People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

Mohamed El Bachir El Ibrahimi University, Bordj Bou-Arreridj
Faculty of Mathematics and Computer Science
Department of Mathematics

Probability and Statistics

Course & Exercises

Prepared by:

Dr.REBIHA ZEGHDANE

Destined for Master 1 students- Mathematical Analysis and Applications (LMD)

First Year, Semester 1

Academic Year: 2025–2026

Contents

List of Tables

List of Figures

Preface

Course Overview: Probability and Statistics

Semester 1 – Master's Program in Mathematical Analysis and Applications

Master's Program Title:	Mathematical Analysis and Applications
Semester:	Semester 1
Course Unit:	Fundamental Unit
Course Title:	Probability and Statistics
Credits:	6
Coefficient:	3
Assessment Method:	Written exam (60%) and continuous assessment (40%)

Mathematics education forms the foundation of science and engineering at both undergraduate and graduate levels, as these fields heavily rely on mathematical modeling. The quality and level of mathematical instruction significantly influence the overall standard of education. This course is designed for Master 1 students specializing in mathematical analysis and applications, intended for the first semester. This course emphasizes the essential mathematics used in core mathematics disciplines. It covers probability and statistics, introducing key concepts such as discrete and continuous probability distributions, sampling distributions, correlation and probability lois analysis, and inferential statistics and its applications.

Topics include the properties of single and multiple random variables across discrete and continuous distributions, correlation and independence properties for both bivariate and multivariate cases, estimation methods, hypothesis testing, and statistical inference for both large and small sample sizes.

Course Objectives

1. To identify random variables that model uncertainty in real-world phenomena, and to distinguish clearly between discrete and continuous random variables.

2. To understand and apply fundamental probability distributions, including discrete distributions such as the Binomial and Poisson distributions, and continuous distributions such as the Normal distribution.

3. To analyze linear relationships between two variables, and to predict the behavior of a dependent variable based on changes in an independent variable.

4. To understand sampling techniques, sampling distributions of sample means and variances, and methods for estimating statistical parameters.

5. To develop a solid foundation in probability theory and statistical inference, enabling informed conclusions about populations based on both small and large samples.

Dr. Rebiha Zeghdane, September 2025

Introduction

Probability and inferential statistics serve as essential pillars across scientific disciplines, engineering fields, economics, and numerous applied research areas. These mathematical frameworks enable rigorous treatment of uncertainty, variability, and partial information frequently encountered in real-world data and phenomena. Often, experimental measurements are subject to noise, and systems are affected by random influences, necessitating the use of probabilistic models to effectively represent and analyze such randomness. Inferential statistics complements this by helping practitioners draw valid conclusions and make data-driven decisions concerning broader populations or processes from limited sample data.

The study of probability starts with modeling random experiments and defining events whose outcomes are not deterministic. Probabilities assign numerical values representing the chance of these outcomes occurring. Foundational concepts include sample spaces, event combinations, conditional probability, the notion of independence, and Bayes' theorem for updating beliefs in light of new evidence. Random variables, both discrete and continuous, translate uncertain quantities into numerical forms, with associated probability distributions such as Bernoulli, Binomial, Poisson, Uniform, Exponential, and Normal distributions serving as flexible tools to model a wide range of phenomena.

Understanding the long-term behavior of stochastic systems is a principal aim of probability theory. Descriptive statistics like expectation, variance, moments, and covariance capture key features of random variables. Limit theorems, including the Law of Large Numbers and the Central Limit theorem, demonstrate why averages computed from large independent data samples tend to follow predictable patterns, often approximating normality. These principles underpin most statistical inference techniques, particularly in scenarios involving extensive data or repeated sampling.

Inferential statistics applies these probabilistic insights to analyze real data sets. Since full population data are rarely available, inference methods estimate unknown parameters and test hypotheses using only sample observations. Estimation processes generate point or interval values that approximate population characteristics such as means and variances, while quantifying confidence in these estimates. Hypothesis testing offers structured approaches to verify scientific or engineering assumptions, employing tools like test statistics and significance levels to assess evidence strength.

Beyond univariate analysis, inferential statistics explores relationships among variables through correlation and regression methods. These analytical techniques enable modeling and prediction of dependent variables influenced by one or more independent variables, vital for diverse applications in engineering, medicine, environmental science, finance, and beyond. By controlling for confounding factors, these approaches facilitate a robust understanding and forecasting in multifaceted data contexts.

This course aims to provide a comprehensive foundation in both probability and inferential statistics, blending theoretical concepts with practical application. Learners will

acquire skills in modeling uncertain systems, calculating probabilities, designing and analyzing experiments, and making sound inferences from observed data. The knowledge gained will prepare students for advanced study and careers involving statistical reasoning, data science, stochastic modeling, and applied mathematics in various scientific and engineering domains.

These lecture notes are intended not only to serve as a dependable reference for students during and after the course, but also to act as a foundation for deeper exploration into advanced topics in probability theory, stochastic analysis, and their many applications in contemporary scientific and technological research. It is my hope that this material will stimulate curiosity, encourage independent study, and provide a stepping stone for those interested in pursuing further work or research in this rich and evolving field.

Chapter 1
Basic Probability Theory

The development of probability theory [1, 2, 4, 9, 11, 13, 16, 18, 20] can be traced back to Blaise Pascal, Pierre Fermat, and Jacques Bernoulli, who studied so-called games of chance. Subsequently, probability theory became an essential tool for analyzing datasets and quantifying the precision with which population parameters can be estimated. This framework forms the basis of mathematical, or inferential, statistics.

We study a random experiment. The sample space, also called the universe, describes all possible outcomes of the experiment. Each of these outcomes is called an *elementary event*. The sample space is often denoted by Ω, and an elementary outcome by ω.

An *event* is a subset of Ω, or equivalently, a union of elementary events. An event is said to be *realized* if one of the elementary events that composes it is realized. Events are sets and are often denoted by capital letters.

Examples:

- **Rolling one die:**
$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$
One may be interested in the event
$$A = \{\text{an even number is obtained}\} = \{2, 4, 6\}.$$

- **Rolling two dice:**
$$\Omega = \{1, \ldots, 6\} \times \{1, \ldots, 6\} = \{(i, j) : 1 \leq i \leq 6,\ 1 \leq j \leq 6\}.$$
Here, an elementary event ω is an ordered pair (i, j), where i represents the outcome of the first die and j that of the second.

- **Tossing a coin three times:** The elementary events describe the outcomes of the experiment as precisely as possible. Thus, an elementary event ω is a triplet (r_1, r_2, r_3) giving the results of the three tosses (in order).
The event
$$B = \{\text{heads on the second toss}\}$$
is given by
$$B = \{(T, H, T), (T, H, H), (H, H, T), (H, H, H)\}.$$
The event B is realized if one of the elementary events listed above occurs.

It is sometimes not necessary to know all these details. One may choose a simpler model by letting ω represent the number of heads obtained. Then,

$$\Omega = \{0, 1, 2, 3\}.$$

This model is much simpler, but it does not allow one to describe events such as B.

In this chapter, we introduce the mathematical framework of probability theory, which allows principled reasoning about uncertainty using set theory.

1.1 Probability Spaces

Our objective is to develop a mathematical framework to describe and analyze uncertain phenomena, such as the outcome of rolling a die, tomorrow's weather, or an NBA game result. To achieve this, we model the phenomenon as an experiment consisting of several mutually exclusive outcomes, which may be finite or infinite in number.

In most cases where the number of outcomes is large, it is more practical to consider sets of outcomes known as *events*. To quantify the likelihood that the experiment's outcome belongs to a particular event, we assign a *probability* to that event. Formally, this involves defining a measure, a function that assigns real numbers to sets which assigns a probability value to each event of interest.

More precisely, this modeling approach is formalized through the concept of a *probability space*.

Definition 1.1 (Probability space)**.** *A* ***probability space*** *is a triple* $(\Omega, \mathcal{F}, P)$ *consisting of:*

- *A* ***sample space*** Ω*, which contains all possible outcomes of the experiment.*
- *A set of* ***events*** $\mathcal{F}$*, which is a* σ*-algebra (see Definition 1.2 below).*
- *A* ***probability measure*** P *that assigns probabilities to the events in* $\mathcal{F}$*.*

Sample spaces may be *discrete* or *continuous*. Examples of discrete sample spaces include the possible outcomes of a coin toss, the score of a basketball game, or the number of people attending a party. Continuous sample spaces are typically intervals of $\mathbb{R}$ or $\mathbb{R}^n$ used to model quantities such as time, position, or temperature.

The term **σ-algebra** is used in measure theory to denote a collection of sets that satisfy certain conditions listed below. It may seem intimidating at first, but it essentially means that if we assign a probability to certain events (for example, it will rain tomorrow or snow tomorrow), we must also assign probabilities to their complements (i.e., it will not rain tomorrow or it will not snow tomorrow) and to their unions (i.e., it will rain or snow tomorrow).

1.1.1 σ-algebra

Definition 1.2 (σ-algebra)**.** *A* σ*-algebra* $\mathcal{F}$ *is a collection of subsets of* Ω *such that:*

1. *If* $A \in \mathcal{F}$*, then its complement* $A^c \in \mathcal{F}$*.*

2. *If $A_1, A_2 \in \mathcal{F}$, then their union $A_1 \cup A_2 \in \mathcal{F}$. This property also holds for countably infinite sequences; that is, if $A_1, A_2, \ldots \in \mathcal{F}$, then*

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

3. *The entire sample space $\Omega \in \mathcal{F}$.*

Examples of σ-algebras

Example 1.1. *Consider a finite sample space $\Omega = \{a, b\}$. The power set of Ω,*

$$\mathcal{P}(\Omega) = \{\emptyset, \{a\}, \{b\}, \{a, b\}\},$$

is a σ-algebra since it contains the empty set, Ω, and is closed under complementation and countable unions.

Example 1.2. *For any sample space Ω, the trivial σ-algebra is*

$$\mathcal{F} = \{\emptyset, \Omega\},$$

which contains only the empty set and the whole space.

Example 1.3. *Let $\Omega = \mathbb{R}$ with the Borel σ-algebra $\mathcal{B}(\mathbb{R})$, defined as the smallest σ-algebra containing all open intervals (a, b) with $a, b \in \mathbb{R}$. This σ-algebra is crucial in probability theory and measure theory.*

Example 1.4. *Let $\Omega = \{1, 2, 3, 4\}$. Consider the collection of subsets*

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

this $\mathcal{F}$ is a σ-algebra because it contains Ω and $\emptyset$, and is closed under complements and finite unions.

