mathbb{E}[X]+b, \qquad \text{Var}(aX+b) = a^2\text{Var}(X), \qquad \text{Var}(X) = \mathbb{E}[X^2]-\mathbb{E}[X]^2.$$

5.3 Higher-Order Moments

Definition 5.9 (Moment of order r)**.** The r**-th moment** of X ($r \in \mathbb{N}^*$) is:

$$m_r = \mathbb{E}[X^r],$$

when this expectation exists. The r**-th central moment** is $\mu_r = \mathbb{E}[(X - \mathbb{E}[X])^r]$.

Definition 5.10 (Skewness and kurtosis)**.**

- The **skewness** is $\gamma_1 = \mu_3/\sigma^3$.
- The **excess kurtosis** is $\gamma_2 = \mu_4/\sigma^4 - 3$.
  For the normal distribution, $\gamma_1 = 0$ and $\gamma_2 = 0$.

Remark 5.11. $\gamma_1 > 0$ indicates a heavier right tail; $\gamma_1 < 0$ a heavier left tail. $\gamma_2 > 0$ indicates tails heavier than the normal (leptokurtic distribution).

5.4 Fundamental Inequalities

Theorem 5.12 (Markov's inequality)**.** *If* $X \geq 0$ *and* $a > 0$*:*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof. $\mathbb{E}[X] = \mathbb{E}[X\,\mathbf{1}_{X\geq a}] + \mathbb{E}[X\,\mathbf{1}_{X<a}] \geq \mathbb{E}[X\,\mathbf{1}_{X\geq a}] \geq a\,\mathbb{P}(X \geq a)$. □

Theorem 5.13 (Chebyshev's inequality)**.** *If* $\operatorname{Var}(X)$ *exists:*

$$\mathbb{P}\big(|X - \mathbb{E}[X]| \geq \varepsilon\big) \leq \frac{\operatorname{Var}(X)}{\varepsilon^2}, \quad \forall\, \varepsilon > 0.$$

Proof. Apply Markov's inequality to $Y = (X - \mathbb{E}[X])^2$ with $a = \varepsilon^2$:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) = \mathbb{P}(Y \geq \varepsilon^2) \leq \frac{\mathbb{E}[Y]}{\varepsilon^2} = \frac{\operatorname{Var}(X)}{\varepsilon^2}.$$ □

Theorem 5.14 (Jensen's inequality)**.** *If* φ *is convex and* $\mathbb{E}[|X|] < \infty$*:*

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)].$$

If φ *is concave, the inequality is reversed.*

Example 5.15. Taking $\varphi(x) = x^2$ (convex), we recover $(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$, i.e. $\operatorname{Var}(X) \geq 0$.

Theorem 5.16 (Cauchy–Schwarz inequality)**.**

$$|\mathbb{E}[XY]|^2 \leq \mathbb{E}[X^2]\,\mathbb{E}[Y^2].$$

Equality holds if and only if $Y = aX + b$ *a.s. for some* $a, b \in \mathbb{R}$*.*

5.5 Conditional Expectation (Introduction)

Definition 5.17 (Discrete conditional expectation)**.** If X and Y are discrete:

$$\mathbb{E}[X \mid Y = y] = \sum_x x \, \mathbb{P}(X = x \mid Y = y).$$

Theorem 5.18 (Law of total expectation)**.**

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \sum_y \mathbb{E}[X \mid Y = y] \, \mathbb{P}(Y = y).$$

Theorem 5.19 (Law of total variance)**.**

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]).$$

Example 5.20. Let $N \sim \mathrm{Poisson}(\lambda)$ and, conditionally on $N = n$, $S = X_1 + \cdots + X_n$ where the X_i are i.i.d. with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Then:

$$\begin{aligned} \mathbb{E}[S] &= \mathbb{E}[\mathbb{E}[S \mid N]] = \mathbb{E}[N\mu] = \lambda\mu, \\ \mathrm{Var}(S) &= \mathbb{E}[\mathrm{Var}(S \mid N)] + \mathrm{Var}(\mathbb{E}[S \mid N]) = \mathbb{E}[N\sigma^2] + \mathrm{Var}(N\mu) = \lambda\sigma^2 + \lambda\mu^2. \end{aligned}$$

5.6 Exercises

Exercise 5.1. Compute $\mathbb{E}[X]$ and $\mathrm{Var}(X)$ for $X \sim \mathcal{N}(\mu, \sigma^2)$ directly by integration.

Exercise 5.2. Let X have density $f(x) = 6x(1-x)\,\mathbf{1}_{[0,1]}(x)$. Compute $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, $\mathrm{Var}(X)$, γ_1, and γ_2.

Exercise 5.3. Prove Markov's inequality. Deduce that if $\mathbb{E}[e^{tX}]$ exists for $t > 0$, then $\mathbb{P}(X \geq a) \leq e^{-ta}\,\mathbb{E}[e^{tX}]$ for all a (Chernoff bound).

Exercise 5.4. Let X be uniformly distributed on $\{1, 2, \ldots, n\}$. Compute $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, and $\mathrm{Var}(X)$.

Exercise 5.5. Prove the Cauchy–Schwarz inequality by considering the polynomial $p(t) = \mathbb{E}[(X + tY)^2] \geq 0$ and studying its discriminant.

Exercise 5.6. A fair die is rolled N times, where $N \sim \mathrm{Poisson}(10)$. Let S be the sum of the results. Compute $\mathbb{E}[S]$ and $\mathrm{Var}(S)$.

Exercise 5.7. Let X be a non-negative random variable. Show that

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\, dt.$$

Hint: interchange the integral and the expectation.

Exercise 5.8. Show that $\mathrm{Var}(X) = \min_{c \in \mathbb{R}} \mathbb{E}[(X - c)^2]$, with the minimum attained at $c = \mathbb{E}[X]$.

Chapter 6

Standard Distributions and Properties

Probability theory has produced, over the centuries, a gallery of remarkable distributions—each born from a concrete problem, and each endowed with specific properties that make it irreplaceable in its domain. The Gaussian normal, the exponential of waiting times, the Poisson of rare events, the Beta of Bayesian statistics, the Gamma of queuing theory: this chapter gathers them in an organised panorama, highlighting their mutual relationships (sums, conditioning, limits) and the properties that characterise them.

Intuition

This chapter presents an organised catalogue of the most common probability distributions, both discrete and continuous, together with their properties, mutual relationships, and domains of application.

6.1 Discrete Distributions

6.1.1 Negative binomial distribution

Definition 6.1 (Negative binomial)**.** $X \sim \mathrm{NB}(r, p)$: number of trials to obtain the r-th success.

$$\mathbb{P}(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k = r, r+1, \ldots$$

$\mathbb{E}[X] = r/p$, $\mathrm{Var}(X) = r(1-p)/p^2$. For $r = 1$ this reduces to the geometric distribution.

6.1.2 Discrete uniform distribution

Definition 6.2 (Discrete uniform)**.** $X \sim \mathcal{U}\{a, a+1, \ldots, b\}$:

$$\mathbb{P}(X = k) = \frac{1}{b-a+1}, \quad k = a, a+1, \ldots, b.$$

$\mathbb{E}[X] = (a+b)/2$, $\mathrm{Var}(X) = ((b-a+1)^2 - 1)/12$.

6.1.3 Poisson distribution — further properties

Proposition 6.3 (Stability under summation)**.** If $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ are independent, then $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Proposition 6.4 (Poisson splitting)**.** If $X \sim \text{Poisson}(\lambda)$ and each event is independently classified as type 1 with probability p or type 2 with probability $1 - p$, then the counts N_1, N_2 are independent with $N_1 \sim \text{Poisson}(\lambda p)$ and $N_2 \sim \text{Poisson}(\lambda(1 - p))$.

6.2 Continuous Distributions

6.2.1 Log-normal distribution

Definition 6.5 (Log-normal)**.** $X \sim \text{LogN}(\mu, \sigma^2)$ if $\ln X \sim \mathcal{N}(\mu, \sigma^2)$.

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0.$$

$\mathbb{E}[X] = e^{\mu+\sigma^2/2}$, $\text{Var}(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$.

6.2.2 Cauchy distribution

Definition 6.6 (Cauchy)**.** $X \sim \text{Cauchy}(x_0, \gamma)$:

$$f_X(x) = \frac{1}{\pi\gamma\left(1 + \left(\frac{x-x_0}{\gamma}\right)^2\right)}.$$

The expectation and variance do not exist.