Example 1.5. *If $\Omega = \mathbb{R}$, the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ is the smallest σ-algebra containing all open intervals (a, b), $a, b \in \mathbb{R}$. It includes all open and closed sets, countable unions, and intersections of intervals, and is fundamental in defining probability measures on real variables.*

Example 1.6. *Given a set $A \subseteq \Omega$, the σ-algebra generated by A is*

$$\mathcal{F} = \{\emptyset, A, A^c, \Omega\},$$

which is the smallest σ-algebra containing A and its complement.

1.1.2 Probability

In this course, we restrict ourselves to studying countable sample spaces. The probability of an event is a numerical value that represents the proportion of times the event will occur when the experiment is repeated under identical conditions. From this definition, we can deduce that a probability must lie between 0 and 1, and that the probability of an event is the sum of the probabilities of each elementary outcome that composes it. Finally, the sum of the probabilities of all the elements of Ω is 1

Important. Recall that an event is nothing more than a subset of Ω. A probability associates to each event a number between 0 and 1. It is therefore a function from the power set of Ω, denoted $\mathcal{P}(\Omega)$, to $[0, 1]$.

Example 1.7. *Let $\Omega = \{0, 1, 2\}$. Let us construct $\mathcal{P}(\Omega)$:*

$$\mathcal{P}(\Omega) = \{\varnothing, \{0\}, \{1\}, \{2\}, \{0, 1\}, \{0, 2\}, \{1, 2\}, \Omega\}.$$

Definition 1.3. *A probability is a function on $\mathcal{P}(\Omega)$, the set of all subsets of Ω, such that:*

- $0 \leq P(A) \leq 1$, *for every event $A \subseteq \Omega$.*
- $P(A) = \sum_{\omega \in A} P(\omega)$, *for every event A.*
- $P(\Omega) = \sum_{\omega \in \Omega} P(\omega) = 1$.

What does it mean to say, an event A has probability?

- 0.95: A is very likely to occur.
- 0.03: A has very little chance of occurring.
- 4.0: incorrect.
- -2: incorrect.
- 0.4: A will occur in a bit less than half of the trials.
- 0.5: one chance out of two.
- 0: no chance that A occurs.

From the definition, we can easily deduce the following proposition, which is very useful for doing some calculations.

Proposition 1.1. *Let A and B be two events.*

1. *If A and B are mutually exclusive (incompatible), then*

$$P(A \cup B) = P(A) + P(B).$$

2. $P(A^c) = 1 - P(A)$.
3. $P(\varnothing) = 0$.
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof. 1. Immediate from the second point of the definition of a probability.

2. Since A and A^c are incompatible,

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c).$$

3. $P(\varnothing) = 1 - P(\varnothing^c) = 1 - P(\Omega) = 0$.

4. The technique is very often the same to compute the probability of a union of sets: we write this union as a union of incompatible sets and then use point (1). Here we write $A \cup B = A \cup (B \setminus A)$ and obtain

$$P(A \cup B) = P(A) + P(B \setminus A).$$

Then we write $B = (B \cap A) \cup (B \setminus A)$ to deduce

$$P(B) = P(B \cap A) + P(B \setminus A).$$

Combining these two equalities yields the formula in (4).

□

We now mention a more general definition of probability, valid for non-countable sample spaces.

Definition 1.4. *Let a random experiment be given and let Ω be the associated sample space. A probability on Ω is a function, defined on the collection of events, that satisfies:*

Axiom 1: $0 \leq P(A) \leq 1$, *for every event A.*

Axiom 2: *For every sequence of events $(A_i)_{i\in\mathbb{N}}$ pairwise incompatible (i.e, the events $(A_i)_{i\in\mathbb{N}}$ are pairwise incompatible if for all $i \neq j$, $A_i \cap A_j = \varnothing$),*

$$P\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} P(A_i).$$

Axiom 3: $P(\Omega) = 1$.

Example 1.8. *(uniform probability)*
Let Ω be a finite set. It sometimes happens, as when tossing a fair die, that all elementary events have the same probability. In this case, we say that the elementary events are equiprobable*. Let p be the probability of each elementary event. Then*

$$1 = P(\Omega) = \sum_{\omega\in\Omega} P(\omega) = \sum_{\omega\in\Omega} p = p \times \mathrm{card}(\Omega).$$

Hence

$$p = P(\omega) = \frac{1}{\mathrm{card}(\Omega)}, \quad \textit{for every } \omega \in \Omega.$$

The probability thus defined on Ω is called a uniform probability*. The probability of an event A is then easily computed:*

$$P(A) = \sum_{\omega\in A} P(\omega) = \frac{\mathrm{card}(A)}{\mathrm{card}(\Omega)}.$$

This formula is valid only when the elementary events are indeed equiprobable. In that case, it suffices to know how to compute the cardinality of the sets considered in order to compute the probabilities. From the definition, we can easily deduce the following proposition, which is very useful for doing some calculations.

1.2 Independence and conditioning

What is the probability of developing lung cancer?

Additional information: you smoke about twenty cigarettes per day. This information changes the probability. The tool that allows such an update is *conditional probability.*

Definition 1.5. *(Conditional Probability) Given two events A and B, with $P(A) > 0$, the probability of B conditional on A, or* given *A, denoted by $P(B \mid A)$, is defined by*

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

Equivalently, one may write

$$P(A \cap B) = P(B \mid A)\, P(A).$$

Moreover, the conditional probability given A, denoted by $P(\cdot \mid A)$, is itself a probability measure and therefore satisfies all the properties of a probability.

Example 1.9. *An urn contains r red balls and v green balls. Two balls are drawn one after the other,* without replacement*. What is the probability of drawing two red balls?*

Let Ω describe the outcomes of the experiment precisely:

$$\Omega = \{\textit{red}, \textit{green}\} \times \{\textit{red}, \textit{green}\}.$$

An elementary event is an ordered pair (x, y), where x is the color of the first ball drawn and y is the color of the second.

Let A be the event "the first ball is red" and B the event "the second ball is red". Then,

$$P(A \cap B) = P(B \mid A)\, P(A) = \frac{r-1}{r+v-1} \cdot \frac{r}{r+v}.$$

1.2.1 Law of Total Probability

Proposition 1.2. *Let A be an event such that $0 < P(A) < 1$. For any event B, we have*

$$P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c).$$

Proof. Since $A \cup A^c = \Omega$,

$$P(B) = P\Big(B \cap (A \cup A^c)\Big) = P\Big((B \cap A) \cup (B \cap A^c)\Big).$$

But $B \cap A$ and $B \cap A^c$ are incompatible. We deduce that

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

The definition of conditional probability then allows us to conclude. □

Example 1.10. *(See Example 1.9). What is the probability that the second ball drawn is red? We keep the same notation as before. Then*

$$P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c) = \frac{r-1}{r+v-1} \cdot \frac{r}{r+v} + \frac{r}{r+v-1} \cdot \frac{v}{r+v} = \frac{r}{r+v}.$$

Definition 1.6. *Let $(A_i)_{i\in I}$ be a family of events. We call it a* partition *of Ω if it satisfies:*

(i) $\bigcup_{i\in I} A_i = \Omega$.

(ii) The sets A_i are pairwise incompatible: for all $i \neq j$, $A_i \cap A_j = \varnothing$.

Proposition 1.3. *(Generalized law of total probability). Let $(A_i)_{i\in I}$ be a partition of Ω such that $P(A_i) > 0$ for all $i \in I$. Then, for every event B,*

$$P(B) = \sum_{i\in I} P(B \mid A_i)\, P(A_i).$$

Remark 1.1. *The law of total probability allows us to follow the stages of the random experiment in chronological order.*

We now introduce a formula that allows us to go backward in time.

Proposition 1.4. *(Bayes' formula). Let A and B be two events such that $0 < P(A) < 1$ and $P(B) > 0$. Then*

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B \mid A)\, P(A) + P(B \mid A^c)\, P(A^c)}.$$

Proof.

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

We conclude by replacing $P(B)$ using its expression given by the law of total probability. □

Proposition 1.5. *(Generalized Bayes' formula). Let $(A_i)_{i\in I}$ be a partition of Ω such that $P(A_i) > 0$ for all $i \in I$. Let B be an event with $P(B) > 0$. Then, for every $i \in I$,*

$$P(A_i \mid B) = \frac{P(B \mid A_i)\, P(A_i)}{\sum_{j\in I} P(B \mid A_j)\, P(A_j)}.$$

Example 1.11. *Two data-entry operators, A and B, input respectively 100 and 200 tables into a computer system. The tables entered by A contain errors in 5.2% of cases, and those entered by B in 6.7% of cases. A table is chosen at random. It contains errors. What is the probability that A entered this table? Define the events:*

$$T_A = \text{“the table was entered by A”}, \quad T_B = T_A^c = \text{“the table was entered by B”},$$

$$F = \text{“the table contains errors”}.$$

By Bayes' formula,

$$P(T_A \mid F) = \frac{P(F \mid T_A)\, P(T_A)}{P(F \mid T_A)\, P(T_A) + P(F \mid T_B)\, P(T_B)} = \frac{0.052 \times \frac{1}{3}}{0.052 \times \frac{1}{3} + 0.067 \times \frac{2}{3}} = 0.279.$$

Definition 1.7. *Let $(\Omega, \mathcal{T}, P)$ be a probability space. Two events $A, B \in \mathcal{T}$ are said to be independent if*

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

Let $(A_i)_{i\in I}$ be a family of events. These events are said to be mutually independent if, for every finite subset $S \subset I$,

$$P\left(\bigcap_{i\in S} A_i\right) = \prod_{i\in S} P(A_i).$$

Example 1.12. *For a family of three events $\{A_1, A_2, A_3\}$, $I = \{1, 2, 3\}$, these conditions are written as:*

$$\begin{aligned} S = \{1\} \quad & P(A_1) = P(A_1), \\ S = \{2\} \quad & P(A_2) = P(A_2), \\ S = \{3\} \quad & P(A_3) = P(A_3), \\ S = \{1,2\} \quad & P(A_1 \cap A_2) = P(A_1)P(A_2), \\ S = \{1,3\} \quad & P(A_1 \cap A_3) = P(A_1)P(A_3), \\ S = \{2,3\} \quad & P(A_2 \cap A_3) = P(A_2)P(A_3), \\ S = \{1,2,3\} \quad & P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3). \end{aligned}$$

Remark 1.2.

- *Thus, A and B are independent if knowing that one occurs does not change the probability of the other.*
- *Two incompatible events A and B with $P(A) > 0$ and $P(B) > 0$ are never independent. Indeed, $A \cap B = \varnothing$ implies $P(A \cap B) = 0 \neq P(A)P(B)$.*

1.2.2 Independent repetitions

When we study a random experiment that can be decomposed into several smaller, independent random experiments, the calculations are easy. Moreover, if each of these smaller random experiments has the uniform probability, then the overall experiment also has the uniform probability.

Proposition 1.6. *Let $\Omega = E \times F$, where E has cardinality n and F has cardinality p. Suppose we choose an element of E with the uniform probability, and independently an element of F, also with the uniform probability. Then each element $\omega = (x, y)$ of Ω has the same probability, namely*

$$P(\omega) = P\Big((x, y)\Big) = \frac{1}{\mathrm{card}(\Omega)} = \frac{1}{np} = P_E(\{x\})\, P_F(\{y\}).$$

Example 1.13. *We toss a fair coin and roll a fair die.*

$$\Omega = \{P, F\} \times \{1, \ldots, 6\}.$$

Since we have the uniform probability on $\{H, T\}$ and on $\{1, \ldots, 6\}$, we finally obtain the uniform probability on Ω, and

$$\forall \omega \in \Omega, \qquad P(\omega) = \frac{1}{\mathrm{card}(\Omega)} = \frac{1}{12}.$$

Proposition 1.7. *We repeat the same random experiment N times independently, modeled by a sample space Ω and a probability measure P. Then the new sample space is*

$$\Omega^N = \Omega \times \cdots \times \Omega,$$

and the associated probability measure is defined by

$$P^{\otimes N}(\omega_1, \ldots, \omega_N) = P(\omega_1) \cdots P(\omega_N).$$

In particular, if P is the uniform probability on Ω, then $P^{\otimes N}$ is the uniform probability on Ω^N.

Example 1.14. *(The Chevalier de Méré) The Chevalier de Méré observed that, when throwing three dice, he obtained the sum* 11 *more frequently than the sum* 12*. However, the number of combinations whose sum is* 12 *is the same as the number of combinations whose sum is* 11*. How can this be explained?*

Solution 1.1. *When throwing three fair dice, all outcomes are* equiprobable *at the level of* ordered triples*. There are* $6^3 = 216$ *possible outcomes.*

Let us list the combinations (unordered triples) whose sum is 11 *and* 12*.*

Sum 11***:***

$$(1,4,6), \quad (1,5,5), \quad (2,3,6), \quad (2,4,5), \quad (3,3,5), \quad (3,4,4).$$

Sum 12***:***

$$(1,5,6), \quad (2,4,6), \quad (2,5,5), \quad (3,3,6), \quad (3,4,5), \quad (4,4,4).$$

Thus, there are 6 *combinations for each sum. However, these combinations do not all have the same number of* permutations*.*

Counting permutations

- *If all three numbers are distinct, there are* $3! = 6$ *permutations.*
- *If two numbers are equal, there are* 3 *permutations.*
- *If all three numbers are equal, there is only* 1 *permutation.*

Sum 11***:***

$$\begin{array}{l} (1,4,6) \to 6 \\ (2,3,6) \to 6 \\ (2,4,5) \to 6 \\ (1,5,5) \to 3 \\ (3,3,5) \to 3 \\ (3,4,4) \to 3 \end{array} \quad \Rightarrow \quad 6+6+6+3+3+3 = 27$$

Sum 12***:***

$$\begin{array}{l} (1,5,6) \to 6 \\ (2,4,6) \to 6 \\ (3,4,5) \to 6 \\ (2,5,5) \to 3 \\ (3,3,6) \to 3 \\ (4,4,4) \to 1 \end{array} \quad \Rightarrow \quad 6+6+6+3+3+1 = 25$$

Conclusion

Although the number of combinations whose sum is 11 *equals the number whose sum is* 12*, the total number of favorable outcomes is greater for sum* 11 *than for sum* 12*. Therefore,*

$$\mathbb{P}(S = 11) = \frac{27}{216} > \frac{25}{216} = \mathbb{P}(S = 12).$$

This explains the observation of the Chevalier de Méré.

1.3 Exercises

Exercise 1.1. *Let P be a probability measure on a set Ω, and let A and B be two events. Assume that*

$$P(A \cup B) = \frac{7}{8}, \qquad P(A \cap B) = \frac{1}{4}, \qquad P(A) = \frac{3}{8}.$$

Compute $P(B)$, $P(A \cap B^c)$, and $P(B \cap A^c)$.

Solution 1.2. *We use the inclusion–exclusion formula:*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Substituting the given values,

$$\frac{7}{8} = \frac{3}{8} + P(B) - \frac{1}{4}.$$

Since $\frac{1}{4} = \frac{2}{8}$, we obtain

$$\frac{7}{8} = \frac{3}{8} + P(B) - \frac{2}{8} = \frac{1}{8} + P(B),$$

hence

$$P(B) = \frac{6}{8} = \frac{3}{4}.$$

Next,

$$P(A \cap B^c) = P(A) - P(A \cap B) = \frac{3}{8} - \frac{1}{4} = \frac{3}{8} - \frac{2}{8} = \frac{1}{8}.$$

Finally,

$$P(B \cap A^c) = P(B) - P(A \cap B) = \frac{3}{4} - \frac{1}{4} = \frac{1}{2}.$$

Conclusion:

$$P(B) = \frac{3}{4}, \qquad P(A \cap B^c) = \frac{1}{8}, \qquad P(B \cap A^c) = \frac{1}{2}.$$

Exercise 1.2. *Suppose that the faces of a die are biased in such a way that each odd number has the same probability of appearing, and this probability is twice that of each even number. The die is rolled. What is the probability of obtaining a number greater than or equal to 4?*

Solution 1.3. *Let p be the probability of each even number. Then each odd number has probability $2p$.*

Since a die has three odd numbers and three even numbers, we must have

$$3(2p) + 3(p) = 1.$$

Hence,

$$9p = 1 \quad \Longrightarrow \quad p = \frac{1}{9}.$$

Thus,

$$P(1) = P(3) = P(5) = \frac{2}{9}, \quad P(2) = P(4) = P(6) = \frac{1}{9}.$$

The event "number ≥ 4" corresponds to $\{4, 5, 6\}$, so

$$P(X \geq 4) = P(4) + P(5) + P(6) = \frac{1}{9} + \frac{2}{9} + \frac{1}{9} = \frac{4}{9}.$$

Conclusion: *The required probability is* $\boxed{\frac{4}{9}}$.

Exercise 1.3. *A parking lot contains twelve aligned parking spaces. Eight cars have parked at random, and it is observed that the four empty spaces are consecutive. Is this surprising?*

Solution 1.4. *There are* 12 *aligned parking spaces, of which* 8 *are occupied and* 4 *are empty. We assume that all possible arrangements of the* 4 *empty spaces among the* 12 *spaces are equally likely.*

The total number of possible configurations of the empty spaces is

$$\binom{12}{4}.$$

We now count the favorable configurations in which the 4 *empty spaces are* consecutive. *Since the block of* 4 *empty spaces can start at positions* $1, 2, \ldots, 9$, *there are* 9, *such configurations. Hence, the probability that the four empty spaces are consecutive is*

$$P = \frac{9}{\binom{12}{4}} = \frac{9}{495} = \frac{3}{165} \approx 0.0182.$$

Conclusion: *The probability is less than* 2%, *which is quite small. Therefore, observing four consecutive empty spaces is indeed* surprising.

Exercise 1.4. *Let M_1, M_2, and M_3 be three people. The first person, M_1, possesses some information encoded as either $+$ or $-$. She transmits it to the second person M_2, who then transmits it to M_3.*

Unfortunately, each time the information is transmitted, there is a probability p that the information is flipped to its opposite. Taking into account the fact that two changes restore the original message, what is the probability that M_3 receives the correct information?

If M_3 transmits the information he has to a fourth person M_4, what is the probability that M_4 receives the correct information?

When $p = 0.2$, what is the numerical value of this probability?

Solution 1.5. *At each transmission, the information is flipped with probability p and remains unchanged with probability $1 - p$.*

From M_1 to M_3:

The information received by M_3 is correct if either:

- *no transmission error occurs, or*
- *two transmission errors occur (since two flips restore the original message).*

Thus,

$$P(M_3 \text{ correct}) = (1-p)^2 + p^2.$$

From M_1 ***to*** M_4***:***

The information received by M_4 *is correct if an even number of flips occurs during the three transmissions. This happens when there are either* 0 *or* 2 *errors.*

Therefore,

$$P(M_4 \text{ correct}) = (1-p)^3 + \binom{3}{2} p^2(1-p) = (1-p)^3 + 3p^2(1-p).$$

Numerical value for $p = 0.2$***:***

$$\begin{aligned} P(M_4 \text{ correct}) &= (0.8)^3 + 3(0.2)^2(0.8) \\ &= 0.512 + 0.096 \\ &= 0.608. \end{aligned}$$

Conclusion:

$$\boxed{\begin{aligned} P(M_3 \text{ correct}) &= (1-p)^2 + p^2, \\ P(M_4 \text{ correct}) &= (1-p)^3 + 3p^2(1-p), \\ P(M_4 \text{ correct})\Big|_{p=0.2} &= 0.608. \end{aligned}}$$

Exercise 1.5. *1. In a class of* 36 *students, what is the probability that at least two students were born on the same day? (Assume that the year has* 365 *days and that all birthdays are independent and equally likely.)*

2. Generalize this result to a class of n *students. Plot the resulting probability.*

Solution 1.6. *1.* ***Class of 36 students***

Let us compute the probability that all *36 students have different birthdays. The first student may be born on any of the 365 days. The second student must avoid this day, the third must avoid the first two days, and so on.*

$$P(\text{all birthdays different}) = \prod_{k=0}^{35} \frac{365-k}{365}.$$

Therefore, the probability that at least two students share the same birthday *is*

$$P(\text{at least one coincidence}) = 1 - \prod_{k=0}^{35} \frac{365-k}{365}.$$

Numerically, this probability is greater than 0.8.

Conclusion: *In a class of 36 students, it is very likely that at least two students were born on the same day.*

2. ***General case: class of n students***

   For a class of n students ($n \leq 365$), the probability that all birthdays are different is

$$P_n(\textit{all different}) = \prod_{k=0}^{n-1} \frac{365-k}{365}.$$

   Hence, the probability that at least two students share a birthday is

$$P_n(\textit{at least one coincidence}) = 1 - \prod_{k=0}^{n-1} \frac{365-k}{365}.$$

   This probability increases rapidly with n, exceeds $1/2$ when $n = 23$, and approaches 1 as n becomes large. A plot of this function shows a sharp increase between $n = 20$ and $n = 40$.