Cauchy has no moments

The Cauchy distribution shows that not every random variable has an expectation. The sample mean of i.i.d. Cauchy variables does not converge: the law of large numbers does not apply.

6.2.3 Weibull distribution

Definition 6.7 (Weibull)**.** $X \sim \text{Weibull}(k, \lambda)$ with $k, \lambda > 0$:

$$f_X(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}\,\mathbf{1}_{(0,+\infty)}(x).$$

For $k = 1$ this is the exponential distribution with parameter $1/\lambda$.

6.2.4 Chi-squared distribution

Definition 6.8 (Chi-squared)**.** If $Z_1, \ldots, Z_n$ are i.i.d. $\mathcal{N}(0, 1)$, then $\chi^2_n = Z_1^2 + \cdots + Z_n^2 \sim \Gamma(n/2, 1/2)$ is the **chi-squared distribution with n degrees of freedom**. $\mathbb{E}[\chi^2_n] = n$, $\text{Var}(\chi^2_n) = 2n$.

6.2.5 Student's t-distribution

Definition 6.9 (Student's t)**.** If $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_n$ are independent, then $T = Z/\sqrt{V/n}$ follows the **Student t-distribution with n degrees of freedom**. $\mathbb{E}[T] = 0$ (for $n > 1$), $\mathrm{Var}(T) = n/(n-2)$ (for $n > 2$). For $n = 1$, one recovers the standard Cauchy distribution.

6.2.6 Fisher's F-distribution

Definition 6.10 (Fisher's F)**.** If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ are independent, then $F = (U/m)/(V/n)$ follows the **F-distribution with (m,n) degrees of freedom**.

6.3 Relationships Between Distributions

Map of distribution relationships

- $\mathrm{Bernoulli}(p) = \mathrm{Bin}(1,p)$.
- $\mathrm{Geom}(p) = \mathrm{NB}(1,p)$.
- $\mathrm{Exp}(\lambda) = \Gamma(1,\lambda)$.
- $\chi^2(n) = \Gamma(n/2, 1/2)$.
- $\mathrm{Bin}(n,p) \xrightarrow{n\to\infty,\, np\to\lambda} \mathrm{Poisson}(\lambda)$.
- $\mathrm{Bin}(n,p) \xrightarrow{n\to\infty} \mathcal{N}(np, np(1-p)) \quad \text{(CLT)}$.
- $\mathrm{Poisson}(\lambda) \xrightarrow{\lambda\to\infty} \mathcal{N}(\lambda,\lambda)$.
- $t_n \xrightarrow{n\to\infty} \mathcal{N}(0,1)$.
- $\Gamma(n,\lambda)$ = distribution of the sum of n i.i.d. $\mathrm{Exp}(\lambda)$ variables.

6.4 General Summary Table

Continuous distributions summary

Distribution	**Density**	**Support**	$\mathbb{E}[X]$	$\mathrm{Var}(X)$
$\mathcal{U}([a,b])$	$\frac{1}{b-a}$	$[a,b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathrm{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	$[0,\infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$	$\mathbb{R}$	μ	σ^2
$\Gamma(\alpha,\lambda)$	$\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$	$(0,\infty)$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
$\mathrm{Beta}(\alpha,\beta)$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	$(0,1)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

6.5 Identification Methods

Remark 6.11. To identify a distribution, proceed as follows:

1. Identify the **support**: $\mathbb{N}$, $\{0, 1, \ldots, n\}$, $(0, \infty)$, $\mathbb{R}$, $(0, 1)$, etc.
2. Recognise the **functional form**: decaying exponential, symmetric bell curve, power law, etc.
3. Check the **parameters** by computing the mean and variance.

6.6 Exercises

Exercise 6.1. Let $X \sim \Gamma(n, \lambda)$ with $n \in \mathbb{N}^*$. Show that X has the same distribution as the sum of n i.i.d. $\text{Exp}(\lambda)$ variables.

Exercise 6.2. Show that if $X \sim \mathcal{N}(0, 1)$, then $X^2 \sim \chi^2(1)$.

Exercise 6.3. Let $X \sim \text{Poisson}(\lambda)$ with $\lambda = 100$. Using the normal approximation, estimate $\mathbb{P}(90 \leq X \leq 110)$.

Exercise 6.4. Let $U \sim \mathcal{U}([0, 1])$. Find the distribution of $X = -\ln U$.

Exercise 6.5. Let $X \sim \text{Beta}(1, 1)$. Show that $X \sim \mathcal{U}([0, 1])$.

Exercise 6.6. Customers arrive at a shop according to a Poisson process with rate $\lambda = 12$ per hour. Each customer independently makes a purchase with probability $p = 0.3$. What is the distribution of the number of purchasing customers per hour?

Exercise 6.7. Show that if $X \sim t_n$ (Student with n df), then $X^2 \sim F(1, n)$ (Fisher).

Exercise 6.8. Let $X \sim \text{LogN}(0, 1)$. Compute $\mathbb{E}[X]$, $\text{Var}(X)$, and the median of X.

Chapter 7

Functions of Random Variables

In science and engineering, one rarely measures the quantity of direct interest. A physicist measures an angle and deduces a velocity; a financial analyst observes a log-return and deduces a price; an engineer measures a voltage and deduces a power. Each time, a transformation g is applied to a random variable X to obtain a new random variable $Y = g(X)$. But what is the distribution of Y? The question, innocent in appearance, leads to rich and varied techniques: the change-of-variable formula, the CDF method, Jacobians for multidimensional transformations. Mastering these tools means graduating from passive observer of randomness to architect capable of manipulating distributions at will.

Intuition

When a transformation is applied to a random variable, the result is a new random variable whose distribution may be very different. This chapter develops the methods for determining the distribution of $Y = g(X)$ from that of X.

7.1 Discrete Case

Proposition 7.1 (Distribution of $g(X)$ — discrete case)**.** If X is discrete taking values $x_1, x_2, \dots$ and $g : \mathbb{R} \to \mathbb{R}$, then $Y = g(X)$ is discrete with

$$\mathbb{P}(Y = y) = \sum_{k:\, g(x_k)=y} \mathbb{P}(X = x_k).$$

Example 7.2. If $X \sim \mathrm{Bin}(n, 1/2)$ and $Y = 2X - n$, then Y takes values $-n, -n+2, \dots, n-2, n$ and

$$\mathbb{P}(Y = 2k - n) = \binom{n}{k} \left(\frac{1}{2}\right)^n, \quad k = 0, 1, \dots, n.$$

7.2 The CDF Method

Theorem 7.3 (CDF method)**.** *To find the distribution of $Y = g(X)$:*

1. *Compute $F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(g(X) \le y)$.*

2. *Differentiate: $f_Y(y) = F_Y'(y)$.*

Example 7.4 ($Y = X^2$ with $X \sim \mathcal{N}(0,1)$)**.** For $y > 0$:

$$F_Y(y) = \mathbb{P}(X^2 \le y) = \mathbb{P}(-\sqrt{y} \le X \le \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y}) - 1.$$

Differentiating:

$$f_Y(y) = 2 \cdot \frac{1}{\sqrt{2\pi}} e^{-y/2} \cdot \frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi}}\, y^{-1/2}\, e^{-y/2}, \quad y > 0.$$

This is the density of $\Gamma(1/2, 1/2) = \chi^2(1)$.

Example 7.5 ($Y = e^X$ with $X \sim \mathcal{N}(\mu, \sigma^2)$)**.** For $y > 0$:

$$F_Y(y) = \mathbb{P}(e^X \le y) = \mathbb{P}(X \le \ln y) = \Phi\left(\frac{\ln y - \mu}{\sigma}\right).$$

Differentiating:

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right), \quad y > 0.$$

This is the log-normal density.

7.3 The Change-of-Variable Formula

Theorem 7.6 (Change of variable — monotone case)**.** *Let X be a continuous variable with density f_X and g a **strictly monotone** C^1 function on the support of X, with C^1 inverse g^{-1}. Then $Y = g(X)$ has density*

$$f_Y(y) = f_X\big(g^{-1}(y)\big) \cdot \left|(g^{-1})'(y)\right|.$$

Proof. Suppose g is strictly increasing (the decreasing case is analogous).

$$F_Y(y) = \mathbb{P}(g(X) \le y) = \mathbb{P}(X \le g^{-1}(y)) = F_X(g^{-1}(y)).$$

Differentiating by the chain rule: $f_Y(y) = f_X(g^{-1}(y)) \cdot (g^{-1})'(y)$. Since g is increasing, $(g^{-1})' > 0$, so the absolute value is unnecessary. In the decreasing case a minus sign appears, absorbed by $|\cdot|$. □

Example 7.7 (Affine transformation)**.** If $Y = aX + b$ with $a \ne 0$: $g^{-1}(y) = (y - b)/a$, $(g^{-1})'(y) = 1/a$. Hence:

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

In particular, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim \mathcal{N}(0,1)$.