Exercise 1.6. *In a factory, machine A produces 60% of the parts, of which 2% are defective. Machine B produces 30% of the parts, of which 3% are defective. Machine C produces 10% of the parts, of which 4% are defective.*

1. *A part is randomly selected from the production. What is the probability that it is defective?*

2. *A randomly selected part is defective. What is the probability that it was manufactured by machine A? by machine B? by machine C?*

Solution 1.7. *Let A, B, and C denote the events that a part is manufactured by machines A, B, and C, respectively, and let D denote the event that a part is defective.*

We are given:

$$P(A) = 0.6, \quad P(B) = 0.3, \quad P(C) = 0.1,$$
$$P(D \mid A) = 0.02, \quad P(D \mid B) = 0.03, \quad P(D \mid C) = 0.04.$$

1. ***Probability that a part is defective***

   By the law of total probability,

$$P(D) = P(D \mid A)P(A) + P(D \mid B)P(B) + P(D \mid C)P(C).$$

   Thus,

$$P(D) = 0.02(0.6) + 0.03(0.3) + 0.04(0.1) = 0.012 + 0.009 + 0.004 = 0.025.$$

   Conclusion: *The probability that a randomly selected part is defective is* $\boxed{0.025}$ *(i.e. 2.5%).*

2. ***Origin of a defective part***

   Using Bayes' theorem,

$$P(A \mid D) = \frac{P(D \mid A)P(A)}{P(D)} = \frac{0.02(0.6)}{0.025} = 0.48,$$

$$P(B \mid D) = \frac{P(D \mid B)P(B)}{P(D)} = \frac{0.03(0.3)}{0.025} = 0.36,$$
$$P(C \mid D) = \frac{P(D \mid C)P(C)}{P(D)} = \frac{0.04(0.1)}{0.025} = 0.16.$$

Conclusion: *Given that the part is defective,*

$$\boxed{P(A \mid D) = 0.48, \quad P(B \mid D) = 0.36, \quad P(C \mid D) = 0.16.}$$

Exercise 1.7. *In a garden center:*

- 25% *of the plants are less than one year old,*
- 60% *are between* 1 *and* 2 *years old,*
- 25% *have yellow flowers,*
- 60% *have pink flowers,*
- 15% *have yellow flowers and are less than one year old,*
- 3% *are more than two years old and have neither yellow nor pink flowers,*
- 15% *of the plants aged* 1 *to* 2 *years have yellow flowers,*
- 15% *of the plants aged* 1 *to* 2 *years have neither yellow nor pink flowers.*

Assume that flowers cannot be both yellow and pink. A plant is chosen at random from the garden center.

1. *What is the probability that it is less than one year old and has pink flowers?*
2. *What is the probability that it has pink flowers, given that it is more than two years old?*
3. *What is the probability that it is more than two years old and has yellow flowers?*

Solution 1.8. *Let us classify the plants according to their* age *and* flower color*. Denote:*

$Y =$ *"yellow flowers"*, $P =$ *"pink flowers"*, $N =$ *"neither yellow nor pink"*,
$A_1 =$ *"less than one year old"*,
$A_2 =$ *"between 1 and 2 years old"*,
$A_3 =$ *"more than two years old"*.

From the data:

$$P(A_1) = 0.25, \quad P(A_2) = 0.60, \quad P(A_3) = 0.15.$$

Information by age group

- ***Age*** < 1 ***year*** *(*A_1*):*

$$P(Y \cap A_1) = 0.15.$$

Hence,

$$P(P \cap A_1) = P(A_1) - P(Y \cap A_1) = 0.25 - 0.15 = 0.10.$$

- ***Age*** 1–2 ***years*** *(*A_2*):*

$$P(Y \mid A_2) = 0.15 \quad \Rightarrow \quad P(Y \cap A_2) = 0.15 \times 0.60 = 0.09,$$

$$P(N \mid A_2) = 0.15 \quad \Rightarrow \quad P(N \cap A_2) = 0.15 \times 0.60 = 0.09.$$

Thus,

$$P(P \cap A_2) = P(A_2) - 0.09 - 0.09 = 0.42.$$

- ***Age*** > 2 ***years*** *(*A_3*):*

$$P(N \cap A_3) = 0.03.$$

Since $P(A_3) = 0.15$*, we have*

$$P(Y \cap A_3) + P(P \cap A_3) = 0.12.$$

Color totals

$$P(Y) = 0.25, \quad P(P) = 0.60.$$

Thus,

$$P(Y \cap A_3) = 0.25 - (0.15 + 0.09) = 0.01,$$
$$P(P \cap A_3) = 0.60 - (0.10 + 0.42) = 0.08.$$

1. ***Less than one year old and pink***

$$P(P \cap A_1) = \boxed{0.10}.$$

2. ***Pink given more than two years old***

$$P(P \mid A_3) = \frac{P(P \cap A_3)}{P(A_3)} = \frac{0.08}{0.15} = \boxed{\frac{8}{15}}.$$

3. ***More than two years old and yellow***

$$P(Y \cap A_3) = \boxed{0.01}.$$

Chapter 2
Random Vectors

2.1 Random Variables

The notion of *randomness* is fundamental in probability theory and statistics. As discussed in the introductory notes on probability theory, statistics is concerned with quantifying and assessing the uncertainty associated with conclusions drawn from random samples of data.

Before studying random vectors [2, 5, 11, 14, 17] it is essential to recall the concept of a *random variable*. Using the basic notions of set theory and probability already introduced, we can now give a precise mathematical definition of randomness.

Definition 2.1 (Random Variable)**.** *A* random variable *(r.v.) X is a function defined on the sample space Ω that assigns a real number to each outcome of a random experiment. That is, to every elementary outcome $\omega \in \Omega$, the random variable associates a numerical value $X(\omega)$.*

Definition 2.2 (Discrete and Continuous Random Variables)**.** *A random variable is said to be* discrete *if the set of values it can take is finite or countably infinite.*

A random variable is said to be continuous *if the set of values it can take is uncountably infinite.*

2.1.1 Probability Mass and Density Functions

For discrete random variables, we can enumerate all of the possible realizations of the random variable, and associate a specific probability with each possible realization. In contrast, for continuous random variables, we cannot enumerate all of the possible realizations of the random variable, so it is impossible to associate a specific probability with each possible realization. As a result, we must define the probabilities of discrete and continuous random variable occurrences using distinct (but related) concepts. Another tool used to characterize the distribution of a random variable is the *cumulative distribution function.*

Definition 2.3. *Let X be a random variable. The cumulative distribution function (cdf) of X is the function $F : \mathbb{R} \to [0, 1]$ defined, for all $x \in \mathbb{R}$, by*

$$F(x) = \mathbb{P}(X \leq x).$$

Example 2.1. *Let X be the random variable representing the* number of Heads *obtained when a fair coin is tossed three times.*

Step 1: Sample space

Each toss has two possible outcomes, Head (H) or Tail (T). Therefore, the sample space contains $2^3 = 8$ *equally likely outcomes:*

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

Step 2: Definition of the random variable

For each outcome $\omega \in \Omega$*, the value* $X(\omega)$ *is the number of Heads appearing in* ω*. Thus, the possible values of* X *are*

$$X(\Omega) = \{0, 1, 2, 3\}.$$

Step 3: Probability distribution of X

We count the number of outcomes corresponding to each value of X*:*

- $X = 0$ *(no Heads):* $\{TTT\}$ *(1 outcome),*
- $X = 1$ *(one Head):* $\{HTT, THT, TTH\}$ *(3 outcomes),*
- $X = 2$ *(two Heads):* $\{HHT, HTH, THH\}$ *(3 outcomes),*
- $X = 3$ *(three Heads):* $\{HHH\}$ *(1 outcome).*

Since all outcomes are equally likely, the probability mass function of X *is*

$$\mathbb{P}(X = 0) = \frac{1}{8}, \qquad \mathbb{P}(X = 1) = \frac{3}{8}, \qquad \mathbb{P}(X = 2) = \frac{3}{8}, \qquad \mathbb{P}(X = 3) = \frac{1}{8}.$$

Step 4: Cumulative distribution function

The cumulative distribution function (c.d.f.) of X *is defined by*

$$F(x) = \mathbb{P}(X \leq x).$$

We compute $F(x)$ *by accumulating the probabilities of all values of* X *less than or equal to* x*:*

$$F(x) = \begin{cases} 0, & x < 0, \\ \mathbb{P}(X = 0) = \dfrac{1}{8}, & 0 \leq x < 1, \\ \mathbb{P}(X \leq 1) = \dfrac{1}{8} + \dfrac{3}{8} = \dfrac{4}{8}, & 1 \leq x < 2, \\ \mathbb{P}(X \leq 2) = \dfrac{4}{8} + \dfrac{3}{8} = \dfrac{7}{8}, & 2 \leq x < 3, \\ 1, & x \geq 3. \end{cases}$$

Remark 2.1. *The function* F *is a non-decreasing, right-continuous step function, with jumps at the integer values* $0, 1, 2,$ *and* 3*, which correspond to the possible values of the discrete random variable* X*.*

Remark 2.2. *Two random variables with the same probability distribution have the same cumulative distribution function.*

Proposition 2.1. *Let F be the cumulative distribution function of a random variable X. Then:*

1. *F is non-decreasing;*
2. *F is right-continuous and admits a left limit at every point x, equal to $\mathbb{P}(X < x)$;*
3. $$\lim_{x\to-\infty} F(x) = 0, \qquad \lim_{x\to+\infty} F(x) = 1.$$

For a discrete random variable, the cumulative distribution function is a step function, with a jump at each value k in the support of X, and the height of each jump is equal to the probability $\mathbb{P}(X = k)$.

Once the distribution of a random variable has been established, one can compute, as in descriptive statistics, a measure of location (the expectation) and a measure of dispersion (the variance).

2.1.2 Expectation and Variance of a Discrete Random Variable

Definition 2.4. *The expectation (or mean) of a discrete random variable X is the real number*

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

where the sum is taken over all values k that X can take.

A remarkable result allows us to compute easily the expectation of a function of X when the distribution of X is known; this is the *transfer theorem*.

Theorem 2.1 (Transfer Theorem)**.** *For any function g,*

$$\mathbb{E}[g(X)] = \sum_k g(k)\,\mathbb{P}(X = k).$$

Proof. Observe that $g(X) = y$ if and only if $X = x$ with $g(x) = y$. Hence,

$$\mathbb{P}(g(X) = y) = \sum_{x:g(x)=y} \mathbb{P}(X = x).$$

Multiplying this equality by y and summing over all y, we obtain

$$\mathbb{E}[g(X)] = \sum_y y\,\mathbb{P}(g(X) = y) = \sum_y \sum_{x:g(x)=y} g(x)\,\mathbb{P}(X = x) = \sum_x g(x)\,\mathbb{P}(X = x).$$

□

Definition 2.5. *The variance of a discrete random variable X is the nonnegative real number*

$$\mathrm{Var}(X) = \mathbb{E}\Big[(X - \mathbb{E}[X])^2\Big] = \sum_k (k - \mathbb{E}[X])^2\,\mathbb{P}(X = k) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

The standard deviation of X is the square root of its variance.

Remark 2.3. *The expectation does not always exist when the set $X(\Omega)$ is infinite. In this case, X has an expectation if*

$$\sum_{k\in X(\Omega)} |k|\,\mathbb{P}(X=k) < +\infty.$$

The expectation of a random variable is the average of the values that X can take, weighted by their probabilities. It is often simply called the *mean* of X and represents a central value around which the possible values of X are distributed. The standard deviation (or variance) measures the dispersion of the random variable X around its mean $\mathbb{E}[X]$.

The expectation and the variance of a random variable depend on X only through its probability distribution: two random variables with the same distribution have the same expectation and the same variance.

Example 2.2. *We consider again the random variable X, the number of Heads obtained when a coin is tossed three times. Its distribution is known. We compute:*

$$\mathbb{E}[X] = \sum_{k=0}^{3} k\,\mathbb{P}(X=k) = 3\cdot\frac{1}{8} + 2\cdot\frac{3}{8} + 1\cdot\frac{3}{8} + 0\cdot\frac{1}{8} = \frac{12}{8} = \frac{3}{2}.$$

Moreover,

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sum_{k=0}^{3} k^2\,\mathbb{P}(X=k) - \left(\mathbb{E}[X]\right)^2$$

$$= 3^2\cdot\frac{1}{8} + 2^2\cdot\frac{3}{8} + 1^2\cdot\frac{3}{8} + 0^2\cdot\frac{1}{8} - \left(\frac{3}{2}\right)^2 = \frac{3}{4}.$$

2.2 Independence and Conditioning

2.2.1 Independence of Events and Random Variables

Definition 2.6. *Two random variables $X, Y : \Omega \to \mathbb{R}$ are independent if, for all Borel sets $A, B \subset \mathbb{R}$, the events $(X \in A)$ and $(Y \in B)$ are independent:*

$$P\Big((X\in A)\cap(Y\in B)\Big) = P(X\in A)P(Y\in B).$$

Let $(X_i)_{i\in I}$ be a family of random variables. It is said to be independent if, for any finite subset $S \subset I$ and any choice of Borel sets $(A_i)_{i\in S}$ in $\mathbb{R}$,

$$P\left(\bigcap_{i\in S}(X_i\in A_i)\right) = \prod_{i\in S} P(X_i\in A_i).$$

To simplify notation, we set:

$$(X\in A, Y\in B) := (X\in A)\cap(Y\in B) = \{\omega\in\Omega \mid X(\omega)\in A \text{ and } Y(\omega)\in B\},$$

$$\Big(X_i\in A_i,\, i\in S\Big) := \bigcap_{i\in S}(X_i\in A_i),$$

$$(X_1\in A_1, X_2\in A_2, \ldots, X_n\in A_n) := (X_1\in A_1)\cap(X_2\in A_2)\cap\cdots\cap(X_n\in A_n).$$

Definition 2.7. *Two random variables X and Y are said to be* independent *if, for all i and j, the events $\{X = i\}$ and $\{Y = j\}$ are independent, that is,*

$$\mathbb{P}(X = i,\ Y = j) = \mathbb{P}(X = i)\,\mathbb{P}(Y = j).$$

Remark 2.4. *If X and Y are not independent, knowing the distribution of X and that of Y is not sufficient to determine the joint distribution of (X, Y), which is given, for all i and j, by*

$$\mathbb{P}((X, Y) = (i, j)) = \mathbb{P}(X = i,\ Y = j).$$

Proposition 2.2 (Law of Total Probability)**.** *Let X and Y be two random variables. For any $i \in X(\Omega)$,*

$$\mathbb{P}(X = i) = \sum_{j \in Y(\Omega)} \mathbb{P}(X = i \mid Y = j)\,\mathbb{P}(Y = j).$$

Proof. The events $\{Y = j\}$, for $j \in Y(\Omega)$, form a partition of the sample space Ω. The result follows directly from the law of total probability. □

Example 2.3. *Two fair dice are rolled, and the random variables X and Y denote the two numbers obtained. Let $Z = X + Y$. What is the probability distribution of Z?*

Theorem 2.2. *1. For any random variables X and Y,*

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

2. If X and Y are independent,

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Proof. For the first point, observe that

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

so

$$\begin{aligned} \mathbb{E}[X + Y] &= \sum_{x,y} (x + y)\,\mathbb{P}(X = x, Y = y) \\ &= \sum_{x,y} x\,\mathbb{P}(X = x, Y = y) + \sum_{x,y} y\,\mathbb{P}(X = x, Y = y) \\ &= \sum_x x\,\mathbb{P}(X = x) + \sum_y y\,\mathbb{P}(Y = y) = \mathbb{E}[X] + \mathbb{E}[Y]. \end{aligned} \tag{2.1}$$

For the second point, we first show that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ when X and Y are independent:

$$\begin{aligned} \mathbb{E}[XY] &= \sum_{x,y} xy\,\mathbb{P}(X = x, Y = y) \\ &= \sum_{x,y} xy\,\mathbb{P}(X = x)\,\mathbb{P}(Y = y) = \left(\sum_x x\,\mathbb{P}(X = x)\right)\left(\sum_y y\,\mathbb{P}(Y = y)\right) \\ &= \mathbb{E}[X]\mathbb{E}[Y]. \end{aligned} \tag{2.2}$$

Then, using the formula $\mathrm{Var}(X + Y) = \mathbb{E}[(X + Y)^2] - \mathbb{E}[X + Y]^2$, the result follows easily. □

2.3 Discrete probability laws

We are interested here in whether or not an event occurs. In other words, we study random experiments that have only two possible outcomes (for example: a patient in a hospital survives or not, a client signs a contract or not, a voter votes Democrat or Republican, etc.).

Consider a random experiment of this type. It is called a *Bernoulli trial*. The trial results in a *success* if the event of interest occurs, and in a *failure* otherwise. We associate with this trial a random variable Y which takes the value

$$Y = \begin{cases} 1, & \text{if the event occurs,} \\ 0, & \text{otherwise.} \end{cases}$$

Thus, this random variable takes only two values (0 and 1), and its probability distribution is given by

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

We then say that Y follows a *Bernoulli distribution* with parameter p, denoted by $B(p)$. The random variable Y has expectation p and variance $p(1-p)$. Indeed,

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

and

$$\mathrm{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \mathbb{E}[Y] - (\mathbb{E}[Y])^2 = p(1-p).$$

A *Bernoulli scheme* consists of repeating the same Bernoulli trial n times under identical conditions.

2.3.1 Bernoulli Scheme

1. Each trial has two possible outcomes: success (S) or failure (F).
2. For each trial, the probability of success is the same, denoted by

$$\mathbb{P}(S) = p, \qquad \mathbb{P}(F) = q = 1 - p.$$

3. The n trials are independent: the probability of success does not change and does not depend on the outcomes of the other trials.

2.3.2 Binomial Distribution

Let X be the random variable representing the number of successes obtained in n trials of a Bernoulli scheme. Then X is said to follow a *binomial distribution* with parameters (n, p), denoted by $B(n, p)$. This distribution is given by

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad \text{for all } 0 \le k \le n,$$

where

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

Indeed, the number of successes is necessarily an integer between 0 and the number of trials n. Moreover, the event $\{X = k\}$ is a union of elementary events, each of which is a sequence of length n of the form

$$\omega = [SS \cdots SE \cdots E],$$

with k successes (S) and $n - k$ failures (F). Each such elementary event has probability

$$\mathbb{P}(\omega) = p^k (1-p)^{n-k}.$$

How many sequences of length n contain k successes and $n - k$ failures? There are $\binom{n}{k}$ such sequences, corresponding to the number of ways to choose k successes among n trials. Therefore,

$$\mathbb{P}(X = k) = \sum_{\omega : X(\omega) = k} \mathbb{P}(\omega) = \operatorname{card}\{\omega : X(\omega) = k\}\, p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}.$$

Proposition 2.3. *Let X be a random variable with distribution $B(n, p)$. Then the expectation and variance of X are given by*

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

Example 2.4 (Sampling with Replacement)**.** *Consider a population consisting of two categories of individuals, A and B. We select an individual uniformly at random from the population and record the outcome of the trial, that is, whether the individual belongs to category A or B. The individual is then replaced in the population, and the experiment is repeated. This procedure defines a Bernoulli scheme.*

Suppose that there are N_A individuals in category A within a population of size N. For each Bernoulli trial, the probability of selecting an individual from category A (which we call a success) is

$$p = \frac{N_A}{N}.$$

The number of individuals in the sample that belong to category A is a random variable, since its value depends on the random selection of the sample.

Let X denote the number of individuals from category A in the sample. According to the above discussion, we are dealing with a Bernoulli scheme in which $p = \frac{N_A}{N}$ is the probability of success and n is the number of trials. One trial corresponds to drawing one individual, and a success corresponds to the event "the individual belongs to category A." Hence, X follows a binomial distribution $B\left(n, \frac{N_A}{N}\right)$.

Example 2.5. *We study the infection of trees in a forest by a parasite. Let p denote the proportion of infected trees. We examine 4 trees. If a tree is infected, we say that we have a success; otherwise, we have a failure. Let X be the number of infected trees among the 4 trees. Then X follows a binomial distribution $B(4, p)$.*

The probability distribution of X is given by

$$\mathbb{P}(X = 0) = \binom{4}{0} q^4 = q^4,$$

$$\mathbb{P}(X = 1) = \binom{4}{1} p q^3 = 4 p q^3,$$

$$\mathbb{P}(X=2)=\binom{4}{2}p^2q^2=6p^2q^2,$$

$$\mathbb{P}(X=3)=\binom{4}{3}p^3q=4p^3q,$$

$$\mathbb{P}(X=4)=\binom{4}{4}p^4=p^4,$$

where $q=1-p$.

For $p=\frac{1}{5}$, the probabilities are summarized in Table 2.1.

Table 2.1: Binomial distribution for $n=4$ and $p=\frac{1}{5}$

Values of X	0	1	2	3	4
Probabilities	0.4096	0.4096	0.1536	0.0256	0.0016

Remark 2.5. *Associate with each Bernoulli trial a random variable* Y_i $(1\le i\le n)$ *such that*

$$Y_i=\begin{cases}1, & \text{if a success is observed in the } i\text{-th trial},\\ 0, & \text{otherwise.}\end{cases}$$

Then the total number of successes, denoted by X*, satisfies*

$$X=\sum_{i=1}^{n}Y_i.$$

In other words, a random variable with distribution $B(n,p)$ *is the sum of* n *independent Bernoulli random variables with distribution* $B(p)$.