7.4 Non-Monotone Transformations

Theorem 7.8 (General change of variable)**.** *If g is C^1 and the support of X can be decomposed into finitely many intervals $I_1, \dots, I_m$ on each of which g is strictly monotone, with inverses $h_1, \dots, h_m$, then*

$$f_Y(y) = \sum_{j=1}^{m} f_X(h_j(y)) \cdot \left|h_j'(y)\right| \mathbf{1}_{g(I_j)}(y).$$

Example 7.9 ($Y = X^2$ with symmetric f_X)**.** $g(x) = x^2$ is decreasing on $(-\infty, 0)$ and increasing on $(0, +\infty)$. The two inverses are $h_1(y) = -\sqrt{y}$ and $h_2(y) = \sqrt{y}$, with $\left|h_j'(y)\right| = 1/(2\sqrt{y})$. Hence:

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}, \quad y > 0.$$

7.5 Moment Generating Functions

Definition 7.10 (Moment generating function (MGF))**.** The **moment generating function** of X is

$$M_X(t) = \mathbb{E}[e^{tX}],$$

defined for those t in a neighbourhood of 0 where this expectation exists.

Proposition 7.11 (Properties of the MGF)**.** (i) $M_X(0) = 1$.

(ii) $M_X^{(n)}(0) = \mathbb{E}[X^n]$ (moments are obtained by differentiation).

(iii) If $M_X(t) = M_Y(t)$ for all t in a neighbourhood of 0, then X and Y have the same distribution.

(iv) If $X \perp\!\!\!\perp Y$, then $M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$.

Example 7.12 (MGF of the normal distribution)**.** If $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

We recover $M_X'(0) = \mu = \mathbb{E}[X]$ and $M_X''(0) = \sigma^2 + \mu^2 = \mathbb{E}[X^2]$.

7.6 Probability Integral Transform

Theorem 7.13 (Probability integral transform)**.** *If X is a continuous variable with strictly increasing CDF F_X, then $U = F_X(X) \sim \mathcal{U}([0,1])$.*

Proof. For $u \in (0,1)$: $\mathbb{P}(U \leq u) = \mathbb{P}(F_X(X) \leq u) = \mathbb{P}(X \leq F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u$. □

Remark 7.14. This result is the theoretical basis for the inverse transform simulation method (Chapter 4) and for goodness-of-fit tests (p-values).

7.7 Exercises

Exercise 7.1. Let $X \sim \mathcal{U}([0,1])$ and $Y = -\ln X$. Find the distribution of Y.

Exercise 7.2. Let $X \sim \text{Exp}(1)$. Find the density of $Y = \sqrt{X}$.

Exercise 7.3. Let $X \sim \mathcal{N}(0,1)$. Find the density of $Y = |X|$ (half-normal distribution).

Exercise 7.4. Let X have density $f_X(x) = 2x\,\mathbf{1}_{[0,1]}(x)$ and $Y = 1 - X^2$. Find the density of Y.

Exercise 7.5. Compute the MGF of $\mathrm{Exp}(\lambda)$ and deduce $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, $\mathrm{Var}(X)$.

Exercise 7.6. Let $X \sim \mathrm{Cauchy}(0,1)$ with density $f(x) = 1/(\pi(1+x^2))$. Show that the MGF of X does not exist (i.e. $\mathbb{E}[e^{tX}] = \infty$ for all $t \neq 0$).

Exercise 7.7 (Box–Muller)**.** Let U_1, U_2 be i.i.d. $\mathcal{U}([0,1])$. Define $Z_1 = \sqrt{-2\ln U_1}\cos(2\pi U_2)$ and $Z_2 = \sqrt{-2\ln U_1}\sin(2\pi U_2)$. Show that Z_1, Z_2 are i.i.d. $\mathcal{N}(0,1)$.

Exercise 7.8. Using the MGF, show that if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Chapter 8

Joint Distributions and Covariance

Until now, we have studied random variables in isolation. But in the real world, quantities are rarely independent: height and weight are correlated, a stock's return depends on the overall market, temperature and atmospheric pressure evolve jointly. To understand these dependencies, one must move from the law of a single variable to the *joint distribution* of several variables. Francis Galton, in the 1880s, was the first to study correlation systematically — by measuring the heights of parents and their children. Karl Pearson then formalized the correlation coefficient that bears his name. This chapter develops the fundamental tools: joint distributions, marginals, conditionals, covariance, and correlation.

Intuition

This chapter studies the joint distributions of two (or more) random variables, marginal and conditional distributions, covariance, and correlation. These tools are essential for modelling dependence between random quantities.

8.1 Joint Distributions

8.1.1 Discrete case

Definition 8.1 (Discrete joint distribution)**.** Let (X, Y) be a pair of discrete random variables. The **joint distribution** is the function

$$p_{X,Y}(x_i, y_j) = \mathbb{P}(X = x_i, Y = y_j), \quad \text{with } \sum_{i,j} p_{X,Y}(x_i, y_j) = 1.$$

8.1.2 Continuous case

Definition 8.2 (Joint density)**.** The pair (X, Y) has a **joint density** $f_{X,Y}$ if, for every Borel set $B \subseteq \mathbb{R}^2$:

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

with $f_{X,Y}(x, y) \geq 0$ and $\iint_{\mathbb{R}^2} f_{X,Y}(x, y)\, dx\, dy = 1$.

Definition 8.3 (Joint CDF)**.**

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

In the continuous case: $f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x\,\partial y}(x,y)$.

8.2 Marginal Distributions

Definition 8.4 (Marginals)**.** The **marginal distributions** are obtained by summing (discrete) or integrating (continuous) over the other variable:

$$\begin{aligned}
&\text{Discrete:} && p_X(x_i) = \sum_j p_{X,Y}(x_i, y_j), && p_Y(y_j) = \sum_i p_{X,Y}(x_i, y_j).\\
&\text{Continuous:} && f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy, && f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx.
\end{aligned}$$

The joint determines the marginals, but not vice versa

From $f_{X,Y}$ one can compute f_X and f_Y. But knowing only f_X and f_Y is not enough to reconstruct $f_{X,Y}$ without information about the dependence between X and Y.

Example 8.5. Let (X, Y) have joint density $f_{X,Y}(x,y) = 6(1-y)$ for $0 < x < y < 1$ and 0 otherwise. Then:

$$f_X(x) = \int_x^1 6(1-y)\,dy = 3(1-x)^2, \quad 0 < x < 1,$$

$$f_Y(y) = \int_0^y 6(1-y)\,dx = 6y(1-y), \quad 0 < y < 1.$$

8.3 Conditional Distributions

Definition 8.6 (Conditional density)**.** The **conditional density** of Y given $X = x$ is

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \quad f_X(x) > 0.$$

Similarly in the discrete case: $\mathbb{P}(Y = y_j \mid X = x_i) = p_{X,Y}(x_i, y_j)/p_X(x_i)$.

8.4 Independence

Theorem 8.7 (Characterisation of independence)**.** *X and Y are independent if and only if:*

- ***Discrete:*** *$p_{X,Y}(x,y) = p_X(x)\,p_Y(y)$ for all (x,y).*
- ***Continuous:*** *$f_{X,Y}(x,y) = f_X(x)\,f_Y(y)$ for all (x,y).*
- ***Via the CDF:*** *$F_{X,Y}(x,y) = F_X(x)\,F_Y(y)$ for all (x,y).*

Proposition 8.8 (Factorisation criterion)**.** The continuous pair (X, Y) is independent if and only if there exist functions g, h such that $f_{X,Y}(x,y) = g(x)\,h(y)$ on the support (which must be a Cartesian product).

8.5 Covariance and Correlation

Definition 8.9 (Covariance)**.**

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

Proposition 8.10 (Properties of covariance)**.** (i) $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

(ii) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$ (symmetry).

(iii) $\mathrm{Cov}(aX+b, cY+d) = ac\,\mathrm{Cov}(X,Y)$ (bilinearity).

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

(v) If $X \perp\!\!\!\perp Y$, then $\mathrm{Cov}(X,Y) = 0$.