Example 2.6. *A new treatment solves the fragility problem of a material in* 50% *of cases. If the treatment is tested on* 15 *objects, what is:*

- *the probability that at most* 6 *objects are resistant,*
- *the probability that the number of resistant objects is between* 6 *and* 10,
- *the probability that at most* 2 *objects remain fragile?*

What is the average number of objects made resistant by the treatment?

We assume that the results concerning the 15 objects are independent. Let X be the number of resistant objects after the treatment. Then X follows a binomial distribution $B(15,0.5)$.

$$\mathbb{P}(X\le 6)=\sum_{k=0}^{6}\mathbb{P}(X=k)=\frac{1}{2^{15}}\sum_{k=0}^{6}\binom{15}{k}$$

$$=\frac{1}{2^{15}}\left(\binom{15}{0}+\binom{15}{1}+\binom{15}{2}+\binom{15}{3}+\binom{15}{4}+\binom{15}{5}+\binom{15}{6}\right)$$

$$=\frac{1}{2^{15}}(1+15+105+455+1365+3003+5005)=0.304.$$

$$\mathbb{P}(6 \leq X \leq 10) = \sum_{k=6}^{10} \mathbb{P}(X = k) = 0.790.$$

At most 2 objects remain fragile means that at least 13 objects are resistant:

$$\mathbb{P}(X \geq 13) = \mathbb{P}(X = 13) + \mathbb{P}(X = 14) + \mathbb{P}(X = 15)$$

$$= \frac{1}{2^{15}} \left(\binom{15}{13} + \binom{15}{14} + \binom{15}{15} \right) = \frac{455 + 15 + 1}{2^{15}} = 0.018.$$

Finally, the expected number of resistant objects is

$$\mathbb{E}[X] = np = 15 \times 0.5 = 7.5.$$

2.3.3 Geometric Distribution

Instead of performing a fixed number of trials in a Bernoulli scheme, the experimenter stops at the first success. The quantity of interest is the number of trials required until the first success is obtained. The number of successes is fixed at 1, but the total number of trials Y is random and can take any integer value greater than or equal to 1. For all $k \geq 1$,

$$\mathbb{P}(Y = k) = p(1-p)^{k-1},$$

where p is the probability of success in each Bernoulli trial.

We say that Y follows a *geometric distribution* with parameter p, denoted by $G(p)$.

Proposition 2.4. *Let Y be a random variable with distribution $G(p)$. Then*

$$\mathbb{E}[Y] = \frac{1}{p}, \qquad \mathrm{Var}(Y) = \frac{1-p}{p^2}.$$

Proof. First, recall that for $x \in [0, 1[$,

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}, \qquad \sum_{k=1}^{\infty} k x^{k-1} = \frac{1}{(1-x)^2}.$$

Using these identities with $x = 1 - p$ yields the desired results. □

and

$$\left(\sum_{k=0}^{\infty} x^k \right)' = \left(\frac{1}{1-x} \right)' = \frac{1}{(1-x)^2}.$$

Hence, for $x = 1 - p$,

$$\mathbb{E}[Y] = \sum_{k=1}^{\infty} k\mathbb{P}(Y = k) = p \sum_{k=1}^{\infty} k(1-p)^{k-1} = p \cdot \frac{1}{p^2} = \frac{1}{p}.$$

A similar computation allows us to determine the variance (exercise).

2.3.4 Poisson Distribution

The Poisson distribution is an approximation of the binomial distribution when n is large and np is small (in practice, $n \geq 50$ and $np \leq 10$). A random variable X following a Poisson distribution with parameter λ, denoted by $\mathcal{P}(\lambda)$, satisfies

$$\forall k \in \mathbb{N}, \qquad \mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

The expectation and the variance of X are both equal to λ.
The Poisson distribution is commonly used to model:

- the number of tasks arriving at a computer server per minute,
- the number of red blood cells in one milliliter of blood,
- the number of workplace accidents in a company during one year.

In the case of approximating the binomial distribution by the Poisson distribution, the parameter of the Poisson distribution is $\lambda = np$.

2.3.5 Uniform Distribution

Apart from the prestige associated with its name, the uniform distribution represents a situation of complete lack of information. Suppose that a random variable X takes the values $1, 2, \ldots, n$, but that we have no information about its probability distribution. In this case, after justification, we assign the same probability to each value:

$$\forall k = 1, \ldots, n, \qquad \mathbb{P}(X = k) = \frac{1}{n}.$$

It can be easily shown that

$$\mathbb{E}[X] = \frac{n+1}{2}, \qquad \mathrm{Var}(X) = \frac{(n+1)(n-1)}{12}.$$

2.4 Exercises

Exercise 2.1. *Let X be a random variable whose probability distribution is given by*

$$\mathbb{P}(X = -1) = 0.2, \quad \mathbb{P}(X = 0) = 0.1, \quad \mathbb{P}(X = 4) = 0.3, \quad \mathbb{P}(X = 5) = 0.4.$$

Compute $\mathbb{P}(X \leq 3)$, $\mathbb{P}(X > 2)$, the expectation, and the variance of X.

Solution 2.1. *The random variable X takes the values $-1, 0, 4$, and 5 with probabilities*

$$\mathbb{P}(X = -1) = 0.2, \quad \mathbb{P}(X = 0) = 0.1, \quad \mathbb{P}(X = 4) = 0.3, \quad \mathbb{P}(X = 5) = 0.4.$$

1. Computation of $\mathbb{P}(X \leq 3)$
The values of X less than or equal to 3 are -1 and 0. Hence,

$$\mathbb{P}(X \leq 3) = \mathbb{P}(X = -1) + \mathbb{P}(X = 0) = 0.2 + 0.1 = 0.3.$$

2. Computation of $\mathbb{P}(X > 2)$
The values of X *strictly greater than* 2 *are* 4 *and* 5*. Thus,*

$$\mathbb{P}(X > 2) = \mathbb{P}(X = 4) + \mathbb{P}(X = 5) = 0.3 + 0.4 = 0.7.$$

3. Expectation of X
The expectation of X *is given by*

$$\mathbb{E}(X) = \sum_x x\,\mathbb{P}(X = x).$$

Therefore,

$$\mathbb{E}(X) = (-1)(0.2) + (0)(0.1) + (4)(0.3) + (5)(0.4) = -0.2 + 0 + 1.2 + 2 = 3.$$

4. Variance of X
First compute $\mathbb{E}(X^2)$*:*

$$\mathbb{E}(X^2) = (-1)^2(0.2) + 0^2(0.1) + 4^2(0.3) + 5^2(0.4) = 0.2 + 0 + 4.8 + 10 = 15.$$

The variance is

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 15 - 3^2 = 6.$$

Exercise 2.2. *Two dice are rolled.*

- *Model the experiment.*
- *What is the probability of obtaining at least one* 4*?*
- *What is the probability that the smaller die shows a number less than or equal to* 4*?*
- *What is the expectation of the sum of the two dice?*

Solution 2.2. *When two dice are rolled, the sample space is*

$$\Omega = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\},$$

which contains 36 *equally likely outcomes.*

2. Probability of obtaining at least one 4
Let A *be the event "at least one die shows* 4*". Its complement* A^c *is the event "no die shows* 4*". Then,*

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \left(\frac{5}{6}\right)^2 = 1 - \frac{25}{36} = \frac{11}{36}.$$

3. Probability that the smaller die is less than or equal to 4
Let $M = \min(X_1, X_2)$*, where* X_1 *and* X_2 *are the results of the two dice. Then,*

$$\mathbb{P}(M \leq 4) = 1 - \mathbb{P}(M \geq 5).$$

The event $M \geq 5$ *occurs only if both dice show* 5 *or* 6*, hence*

$$\mathbb{P}(M \geq 5) = \left(\frac{2}{6}\right)^2 = \frac{1}{9}.$$

Thus,

$$\mathbb{P}(M \leq 4) = 1 - \frac{1}{9} = \frac{8}{9}.$$

4. Expectation of the sum of the two dice

Let $S = X_1 + X_2$. Since the dice are independent and identically distributed,

$$\mathbb{E}(S) = \mathbb{E}(X_1) + \mathbb{E}(X_2).$$

For a fair die,

$$\mathbb{E}(X_1) = \mathbb{E}(X_2) = \frac{1+2+3+4+5+6}{6} = 3.5.$$

Therefore,

$$\mathbb{E}(S) = 3.5 + 3.5 = 7.$$

Conclusion:

$$\boxed{\mathbb{P}(\textit{at least one } 4) = \frac{11}{36}, \quad \mathbb{P}(\min \leq 4) = \frac{8}{9}, \quad \mathbb{E}(S) = 7.}$$

Exercise 2.3. *Assume that the number of accidents occurring daily on a highway is a random variable following a Poisson distribution with parameter $\lambda = 3$.*

- *Compute $\mathbb{P}(X = k)$ for $k = 0, \ldots, 6$.*
- *Provide a graphical representation of the distribution.*
- *What is the probability that at least two accidents occur on a given day?*

Solution 2.3. *Let X denote the number of accidents occurring daily on the highway. We are given that $X \sim Poisson(\lambda = 3)$.*

1. *The probability mass function of a Poisson random variable is*

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

For $\lambda = 3$, we compute

$$\begin{aligned}
\mathbb{P}(X = 0) &= \frac{3^0 e^{-3}}{0!} = e^{-3} \approx 0.0498, \\
\mathbb{P}(X = 1) &= \frac{3^1 e^{-3}}{1!} = 3e^{-3} \approx 0.1494, \\
\mathbb{P}(X = 2) &= \frac{3^2 e^{-3}}{2!} = \frac{9e^{-3}}{2} \approx 0.2240, \\
\mathbb{P}(X = 3) &= \frac{3^3 e^{-3}}{3!} = \frac{27e^{-3}}{6} \approx 0.2240, \\
\mathbb{P}(X = 4) &= \frac{3^4 e^{-3}}{4!} = \frac{81e^{-3}}{24} \approx 0.1680, \\
\mathbb{P}(X = 5) &= \frac{3^5 e^{-3}}{5!} = \frac{243e^{-3}}{120} \approx 0.1008, \\
\mathbb{P}(X = 6) &= \frac{3^6 e^{-3}}{6!} = \frac{729e^{-3}}{720} \approx 0.0504.
\end{aligned}$$

2. *A graphical representation can be made using a bar plot of* $\mathbb{P}(X = k)$ *for* $k = 0, \ldots, 6$.