(vi) $\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum_{i<j}\mathrm{Cov}(X_i,X_j)$.

Definition 8.11 (Correlation coefficient)**.**

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

Theorem 8.12 (Properties of ρ)**.** *(i)* $-1 \le \rho(X,Y) \le 1$ *(consequence of Cauchy–Schwarz).*

(ii) $|\rho(X,Y)| = 1$ *iff* $Y = aX+b$ *a.s. with* $a \neq 0$.

(iii) $\rho(X,Y) = 0$ *means* X *and* Y *are* ***uncorrelated****.*

Uncorrelated $\neq$ independent

$\mathrm{Cov}(X,Y) = 0$ does not imply independence. Example: if $X \sim \mathcal{N}(0,1)$ and $Y = X^2$, then $\mathrm{Cov}(X,Y) = \mathbb{E}[X^3] = 0$ (by symmetry), yet X and Y are clearly not independent.

8.6 Bivariate Normal Distribution

Definition 8.13 (Bivariate normal)**.** The vector (X,Y) has a **bivariate normal** distribution $\mathcal{N}_2(\boldsymbol{\mu},\Sigma)$ with $\boldsymbol{\mu} = (\mu_X,\mu_Y)^T$ and $\Sigma = \left(\begin{smallmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{smallmatrix}\right)$ if

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

Theorem 8.14 (Properties of the bivariate normal)**.** *(i) The marginals are* $X \sim \mathcal{N}(\mu_X,\sigma_X^2)$ *and* $Y \sim \mathcal{N}(\mu_Y,\sigma_Y^2)$*.*

(ii) $\rho(X,Y) = \rho$ *(the distribution parameter).*

(iii) X *and* Y *are independent if and only if* $\rho = 0$*.*

(iv) Every linear combination $aX+bY$ *is normal.*

Remark 8.15. Property (iii) is specific to the normal distribution: in general, uncorrelated does not imply independent, but for a Gaussian vector it does.

8.7 Sum of Random Variables

Theorem 8.16 (Convolution)**.** *If X and Y are independent with densities f_X and f_Y, then $Z = X + Y$ has density given by the **convolution**:*

$$f_Z(z) = (f_X * f_Y)(z) = \int_{-\infty}^{+\infty} f_X(t)\, f_Y(z-t)\, dt.$$

Example 8.17. If $X \sim \mathrm{Exp}(\lambda)$ and $Y \sim \mathrm{Exp}(\lambda)$ are independent, then for $z > 0$:

$$f_Z(z) = \int_0^z \lambda e^{-\lambda t}\, \lambda e^{-\lambda(z-t)}\, dt = \lambda^2 e^{-\lambda z} \int_0^z dt = \lambda^2 z\, e^{-\lambda z}.$$

This is the density of $\Gamma(2, \lambda)$.

8.8 Exercises

Exercise 8.1. Let (X, Y) have density $f(x, y) = c\,xy$ for $0 < x < 1$, $0 < y < 1$ and 0 otherwise. Find c, the marginals, and check independence.

Exercise 8.2. Let (X, Y) have density $f(x, y) = 2$ for $0 < x < y < 1$, 0 otherwise. Compute f_X, f_Y, $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\mathrm{Cov}(X, Y)$, $\rho(X, Y)$.

Exercise 8.3. Show that if $(X, Y) \sim \mathcal{N}_2$ with $\rho = 0$, then X and Y are independent (factorise the joint density).

Exercise 8.4. Let $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$. Verify that $\mathrm{Cov}(X, Y) = 0$ but X and Y are not independent.

Exercise 8.5. Find by convolution the density of $Z = X + Y$ where X, Y are i.i.d. $\mathcal{N}(0, 1)$.

Exercise 8.6. Let (X, Y) be uniform on the disc $\{(x, y) : x^2 + y^2 \leq 1\}$. Find the marginals, $\mathbb{E}[X]$, $\mathrm{Var}(X)$, $\mathrm{Cov}(X, Y)$.

Exercise 8.7. A point (X, Y) is chosen uniformly in the triangle with vertices $(0, 0)$, $(1, 0)$, $(0, 1)$. Find the joint density, the marginals, and compute $\rho(X, Y)$.

Exercise 8.8. Let X_1, X_2, X_3 be i.i.d. $\mathrm{Exp}(1)$. Compute $\mathrm{Cov}(X_1 + X_2,\ X_2 + X_3)$.

Chapter 9

Laws of Large Numbers

Why does the observed frequency of an event stabilize around its theoretical probability when one repeats the experiment many times? This question, at the heart of probabilistic thinking, received its first answer from Jacob Bernoulli in 1713, in his posthumously published *Ars Conjectandi*. Bernoulli showed that the empirical frequency converges to the probability — but his proof was long and laborious. Chebyshev, in 1867, dramatically simplified the proof using his inequality. Kolmogorov, in 1933, crowned the edifice with the *strong law of large numbers*, which guarantees almost sure convergence. This chapter traces this progression, from concentration inequalities to the various modes of stochastic convergence.

Intuition

The law of large numbers is one of the most fundamental results in probability theory. It justifies the frequentist interpretation of probability: the empirical frequency of an event converges to its theoretical probability as the number of experiments increases.

9.1 Modes of Convergence

Definition 9.1 (Convergence in probability)**.** A sequence (X_n) **converges in probability** to X, written $X_n \xrightarrow{\mathbb{P}} X$, if for every $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

Definition 9.2 (Almost sure convergence)**.** (X_n) **converges almost surely** to X, written $X_n \xrightarrow{\text{a.s.}} X$, if

$$\mathbb{P}\Big(\lim_{n\to\infty} X_n = X\Big) = 1,$$

or equivalently, $\mathbb{P}(\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} \{|X_n - X| > \varepsilon\}) = 0$ for every $\varepsilon > 0$.

Definition 9.3 (Convergence in L^p)**.** (X_n) **converges in** L^p ($p \geq 1$) to X, written $X_n \xrightarrow{L^p} X$, if

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0.$$

Theorem 9.4 (Hierarchy of convergences)**.**

$$X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{\mathbb{P}} X, \qquad X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{\mathbb{P}} X.$$

CHAPTER 9. LAWS OF LARGE NUMBERS

The converses are false in general. However, $X_n \xrightarrow{\mathbb{P}} X$ *implies the existence of a subsequence* $X_{n_k} \xrightarrow{a.s.} X$*.*

a.s. convergence and L^p convergence

Almost sure convergence does not imply L^p convergence, nor the converse. Each has its own conditions.

9.2 Weak Law of Large Numbers

Theorem 9.5 (Weak law of large numbers (WLLN))**.** *Let* $(X_n)_{n\geq 1}$ *be a sequence of* ***independent and identically distributed*** *(i.i.d.) random variables with* $\mathbb{E}[X_1] = \mu$ *and* $\operatorname{Var}(X_1) = \sigma^2 < \infty$*. Set* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$*. Then*

$$\bar{X}_n \xrightarrow{\mathbb{P}} \mu, \quad \textit{i.e.} \quad \forall\, \varepsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}(\left|\bar{X}_n - \mu\right| > \varepsilon) = 0.$$

Proof. This is a direct application of Chebyshev's inequality.

$$\mathbb{P}(\left|\bar{X}_n - \mu\right| > \varepsilon) \leq \frac{\operatorname{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\operatorname{Var}\left(\frac{1}{n}\sum X_i\right)}{\varepsilon^2} = \frac{1}{n^2}\cdot\frac{n\sigma^2}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow[n\to\infty]{} 0. \qquad \square$$

Remark 9.6. The WLLN remains true under the weaker assumption that the X_i are uncorrelated (not necessarily independent) with the same first two moments.

Meaning of the WLLN

The sample mean $\bar{X}_n$ concentrates around μ as n grows. This is the theoretical foundation of Monte Carlo methods and statistical estimation.

9.3 Strong Law of Large Numbers

Theorem 9.7 (Strong law of large numbers (SLLN))**.** *Let* $(X_n)_{n\geq 1}$ *be i.i.d. with* $\mathbb{E}[|X_1|] < \infty$ *and* $\mathbb{E}[X_1] = \mu$*. Then*

$$\bar{X}_n \xrightarrow{a.s.} \mu, \quad \textit{i.e.} \quad \mathbb{P}\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1.$$

Remark 9.8. The SLLN requires only the existence of the first moment (no finite variance). The full proof uses truncation and the Borel–Cantelli lemma.