3. *The probability that at least two accidents occur is*

$$\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X < 2) = 1 - \Big(\mathbb{P}(X = 0) + \mathbb{P}(X = 1)\Big) = 1 - (0.0498 + 0.1494) \approx 0.8008.$$

Exercise 2.4. *For each of the following random experiments:*

- *Rolling a single die.*
- *Number of heads in three coin tosses*
- *Number of accidents per day (Poisson process*

1. *define the sample space* Ω *and the associated probability measure,*
2. *identify the random variable under study,*
3. *specify the name of its distribution, its parameters, and give the expression of* $\mathbb{P}(X = k)$.

Solution 2.4. *For each random experiment, we proceed as follows:*

1. ***Rolling a single die***

   - *Sample space:* $\Omega = \{1, 2, 3, 4, 5, 6\}$ *Probability measure:* $\mathbb{P}(\{i\}) = \frac{1}{6}$ *for* $i = 1, \ldots, 6$.
   - *Random variable:* X *= number obtained on the die.*
   - *Distribution:* X *follows a* ***discrete uniform distribution*** *on* $\{1, \ldots, 6\}$*, with*
   $$\mathbb{P}(X = k) = \frac{1}{6}, \quad k = 1, 2, \ldots, 6.$$

2. ***Number of heads in three coin tosses***

   - *Sample space:* $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ *Probability measure: Each outcome has probability* $\frac{1}{8}$.
   - *Random variable:* X *= number of heads observed.*
   - *Distribution:* X *follows a* ***Binomial distribution*** $B(n = 3, p = 0.5)$*:*
   $$\mathbb{P}(X = k) = \binom{3}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{3-k}, \quad k = 0, 1, 2, 3.$$

3. ***Number of accidents per day (Poisson process)***

   - *Sample space:* $\Omega = \{0, 1, 2, 3, \ldots\}$ *(number of accidents). Probability measure:* $\mathbb{P}(X = k)$ *as given by the Poisson formula.*
   - *Random variable:* X *= number of accidents in one day.*
   - *Distribution:* X *follows a* ***Poisson distribution*** *with parameter* $\lambda > 0$*:*
   $$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

4. ***Example 4: Drawing balls from an urn***

- *Sample space: depends on the experiment, e.g.,* $\Omega = \{white, black\}$ *for one draw. Probability measure:* $\mathbb{P}(white) = p$, $\mathbb{P}(black) = 1 - p$.
- *Random variable:* X *= number of white balls drawn.*
- *Distribution: If drawing* n *balls with replacement,* X *follows a* ***Binomial distribution*** $B(n, p)$*:*

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

Exercise 2.5. *a) A die is rolled* 20 *times. What is the distribution of the number of times the outcome* 5 *is obtained? What is the probability of obtaining fewer than* 3 *occurrences of* 5*?*

b) An urn contains one white ball and one black ball. A ball is drawn at random, then replaced, and an additional ball of the same color is added to the urn. What is the distribution of the number of white balls in the urn?

c) The experiment is repeated. What is the new distribution of the number of white balls in the urn? Also give the distribution of the number of black balls and its expectation.

d) On the side of the A7 highway, a student is hitchhiking. During this season, one out of twenty drivers stops to pick up a hitchhiker. What is the distribution of the number of vehicles the student observes before finding a driver? What is the probability that the student gets into the fourth passing car? What is the probability that at least six cars pass without stopping?

Solution 2.5. *a) Let* X *be the number of times a* 5 *occurs in* 20 *rolls of a die. Since each roll is independent and the probability of obtaining a* 5 *is* $p = \frac{1}{6}$*,* X *follows a* ***Binomial distribution****:*

$$X \sim B(n = 20, p = \frac{1}{6}).$$

The probability of obtaining fewer than 3 *occurrences of* 5 *is*

$$\mathbb{P}(X < 3) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2),$$

where

$$\mathbb{P}(X = k) = \binom{20}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{20-k}, \quad k = 0, 1, 2.$$

b) Initially, the urn contains one white and one black ball. Let Y *be the number of white balls after one draw with replacement and adding another ball of the same color. Then the number of white balls can be:*

$$\mathbb{P}(Y = 1) = \frac{1}{2} \quad \textit{(if black was drawn)}, \qquad \mathbb{P}(Y = 2) = \frac{1}{2} \quad \textit{(if white was drawn)}.$$

c) If the experiment is repeated, let Z be the number of white balls after the second draw. Using conditional probabilities:

$$\begin{aligned}
\mathbb{P}(Z=1) &= \mathbb{P}(Y=1)\cdot\frac{1}{2} = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4},\\
\mathbb{P}(Z=2) &= \mathbb{P}(Y=1)\cdot\frac{1}{2} + \mathbb{P}(Y=2)\cdot\frac{1}{3} = \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{3} = \frac{5}{12},\\
\mathbb{P}(Z=3) &= \mathbb{P}(Y=2)\cdot\frac{2}{3} = \frac{1}{2}\cdot\frac{2}{3} = \frac{1}{3}.
\end{aligned}$$

Similarly, the distribution of the number of black balls W is:

$$\mathbb{P}(W=1) = \frac{5}{12}, \quad \mathbb{P}(W=2) = \frac{1}{3}, \quad \mathbb{P}(W=3) = \frac{1}{4}.$$

The expected number of white balls is

$$\mathbb{E}[Z] = 1\cdot\frac{1}{4} + 2\cdot\frac{5}{12} + 3\cdot\frac{1}{3} = \frac{1}{4} + \frac{10}{12} + 1 = \frac{31}{12} \approx 2.583.$$

*d) Let X be the number of cars observed before a driver stops. The success probability is $p = \frac{1}{20}$. Then X follows a **geometric distribution**:*

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

- Probability that the student gets into the fourth car:

$$\mathbb{P}(X=4) = (1-p)^3 p = \left(\frac{19}{20}\right)^3 \cdot \frac{1}{20}.$$

- Probability that at least six cars pass without stopping:

$$\mathbb{P}(X \geq 6) = (1-p)^5 = \left(\frac{19}{20}\right)^5.$$

Exercise 2.6. *Let X and Y be two independent random variables with respective distributions $B(n,p)$ and $B(n',p)$. What is the distribution of the sum $X+Y$?*

Solution 2.6. *Let $X \sim B(n,p)$ and $Y \sim B(n',p)$ be two independent binomial random variables.*

The sum

$$S = X + Y$$

*also follows a **binomial distribution** because the sum of independent binomial random variables with the same success probability is again binomial with the total number of trials equal to the sum of the individual trials.*

Hence,

$$S \sim B(n+n',p).$$

The probability mass function of S is

$$\begin{aligned}
\mathbb{P}(S=k) &= \sum_{i=0}^{k} \mathbb{P}(X=i)\mathbb{P}(Y=k-i) \qquad (2.3)\\
&= \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \binom{n'}{k-i} p^{k-i} (1-p)^{n'-(k-i)}\\
&= \binom{n+n'}{k} p^k (1-p)^{n+n'-k}.
\end{aligned}$$

2.5 Standard continuous probability distributions

Discrete random variables are used to count events that occur randomly, whereas continuous random variables are used to measure *continuous* quantities (distance, mass, pressure, etc.). A probability density function f describes the distribution of a continuous random variable X in the following sense:

$$\forall a, b \in \mathbb{R}, \qquad \mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\, dx,$$

and

$$\forall x \in \mathbb{R}, \qquad F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X < x) = \int_{-\infty}^{x} f(u)\, du,$$

where F denotes the cumulative distribution function of X.

It follows that a probability density function must satisfy

$$\forall x \in \mathbb{R}, \qquad f(x) \geq 0, \qquad \text{and} \qquad \int_{\mathbb{R}} f(x)\, dx = 1.$$

Definition 2.8. *A* probability density function *is any nonnegative real-valued function whose integral over* $\mathbb{R}$ *is equal to* 1*.*

Remark 2.6. *1. For a continuous random variable* X*, the density function* f *does* not *represent the probability of the event* $\{X = x\}$*, since*

$$\mathbb{P}(X = x) = 0.$$

Instead, one should keep in mind the approximation

$$\mathbb{P}(x \leq X \leq x + \Delta x) \approx f(x)\, \Delta x,$$

for small Δx*.*

2. The distribution of a random variable X *can be characterized by:*

- *its probability density function* f*,*
- *the probabilities* $\mathbb{P}(a \leq X \leq b)$ *for all* $a, b \in \mathbb{R}$*,*
- *the cumulative distribution function* $F(x) = \mathbb{P}(X \leq x)$ *for all* $x \in \mathbb{R}$*.*

3. For a continuous random variable X*, one has*

$$\mathbb{P}(X = x) = 0 \quad \text{for all } x \in \mathbb{R}.$$

Proposition 2.5. *Let* X *be a continuous random variable with probability density function* f *and cumulative distribution function* F*. Then:*

- F *is continuous and non-decreasing,*
- F *is differentiable at every point where* f *is continuous, and*

$$F'(x) = f(x),$$

- *for all* $a \leq b$*,*

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

2.5.1 Expectation and Variance of Continuous Random Variables

Definition 2.9. *The expectation of a continuous random variable X is defined by*

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

whenever this integral exists. Moreover, the variance of X is given by

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

Proposition 2.6. *1. The expectation of a function $Y = \varphi(X)$ is given by*

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

2. For all real numbers a and b,

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

3. If X and Y are two continuous random variables, then

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

If, in addition, X and Y are independent, then

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Example 2.7. *Let X be a random variable with probability density function f defined by*

$$f(x) = \begin{cases} \dfrac{1}{2}, & \text{if } x \in [0, 2], \\ 0, & \text{otherwise.} \end{cases}$$

This is indeed a probability density function, since $f(x) \geq 0$ and

$$\int_{\mathbb{R}} f(x)\, dx = 1.$$

Compute $\mathbb{P}\left(\frac{1}{3} \leq X \leq \frac{2}{3}\right)$. Determine $\mathbb{E}[X]$ and $\mathrm{Var}(X)$.

Solution 2.7. *We are given a random variable X with probability density function*

$$f(x) = \begin{cases} \dfrac{1}{2}, & x \in [0, 2], \\ 0, & \text{otherwise.} \end{cases}$$

1. ***Verification of a valid PDF:*** *Clearly, $f(x) \geq 0$ for all x. Moreover,*

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_0^2 \frac{1}{2}\, dx = \frac{1}{2} \cdot 2 = 1,$$

so $f(x)$ is indeed a valid probability density function.