Proof under the assumption $\mathbb{E}[X_1^4] < \infty$*.* We give the proof under the stronger assumption $\mathbb{E}[X_1^4] < \infty$. WLOG assume $\mu = 0$. Set $S_n = \sum_{i=1}^n X_i$.

$$\mathbb{E}[S_n^4] = \mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^4\right] = \sum_i \mathbb{E}[X_i^4] + \binom{4}{2}\sum_{i\neq j}\mathbb{E}[X_i^2]\mathbb{E}[X_j^2]$$

(cross terms with odd powers vanish by independence and $\mathbb{E}[X_i] = 0$). Hence $\mathbb{E}[S_n^4] = n\mathbb{E}[X_1^4] + 3n(n-1)(\mathbb{E}[X_1^2])^2 \leq Cn^2$ for some constant C. By Markov's inequality:

$$\mathbb{P}(\left|\bar{X}_n\right| > \varepsilon) = \mathbb{P}(|S_n| > n\varepsilon) \leq \frac{\mathbb{E}[S_n^4]}{n^4\varepsilon^4} \leq \frac{C}{n^2\varepsilon^4}.$$

Since $\sum_n 1/n^2 < \infty$, the first Borel–Cantelli lemma gives $\mathbb{P}(\limsup\{\left|\bar{X}_n\right| > \varepsilon\}) = 0$ for every $\varepsilon > 0$, so $\bar{X}_n \to 0$ a.s. $\square$

9.4 Applications

9.4.1 Frequentist interpretation

Corollary 9.9. *Let A be an event with probability p. If the experiment is repeated n times independently, the empirical frequency $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{A_i}$ converges almost surely to p.*

9.4.2 Monte Carlo method

Example 9.10 (Estimating π)**.** Let $(X_i, Y_i)_{i\geq 1}$ be i.i.d. uniform on $[0,1]^2$. Set $Z_i = \mathbf{1}_{X_i^2+Y_i^2\leq 1}$. Then $\mathbb{E}[Z_i] = \pi/4$ and by the law of large numbers:

$$\frac{4}{n}\sum_{i=1}^{n} Z_i \xrightarrow{\text{a.s.}} \pi.$$

9.4.3 Glivenko–Cantelli theorem

Theorem 9.11 (Glivenko–Cantelli)**.** *Let (X_i) be i.i.d. with CDF F and let $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i\leq x}$ be the empirical CDF. Then*

$$\sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| \xrightarrow{a.s.} 0.$$

9.5 Convergence in Distribution (Introduction)

Definition 9.12 (Convergence in distribution)**.** (X_n) **converges in distribution** to X, written $X_n \xrightarrow{\mathcal{L}} X$, if for every bounded continuous $f : \mathbb{R} \to \mathbb{R}$:

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)].$$

Equivalently: $F_{X_n}(x) \to F_X(x)$ at every continuity point x of F_X.

Theorem 9.13 (Complete hierarchy)**.**

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{\mathcal{L}} X.$$

Convergence in distribution is the weakest. It concerns only the laws, not the variables themselves.

Summary of convergences

Convergence	**Definition**
Almost sure	$\mathbb{P}(\lim X_n = X) = 1$
In probability	$\mathbb{P}(\lvert X_n - X\rvert > \varepsilon) \to 0$
In L^p	$\mathbb{E}[\lvert X_n - X\rvert^p] \to 0$
In distribution	$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for bounded continuous f

9.6 Exercises

Exercise 9.1. Let $X_n \sim \mathcal{U}([0,1/n])$. Show that $X_n \xrightarrow{\mathbb{P}} 0$ and $X_n \xrightarrow{L^2} 0$. Is the convergence almost sure?

Exercise 9.2. Let (X_n) be i.i.d. with $\mathbb{P}(X_n = 1) = \mathbb{P}(X_n = -1) = 1/2$. Show that $\bar{X}_n \to 0$ a.s. and that $\mathbb{P}(\bar{X}_n = 0) \to 0$ when n is odd. Comment.

Exercise 9.3. Give an example of a sequence (X_n) converging in probability to 0 but not almost surely. *Hint:* consider indicators of intervals sliding over $[0,1]$.

Exercise 9.4. Draw n numbers independently and uniformly from $[0,1]$. Let $M_n = \max(X_1, \dots, X_n)$. Show that $M_n \to 1$ a.s.

Exercise 9.5. Let (X_n) be i.i.d. Exp(1). Show that $\frac{1}{n}\max(X_1, \dots, X_n) \to 0$ a.s., and even that $\max(X_1, \dots, X_n)/\ln n \to 1$ a.s.

Exercise 9.6. Let (X_n) be i.i.d. standard Cauchy. Show that $\bar{X}_n$ does not converge in probability to any constant. *Hint:* use characteristic functions.

Exercise 9.7 (Monte Carlo)**.** We wish to estimate $I = \int_0^1 e^{-x^2}\,dx$ by Monte Carlo. Write the Monte Carlo estimator and, using Chebyshev's inequality, determine the number of samples needed so that the error is less than 0.01 with probability 0.95.

Exercise 9.8. Let (X_n) be i.i.d. with $\mathbb{E}[X_1] = 0$, $\mathrm{Var}(X_1) = 1$. Show by the WLLN that $S_n/n \to 0$ in probability, then that $S_n^2/n^2 \to 0$ in probability. Deduce a result about $S_n/n^{3/4}$.

Chapter 10

Central Limit Theorem

Intuition

The central limit theorem (CLT) is arguably the most celebrated result in probability theory. It states that the normalised sum of i.i.d. random variables converges in distribution to the normal law, regardless of the common distribution, provided the variance is finite.

10.1 Statement of the Theorem

Theorem 10.1 (Central limit theorem)**.** *Let* $(X_n)_{n\geq 1}$ *be i.i.d. with* $\mathbb{E}[X_1] = \mu$ *and* $\mathrm{Var}(X_1) = \sigma^2 \in (0, \infty)$*. Set* $S_n = \sum_{i=1}^n X_i$*. Then*

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

That is, for every $x \in \mathbb{R}$*:*

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt.$$

Why is the normal distribution so ubiquitous?

The CLT explains why so many natural phenomena are approximately normal: any quantity resulting from the sum of many small independent contributions will be approximately Gaussian.

10.2 Proof via Characteristic Functions

We refer to Chapter 11 for the full theory of characteristic functions. Here we sketch the proof, taking the Lévy continuity theorem for granted.

Proof of the CLT. WLOG set $\mu = 0$ (otherwise replace X_i by $X_i - \mu$). Let $\varphi(t) = \mathbb{E}[e^{itX_1}]$ be the common characteristic function.

By assumption, $\varphi(0) = 1$, $\varphi'(0) = i\mathbb{E}[X_1] = 0$, and $\varphi''(0) = -\mathbb{E}[X_1^2] = -\sigma^2$. Taylor expansion gives:

$$\varphi(t) = 1 - \frac{\sigma^2 t^2}{2} + o(t^2) \quad (t \to 0).$$

The characteristic function of $Z_n = S_n/(\sigma\sqrt{n})$ is:

$$\begin{aligned}\varphi_{Z_n}(t) &= \mathbb{E}\left[\exp\left(it\cdot\frac{S_n}{\sigma\sqrt{n}}\right)\right] = \left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \\ &= \left[1 - \frac{\sigma^2}{2}\cdot\frac{t^2}{\sigma^2 n} + o\left(\frac{1}{n}\right)\right]^n = \left[1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right]^n.\end{aligned}$$

Using $\lim_{n\to\infty}(1+a/n)^n = e^a$:

$$\varphi_{Z_n}(t) \xrightarrow[n\to\infty]{} e^{-t^2/2}.$$

Since $e^{-t^2/2}$ is the characteristic function of $\mathcal{N}(0,1)$, Lévy's continuity theorem gives $Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$. □

10.3 Practical Applications

10.3.1 Normal approximation to the binomial

Corollary 10.2 (De Moivre–Laplace theorem)**.** *If* $S_n \sim \text{Bin}(n,p)$ *with* $0 < p < 1$*, then*

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Example 10.3. A fair coin is tossed 100 times. What is the probability of getting between 45 and 55 heads?

With $S_{100} \sim \text{Bin}(100, 0.5)$, $\mu = 50$, $\sigma = 5$:

$$\begin{aligned}\mathbb{P}(45 \le S_{100} \le 55) &= \mathbb{P}\left(\frac{45-50}{5} \le Z \le \frac{55-50}{5}\right) \approx \mathbb{P}(-1 \le Z \le 1) \\ &= 2\Phi(1) - 1 \approx 0.6827.\end{aligned}$$

Continuity correction

To improve the approximation, use the **continuity correction**:

$$\mathbb{P}(a \le S_n \le b) \approx \Phi\left(\frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}}\right).$$

10.3.2 Confidence intervals

Corollary 10.4 (Confidence interval for the mean)**.** *If* (X_i) *are i.i.d. with* $\mathbb{E}[X_1] = \mu$ *and known* $\text{Var}(X_1) = \sigma^2$*, a* ***95% confidence interval*** *for* μ *is*

$$\left[\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right].$$

Common quantiles for confidence intervals

Confidence level	α	$z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.960
99%	0.01	2.576

10.3.3 Normal approximation to the Poisson

Corollary 10.5. *If $X \sim \text{Poisson}(\lambda)$ with λ large, then $(X - \lambda)/\sqrt{\lambda} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.*

10.4 Multivariate CLT

Theorem 10.6 (Multivariate CLT)**.** *Let (X_n) be i.i.d. random vectors in $\mathbb{R}^d$ with $\mathbb{E}[X_1] = \boldsymbol{\mu}$ and $\text{Cov}(X_1) = \Sigma$. Then*

$$\sqrt{n}\,(\bar{X}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, \Sigma).$$

10.5 The Delta Method

Theorem 10.7 (Delta method)**.** *If $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$ and g is differentiable at μ with $g'(\mu) \neq 0$, then*

$$\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2[g'(\mu)]^2).$$

Example 10.8. If $\bar{X}_n \to p$ and $g(x) = x(1 - x)$, then $g'(p) = 1 - 2p$ and

$$\sqrt{n}(\bar{X}_n(1 - \bar{X}_n) - p(1 - p)) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\, p(1 - p)(1 - 2p)^2).$$

10.6 Rate of Convergence: Berry–Esseen

Theorem 10.9 (Berry–Esseen)**.** *Under the CLT hypotheses, if $\mathbb{E}[|X_1|^3] < \infty$:*

$$\sup_{x \in \mathbb{R}} |\mathbb{P}(Z_n \leq x) - \Phi(x)| \leq \frac{C\,\mathbb{E}[|X_1 - \mu|^3]}{\sigma^3 \sqrt{n}},$$

where C is a universal constant ($C \leq 0.4748$).

Remark 10.10. Berry–Esseen shows that convergence in the CLT is of order $1/\sqrt{n}$. This justifies using the normal approximation once n is moderately large (typically $n \geq 30$).

10.7 Exercises

Exercise 10.1. Let (X_i) be i.i.d. with $\mathbb{E}[X_1] = 2$, $\text{Var}(X_1) = 9$. Approximate $\mathbb{P}(S_{100} > 220)$.

Exercise 10.2. A fair die is rolled 360 times. Approximate the probability that the sum exceeds 1300.

Exercise 10.3. Let $S_n \sim \text{Bin}(400, 0.4)$. Compute approximately $\mathbb{P}(S_n \geq 180)$ with and without continuity correction.

Exercise 10.4. A physical quantity is measured $n = 50$ times. The measurements are i.i.d. with variance $\sigma^2 = 4$. What is the 99% confidence interval for the mean?

Exercise 10.5. How many polls are needed to estimate a proportion p to within 2% with 95% confidence? (The value of p is unknown.)

Exercise 10.6. Using the delta method, find the limiting distribution of $\sqrt{n}(\ln \bar{X}_n - \ln \mu)$ when the X_i are i.i.d. positive with $\mathbb{E}[X_1] = \mu > 0$ and $\text{Var}(X_1) = \sigma^2$.

Exercise 10.7. Verify the CLT for $X_i \sim \text{Exp}(\lambda)$ by directly computing the characteristic function of Z_n and showing it converges to $e^{-t^2/2}$.

Exercise 10.8. Let (X_i) be i.i.d. Poisson(4). Approximate $\mathbb{P}\big(\sum_{i=1}^{100} X_i \leq 380\big)$.

Chapter 11

Characteristic Functions

Intuition

The characteristic function is a powerful analytic tool that transforms probabilistic problems into problems of complex analysis. It always exists (unlike the MGF), uniquely determines the distribution, and provides an elegant route to the central limit theorem.

11.1 Definition and Basic Properties

Definition 11.1 (Characteristic function)**.** The **characteristic function** of a real-valued random variable X is

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \mathbb{E}[\cos(tX)] + i\,\mathbb{E}[\sin(tX)], \quad t \in \mathbb{R}.$$

Theorem 11.2 (Fundamental properties)**.** *Let φ_X be the characteristic function of X. Then:*

(i) $\varphi_X(0) = 1$.

(ii) $|\varphi_X(t)| \le 1$ *for all* $t \in \mathbb{R}$.

(iii) φ_X *is* ***uniformly continuous*** *on* $\mathbb{R}$.

(iv) $\varphi_X(-t) = \overline{\varphi_X(t)}$ *(complex conjugate).*

(v) If $Y = aX + b$*, then* $\varphi_Y(t) = e^{ibt}\,\varphi_X(at)$.

(vi) If $X \perp\!\!\!\perp Y$*, then* $\varphi_{X+Y}(t) = \varphi_X(t)\,\varphi_Y(t)$.

Proof of (iii).

$$|\varphi_X(t+h) - \varphi_X(t)| = \left|\mathbb{E}[e^{itX}(e^{ihX} - 1)]\right| \le \mathbb{E}[\left|e^{ihX} - 1\right|].$$

Since $\left|e^{ihX} - 1\right| \le 2$ and $e^{ihX} - 1 \to 0$ as $h \to 0$, dominated convergence gives $\mathbb{E}[\left|e^{ihX} - 1\right|] \to 0$ uniformly in t. □

11.2 Moments from the Characteristic Function

Theorem 11.3 (Moments and derivatives). *If $\mathbb{E}[|X|^n] < \infty$, then φ_X is n times differentiable and*

$$\varphi_X^{(k)}(0) = i^k\,\mathbb{E}[X^k], \quad k = 0, 1, \ldots, n.$$

In particular: $\mathbb{E}[X] = \varphi_X'(0)/i$ *and* $\mathbb{E}[X^2] = -\varphi_X''(0)$*, whence* $\mathrm{Var}(X) = -\varphi_X''(0) - [\varphi_X'(0)/i]^2$.

11.3 Characteristic Functions of Standard Distributions

Table of characteristic functions

Distribution of X	$\varphi_X(t)$
Bernoulli(p)	$1 - p + pe^{it}$
Bin(n, p)	$(1 - p + pe^{it})^n$
Poisson(λ)	$\exp(\lambda(e^{it} - 1))$
Geom(p)	$\dfrac{pe^{it}}{1 - (1-p)e^{it}}$
$\mathcal{U}([a, b])$	$\dfrac{e^{itb} - e^{ita}}{it(b-a)}$
Exp(λ)	$\dfrac{\lambda}{\lambda - it}$
$\mathcal{N}(\mu, \sigma^2)$	$\exp\left(i\mu t - \dfrac{\sigma^2 t^2}{2}\right)$
$\Gamma(\alpha, \lambda)$	$\left(\dfrac{\lambda}{\lambda - it}\right)^\alpha$
Cauchy$(0, 1)$	$e^{-\lvert t\rvert}$

Example 11.4 (Computation for the normal distribution). If $X \sim \mathcal{N}(0, 1)$:

$$\begin{aligned}\varphi_X(t) &= \int_{-\infty}^{+\infty} e^{itx}\,\frac{1}{\sqrt{2\pi}}\,e^{-x^2/2}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-(x^2-2itx)/2}\,dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-(x-it)^2/2}\,e^{-t^2/2}\,dx = e^{-t^2/2}.\end{aligned}$$

For $X \sim \mathcal{N}(\mu, \sigma^2)$, write $X = \sigma Z + \mu$, so $\varphi_X(t) = e^{i\mu t}\,e^{-\sigma^2 t^2/2} = e^{i\mu t - \sigma^2 t^2/2}$.

11.4 Uniqueness and Inversion Theorems

Theorem 11.5 (Uniqueness theorem). *Two random variables X and Y have the same distribution if and only if $\varphi_X(t) = \varphi_Y(t)$ for all $t \in \mathbb{R}$.*

Theorem 11.6 (Inversion formula)**.** *If $\varphi_X \in L^1(\mathbb{R})$ (i.e. $\int |\varphi_X(t)|\,dt < \infty$), then X has a continuous density given by*

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx}\,\varphi_X(t)\,dt.$$

Remark 11.7. The inversion formula is the probabilistic analogue of the inverse Fourier transform. The characteristic function is simply the Fourier transform of the distribution of X.

11.5 Lévy's Continuity Theorem

Theorem 11.8 (Lévy's continuity theorem)**.** *Let (X_n) be a sequence of random variables with $\varphi_n = \varphi_{X_n}$.*

(i) If $X_n \xrightarrow{\mathcal{L}} X$, then $\varphi_n(t) \to \varphi_X(t)$ for all t.

(ii) Conversely, if $\varphi_n(t) \to \psi(t)$ for all t and ψ is ***continuous at*** 0*, then ψ is the characteristic function of some random variable X and $X_n \xrightarrow{\mathcal{L}} X$.*

Remark 11.9. This theorem is used in the proof of the CLT (Chapter 10): one shows pointwise convergence $\varphi_{Z_n}(t) \to e^{-t^2/2}$, and continuity at 0 of the limit is obvious.

11.6 Application: Stability of the Normal Distribution

Theorem 11.10 (Stability of the normal)**.** *If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then*

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2).$$

Proof.

$$\begin{aligned} \varphi_{X_1+X_2}(t) &= \varphi_{X_1}(t)\,\varphi_{X_2}(t) \\ &= e^{i\mu_1 t - \sigma_1^2 t^2/2}\,e^{i\mu_2 t - \sigma_2^2 t^2/2} \\ &= e^{i(\mu_1+\mu_2)t - (\sigma_1^2+\sigma_2^2)t^2/2}, \end{aligned}$$

which is the CF of $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. The result follows by uniqueness. □

11.7 Application: Stability of the Poisson Distribution

Proposition 11.11. If $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ are independent, then $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Proof. $\varphi_{X_1+X_2}(t) = e^{\lambda_1(e^{it}-1)} \cdot e^{\lambda_2(e^{it}-1)} = e^{(\lambda_1+\lambda_2)(e^{it}-1)}$, the CF of $\text{Poisson}(\lambda_1 + \lambda_2)$. □

11.8 Characteristic Functions and Convergence in Distribution

Example 11.12 (Poisson to normal convergence)**.** Let $X_n \sim \text{Poisson}(n)$. Set $Z_n = (X_n - n)/\sqrt{n}$.

$$\begin{aligned}\varphi_{Z_n}(t) &= e^{-it\sqrt{n}}\,\varphi_{X_n}(t/\sqrt{n}) = e^{-it\sqrt{n}}\,e^{n(e^{it/\sqrt{n}}-1)}\\ &= \exp\left(n\left(e^{it/\sqrt{n}} - 1 - \frac{it}{\sqrt{n}}\right)\right).\end{aligned}$$

Taylor expansion: $e^{it/\sqrt{n}} = 1 + it/\sqrt{n} - t^2/(2n) + o(1/n)$, so

$$\varphi_{Z_n}(t) = \exp\left(n\left(-\frac{t^2}{2n} + o(1/n)\right)\right) \to e^{-t^2/2}.$$

By Lévy's theorem, $Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

11.9 Exercises

Exercise 11.1. Compute the characteristic function of $\mathcal{U}(\{-1,0,1\})$.

Exercise 11.2. Let $X \sim \text{Exp}(\lambda)$. Compute $\varphi_X(t)$ and deduce $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, $\text{Var}(X)$.

Exercise 11.3. Show that $\varphi_X(t) = e^{-|t|}$ is the characteristic function of the standard Cauchy distribution. Deduce that if X_1, X_2 are i.i.d. standard Cauchy, then $(X_1 + X_2)/2$ is again standard Cauchy.

Exercise 11.4. Using the inversion formula, recover the density of $\mathcal{N}(0,1)$ from its CF $\varphi(t) = e^{-t^2/2}$.

Exercise 11.5. Show that the CF of $\Gamma(\alpha, \lambda)$ is $(\lambda/(\lambda - it))^\alpha$. Deduce that the sum of n i.i.d. $\text{Exp}(\lambda)$ variables follows a $\Gamma(n, \lambda)$ distribution.

Exercise 11.6. Let $X_n \sim \text{Bin}(n, \lambda/n)$ with $\lambda > 0$ fixed. By computing $\varphi_{X_n}(t)$, recover the Poisson approximation: $X_n \xrightarrow{\mathcal{L}} \text{Poisson}(\lambda)$.

Exercise 11.7. Show that if φ_X is real-valued, then X and $-X$ have the same distribution (the distribution of X is symmetric about 0).

Exercise 11.8. Let (X_n) be i.i.d. with $\varphi_{X_1}(t) = 1/(1+t^2)$ (Cauchy). Show that $\bar{X}_n$ has the same distribution as X_1 and explain why the CLT does not apply.

Chapter 12

Introduction to Markov Chains

12.1 Introduction

Markov chains are one of the most fundamental stochastic models. They describe the random evolution of a system whose future depends on the past only through the present (the Markov property). Their applications span physics (random walks), biology (population dynamics), computer science (Monte Carlo algorithms, PageRank), and finance.

The Markov property: memorylessness

Imagine a token on a board that moves randomly. At each step, the probability of moving to a neighboring cell depends only on the current position, not on the past trajectory. This is the Markov property: "the future is independent of the past, given the present."

12.2 Definitions

Definition 12.1 (Markov chain)**.** Let E be a countable set (the *state space*). A stochastic process $(X_n)_{n\geq 0}$ taking values in E is a *Markov chain* if for all $n \geq 0$ and all states $i_0, \dots, i_{n-1}, i, j \in E$:

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i),$$

whenever the conditioning is well-defined ($\mathbb{P}(X_n = i, \dots) > 0$).

Definition 12.2 (Homogeneous chain and transition matrix)**.** The chain is *homogeneous* (or *time-homogeneous*) if the transition probabilities do not depend on n:

$$p_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i), \qquad \forall n \geq 0.$$

The matrix $P = (p_{ij})_{i,j\in E}$ is called the *transition matrix*. It satisfies:

(i) $p_{ij} \geq 0$ for all i, j;

(ii) $\sum_{j\in E} p_{ij} = 1$ for all i (stochastic matrix).

Definition 12.3 (Initial distribution)**.** The *initial distribution* is the law of X_0, i.e. the vector $\mu_0 = (\mathbb{P}(X_0 = i))_{i\in E}$. The law of the chain is entirely determined by μ_0 and P.

Proposition 12.4 (Joint distribution)**.** For all $i_0, i_1, \ldots, i_n \in E$:

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \mu_0(i_0)\, p_{i_0 i_1}\, p_{i_1 i_2} \cdots p_{i_{n-1} i_n}.$$

Example 12.5 (Random walk on $\mathbb{Z}$)**.** Let $X_0 = 0$ and $X_{n+1} = X_n + \xi_{n+1}$ where the (ξ_n) are i.i.d. with $\mathbb{P}(\xi = +1) = p$ and $\mathbb{P}(\xi = -1) = 1 - p$. This is a Markov chain on $E = \mathbb{Z}$ with $p_{i,i+1} = p$ and $p_{i,i-1} = 1 - p$.

Example 12.6 (Ehrenfest model)**.** $2N$ particles are distributed between two boxes. At each step, a particle is chosen uniformly at random and moved to the other box. If X_n is the number of particles in the first box, (X_n) is a Markov chain on $\{0, 1, \ldots, 2N\}$ with $p_{k,k-1} = k/(2N)$ and $p_{k,k+1} = 1 - k/(2N)$.

12.3 n-step transition probabilities

Definition 12.7. The *n-step transition probabilities* are:

$$p_{ij}^{(n)} = \mathbb{P}(X_n = j \mid X_0 = i).$$

Theorem 12.8 (Chapman-Kolmogorov equation)**.** *For all $m, n \geq 0$ and all $i, j \in E$:*

$$p_{ij}^{(m+n)} = \sum_{k \in E} p_{ik}^{(m)}\, p_{kj}^{(n)}.$$

In matrix notation: $P^{(m+n)} = P^m \cdot P^n$, i.e. $P^{(n)} = P^n$.

Proof. By the law of total probability:

$$\begin{aligned} p_{ij}^{(m+n)} &= \mathbb{P}(X_{m+n} = j \mid X_0 = i) = \sum_{k \in E} \mathbb{P}(X_{m+n} = j \mid X_m = k)\, \mathbb{P}(X_m = k \mid X_0 = i) \\ &= \sum_{k \in E} p_{kj}^{(n)}\, p_{ik}^{(m)}. \end{aligned}$$

□

12.4 Classification of states

Definition 12.9 (Accessibility and communication)**.**

- State j is *accessible* from i, written $i \to j$, if there exists $n \geq 0$ such that $p_{ij}^{(n)} > 0$.
- States i and j *communicate*, written $i \leftrightarrow j$, if $i \to j$ and $j \to i$.

The relation $\leftrightarrow$ is an equivalence relation. Its classes are called *communicating classes*.

Definition 12.10 (Irreducible chain)**.** The chain is *irreducible* if all states communicate, i.e. there is a single communicating class.

Definition 12.11 (Return time and recurrence)**.** The *first return time* to i is $T_i = \inf\{n \geq 1 : X_n = i\}$.

- State i is *recurrent* if $\mathbb{P}_i(T_i < \infty) = 1$.

- State i is *transient* if $\mathbb{P}_i(T_i < \infty) < 1$.

Theorem 12.12 (Characterization of recurrence)**.** *State i is recurrent if and only if $\sum_{n=0}^{\infty} p_{ii}^{(n)} = +\infty$. State i is transient if and only if $\sum_{n=0}^{\infty} p_{ii}^{(n)} < +\infty$.*

Proof. Let $N_i = \sum_{n=1}^{\infty} \nVdash_{\{X_n=i\}}$ be the number of returns to i. Then $\mathbb{E}_i[N_i] = \sum_{n=1}^{\infty} p_{ii}^{(n)}$. Moreover, $\mathbb{P}_i(N_i \geq k) = \mathbb{P}_i(T_i < \infty)^k$, so N_i has a geometric distribution. If i is recurrent, $\mathbb{P}_i(T_i < \infty) = 1$, so $\mathbb{E}_i[N_i] = +\infty$. If i is transient, $\mathbb{P}_i(T_i < \infty) < 1$, so $\mathbb{E}_i[N_i] = \mathbb{P}_i(T_i < \infty)/(1 - \mathbb{P}_i(T_i < \infty)) < +\infty$. □

Proposition 12.13 (Class property)**.** Recurrence and transience are class properties: if i communicates with j, then i is recurrent if and only if j is recurrent.

Definition 12.14 (Absorbing state)**.** A state i is *absorbing* if $p_{ii} = 1$. An absorbing state is recurrent.

Definition 12.15 (Period)**.** The *period* of state i is $d(i) = \gcd\{n \geq 1 : p_{ii}^{(n)} > 0\}$. State i is *aperiodic* if $d(i) = 1$. The period is a class property.

12.5 Stationary distributions

Definition 12.16 (Stationary distribution)**.** A probability vector $\pi = (\pi_i)_{i \in E}$ (with $\pi_i \geq 0$ and $\sum_i \pi_i = 1$) is a *stationary distribution* (or *invariant distribution*) for P if:

$$\pi P = \pi, \qquad \text{i.e.} \quad \pi_j = \sum_{i \in E} \pi_i \, p_{ij}, \quad \forall j \in E.$$

Theorem 12.17 (Existence for irreducible positive recurrent chains)**.** *Let (X_n) be an irreducible recurrent Markov chain. Then:*

1. *There exists an invariant measure π (unique up to a multiplicative constant) given by $\pi_j = 1/\mathbb{E}_j[T_j]$.*
2. *π is a probability distribution (i.e. $\sum_j \pi_j = 1$) if and only if the chain is* positive recurrent*, i.e. $\mathbb{E}_i[T_i] < \infty$ for every (equivalently, for some) state i.*

Remark 12.18. If the state space E is finite and the chain is irreducible, then it is automatically positive recurrent and admits a unique stationary distribution.

12.6 Ergodic theorem

Theorem 12.19 (Ergodic theorem for Markov chains)**.** *Let (X_n) be an irreducible, positive recurrent, aperiodic Markov chain with stationary distribution π. Then for every initial state i and every state j:*

$$\lim_{n \to \infty} p_{ij}^{(n)} = \pi_j.$$

Moreover, for every bounded function $f : E \to \mathbb{R}$:

$$\frac{1}{n} \sum_{k=0}^{n-1} f(X_k) \xrightarrow[n \to \infty]{a.s.} \sum_{j \in E} f(j) \, \pi_j.$$

CHAPTER 12. INTRODUCTION TO MARKOV CHAINS

Key Markov chain formulas

- **Chapman-Kolmogorov:** $P^{(m+n)} = P^m \cdot P^n$
- **Distribution at time n:** $\mu_n = \mu_0 P^n$
- **Stationarity:** $\pi P = \pi$
- **Link with return time:** $\pi_j = 1/\mathbb{E}_j[T_j]$
- **Convergence:** $\lim_{n\to\infty} P^n = \mathbf{1}\pi^T$ (if irreducible, positive recurrent, aperiodic)

12.7 Reversibility and detailed balance

Definition 12.20 (Reversibility)**.** A Markov chain with transition matrix P and stationary distribution π is *reversible* if the *detailed balance equations* are satisfied:

$$\pi_i\, p_{ij} = \pi_j\, p_{ji}, \qquad \forall i, j \in E.$$

Proposition 12.21. If π satisfies the detailed balance equations, then π is a stationary distribution.

Proof. Summing over i: $\sum_i \pi_i\, p_{ij} = \sum_i \pi_j\, p_{ji} = \pi_j \sum_i p_{ji} = \pi_j$, since $\sum_i p_{ji} = 1$. □

Example 12.22 (Random walk on a graph)**.** Let $G = (V, E)$ be a connected finite undirected graph. The simple random walk on G has transition matrix $p_{ij} = 1/\deg(i)$ if (i, j) is an edge. The stationary distribution is $\pi_i = \deg(i)/(2\,|E|)$, and the detailed balance equations are satisfied.

Detailed balance $\neq$ stationarity

Detailed balance is a *sufficient* but not necessary condition for stationarity. There exist Markov chains with a stationary distribution that does not satisfy detailed balance (e.g. chains with a directed cycle).

12.8 Exercises

Exercise 12.1. Let (X_n) be the symmetric random walk on $\mathbb{Z}$ ($p = 1/2$). Show that every state is recurrent. *Hint*: use $p_{00}^{(2n)} = \binom{2n}{n} 2^{-2n} \sim (\pi n)^{-1/2}$.

Exercise 12.2. Show that the symmetric random walk on $\mathbb{Z}^3$ is transient. *Hint*: show that $\sum_n p_{00}^{(n)} < \infty$.

Exercise 12.3. Let $E = \{1, 2, 3\}$ and $P = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \end{pmatrix}$.

1. Verify that the chain is irreducible and aperiodic.
2. Compute the stationary distribution π.

3. Is the chain reversible?

Exercise 12.4 (Queueing model)**.** Let $N \geq 1$ be the capacity of a queue. At each step, a customer arrives with probability λ and a customer is served with probability μ (if the queue is non-empty), independently. Let X_n be the number of customers at time n. Show that (X_n) is a Markov chain, write its transition matrix, and find its stationary distribution when $\lambda < \mu$.

Exercise 12.5 (Metropolis-Hastings algorithm)**.** Let π be a target probability distribution on a finite set E and let Q be an irreducible transition matrix (the *proposal*). Define the transition matrix:

$$p_{ij} = q_{ij} \min\left(1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right) \quad (i \neq j), \qquad p_{ii} = 1 - \sum_{j \neq i} p_{ij}.$$

Show that π is a stationary distribution for P and that detailed balance holds.

Exercise 12.6. Let (X_n) be an irreducible Markov chain on a finite state space E with $|E| = N$. Show that there exists a unique stationary vector π and that $\pi_i > 0$ for all i.

CHAPTER 12. INTRODUCTION TO MARKOV CHAINS

Bibliography

[1] Billingsley, P., *Probability and Measure*, 3rd ed., Wiley, 1995.

[2] Feller, W., *An Introduction to Probability Theory and Its Applications, Vol. I*, 3rd ed., Wiley, 1968.

[3] Feller, W., *An Introduction to Probability Theory and Its Applications, Vol. II*, 2nd ed., Wiley, 1971.

[4] Durrett, R., *Probability: Theory and Examples*, 5th ed., Cambridge, 2019.

[5] Dacunha-Castelle, D. and Duflo, M., *Probabilités et Statistiques*, 2 vols., Masson, 1982–1983.