2. ***Probability computation:***

$$\mathbb{P}\Big(\frac{1}{3} \le X \le \frac{2}{3}\Big) = \int_{1/3}^{2/3} f(x)\,dx = \int_{1/3}^{2/3} \frac{1}{2}\,dx = \frac{1}{2}\left(\frac{2}{3} - \frac{1}{3}\right) = \frac{1}{6}.$$

3. ***Expectation:***

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\,dx = \int_0^2 x \cdot \frac{1}{2}\,dx = \frac{1}{2}\int_0^2 x\,dx = \frac{1}{2}\cdot\frac{2^2}{2} = 1.$$

4. ***Variance:*** *First, compute* $\mathbb{E}[X^2]$*:*

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_0^2 x^2 \cdot \frac{1}{2}\,dx = \frac{1}{2}\int_0^2 x^2 dx = \frac{1}{2}\cdot\frac{8}{3} = \frac{4}{3}.$$

Then,

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{4}{3} - 1^2 = \frac{1}{3}.$$

2.6 Continuous Probability Laws

Continuous probability laws describe random variables that can take any value in a range (like height or time) using a probability density function, where probability is the area under the curve over an interval, not a single point (which is zero). Key laws involve properties like total area and calculating probabilities through integration, with common examples including the Uniform, Normal, and Exponential distributions, crucial for modeling real-world measurements.

2.6.1 Uniform Distribution

Definition 2.10. *A random variable* X *follows a* uniform distribution *on the interval* $[a, b]$ *if it has the probability density function*

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{if } a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}$$

Remark 2.7. *It is easy to verify that* f *is a probability density function.*

The cumulative distribution function F of X is given by

$$F(x) = \mathbb{P}(X \le x) = \begin{cases} 0, & \text{if } x < a, \\ \dfrac{x-a}{b-a}, & \text{if } a \le x \le b, \\ 1, & \text{if } x > b. \end{cases}$$

The expectation and variance of X are

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

Exercise 2.7. *Let* X *be a random variable uniformly distributed on* $[0, 10]$*. Compute* $\mathbb{P}(X < 3)$*,* $\mathbb{P}(X > 6)$*, and* $\mathbb{P}(3 < X < 8)$*.*

2.6.2 The Normal Distribution

Standard Normal Distribution

The normal distribution is the most important probability distribution. It plays a central role in many probabilistic models and in statistics as a whole. It possesses several remarkable properties that make it particularly convenient to use.

Definition 2.11. *A random variable X is said to follow a* normal distribution *(or* Gaussian distribution*, or* Laplace–Gauss distribution*)* $\mathcal{N}(0,1)$ *if its probability density function f is given by*

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \qquad \text{for all } x \in \mathbb{R}.$$

Remark 2.8. *1. Let us verify that f has integral equal to 1. Let*

$$I = \int_{\mathbb{R}} f(x)\,dx.$$

Then

$$I^2 = \left(\int_{\mathbb{R}} f(x)\,dx\right)\left(\int_{\mathbb{R}} f(y)\,dy\right) = \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} \exp\left(-\frac{x^2+y^2}{2}\right) dx\,dy.$$

Using the polar change of variables $x = r\cos\theta$, $y = r\sin\theta$, we have $dx\,dy = r\,dr\,d\theta$, and thus

$$I^2 = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{+\infty} \exp\left(-\frac{r^2}{2}\right) r\,dr\,d\theta = 1.$$

Hence, $I = 1$, and f is indeed a probability density function.

2. If X follows a standard normal distribution $\mathcal{N}(0,1)$, then for all real numbers $a < b$,

$$\mathbb{P}(a \le X \le b) = \frac{1}{\sqrt{2\pi}}\int_a^b \exp\left(-\frac{x^2}{2}\right) dx = \Phi(b) - \Phi(a),$$

where Φ denotes the cumulative distribution function of X. Recall that Φ is an antiderivative of the density function f. However, there is no closed-form expression for this antiderivative. Therefore, the values of Φ must be read from statistical tables.

3. The graph of the density function of the standard normal distribution $\mathcal{N}(0,1)$ is called the bell curve*. It tends to 0 as $x \to \pm\infty$, is increasing on $\mathbb{R}^-$ and decreasing on $\mathbb{R}^+$. Consequently, it attains its maximum at $x = 0$. Moreover, it is symmetric with respect to 0, which is its center of symmetry.*

Proposition 2.7. *Let X be a random variable with standard normal distribution $\mathcal{N}(0,1)$. Then X is* centered*, that is, it has zero mean, and* reduced*, that is, it has variance 1. Moreover, the random variable $-X$ also follows a standard normal distribution.*

Proof. The computation of the expectation is immediate once we observe that the function $xf(x)$ is odd, where f is the density of X. Hence,

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x)\,dx = 0.$$

The variance is obtained by an integration by parts:

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

Finally, we show that X and $-X$ have the same distribution. For all $a, b \in \mathbb{R}$,

$$\mathbb{P}(a \leq -X \leq b) = \mathbb{P}(-b \leq X \leq -a) = \int_{-b}^{-a} f(x)\, dx.$$

Performing the change of variables $y = -x$, we obtain

$$\mathbb{P}(a \leq -X \leq b) = \int_a^b f(-y)\, dy = \int_a^b f(y)\, dy,$$

by symmetry of the density function f. Therefore,

$$\mathbb{P}(a \leq -X \leq b) = \mathbb{P}(a \leq X \leq b),$$

which proves that $-X$ also follows a standard normal distribution. □

Use of Normal Distribution Tables

To compute probabilities of the form $\mathbb{P}(a \leq X \leq b)$ or $\mathbb{P}(X \leq x)$, one may use numerical computation on a computer or, more simply, standard normal distribution tables, which provide the values of $\mathbb{P}(X \leq x)$ for all positive decimal values of x (typically given to two decimal places).

Note that

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

Thus, if a and b are positive, the required probabilities can be directly read from the table.

To compute $\mathbb{P}(X \leq -x)$ for $x > 0$, we use the fact that X and $-X$ have the same distribution:

$$\mathbb{P}(X \leq -x) = \mathbb{P}(-X \geq x) = \mathbb{P}(X \geq x) = 1 - \mathbb{P}(X \leq x).$$

Examples with the Standard Normal Distribution

Example 2.8. *We want to calculate* $\mathbb{P}[-1 \leq X \leq 1]$ *for* $X \sim \mathcal{N}(0,1)$*:*

$$\mathbb{P}[-1 \leq X \leq 1] = \mathbb{P}(X \leq 1) - \mathbb{P}(X \leq -1) = \mathbb{P}(X \leq 1) - [1 - \mathbb{P}(X \leq 1)] = 2\mathbb{P}(X \leq 1) - 1.$$

Using the standard normal table, $\mathbb{P}(X \leq 1) = 0.8413$*, hence*

$$\mathbb{P}[-1 \leq X \leq 1] = 2 \times 0.8413 - 1 = 0.6826.$$

Example 2.9. *We want to find* $u \in \mathbb{R}$ *such that*

$$\mathbb{P}[-u \leq X \leq u] = 0.90, \quad X \sim \mathcal{N}(0,1).$$

Using symmetry of the standard normal distribution,

$$\mathbb{P}[-u \leq X \leq u] = 2\mathbb{P}(X \leq u) - 1.$$

Hence,

$$\mathbb{P}(X \leq u) = 0.95 \quad \Rightarrow \quad u = 1.6446.$$

General Normal Distribution

Proposition 2.8. *Let $m \in \mathbb{R}$ and $\sigma > 0$. A random variable X follows a* normal distribution $\mathcal{N}(m, \sigma^2)$ *if its probability density function f is given by*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right], \qquad x \in \mathbb{R}.$$

Remark. - The standard normal distribution is a special case with $m = 0$ and $\sigma = 1$. - If $\sigma > 1$, the distribution is wider (more spread out); if $\sigma < 1$, it is narrower. - The cumulative distribution function is given by

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

Proposition 2.9. *Let $X \sim \mathcal{N}(m, \sigma^2)$ and define*

$$Z = \frac{X - m}{\sigma}.$$

Then Z follows the standard normal distribution $\mathcal{N}(0, 1)$.

Proof. Let $a, b \in \mathbb{R}$ with $a < b$. Then

$$\mathbb{P}(a \leq Z \leq b) = \mathbb{P}(a \leq \frac{X-m}{\sigma} \leq b) = \mathbb{P}(m+a\sigma \leq X \leq m+b\sigma) = \int_{m+a\sigma}^{m+b\sigma} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right] dx.$$

Perform the change of variable $z = \frac{x-m}{\sigma}$, then $dx = \sigma dz$:

$$\mathbb{P}(a \leq Z \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz.$$

Since this is true for all $a < b$, the density of Z is the standard normal density. Hence, $Z \sim \mathcal{N}(0, 1)$. □

Proposition 2.10. *If $X \sim \mathcal{N}(m, \sigma^2)$, then*

$$\mathbb{E}[X] = m \quad \textit{and} \quad \mathrm{Var}(X) = \sigma^2.$$

Remark 2.9. *By using the previous proposition, we can always reduce to the standard normal distribution to compute probabilities. Let $X \sim \mathcal{N}(m, \sigma^2)$ and $a, b \in \mathbb{R}$. Then*

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}\left(\frac{a-m}{\sigma} \leq Z \leq \frac{b-m}{\sigma}\right),$$

where $Z \sim \mathcal{N}(0, 1)$. Hence,

$$\mathbb{P}(a \leq X \leq b) = \Phi\Big(\frac{b-m}{\sigma}\Big) - \Phi\Big(\frac{a-m}{\sigma}\Big),$$

where Φ is the cumulative distribution function of the standard normal.

Remark 2.10. *If $Z \sim \mathcal{N}(0, 1)$, then $-Z \sim \mathcal{N}(0, 1)$ as well. Therefore, for $u > 0$,*

$$\mathbb{P}(-u \leq Z \leq u) = \mathbb{P}(Z \leq u) - \mathbb{P}(Z \leq -u) = \mathbb{P}(Z \leq u) - (1 - \mathbb{P}(Z \leq u)) = 2\mathbb{P}(Z \leq u) - 1.$$

Example 2.10. *Let $X \sim \mathcal{N}(15, 16)$. Compute $\mathbb{P}(10 \leq X \leq 22)$.*

Standardizing:

$$\mathbb{P}(10 \leq X \leq 22) = \mathbb{P}\left(\frac{10-15}{4} \leq \frac{X-15}{4} \leq \frac{22-15}{4}\right) = \mathbb{P}(-1.25 \leq Y \leq 1.75),$$

where $Y \sim \mathcal{N}(0, 1)$. Then

$$\mathbb{P}(10 \leq X \leq 22) = \mathbb{P}(Y \leq 1.75) - \mathbb{P}(Y \leq -1.25) = 0.9599 + 0.8944 - 1 = 0.8543.$$

Example 2.11. *Let $Y \sim \mathcal{N}(0, 1)$. Then:*

$$\mathbb{P}(-1 \leq Y \leq 1) = 0.6826, \quad \mathbb{P}(-2 \leq Y \leq 2) = 0.9544, \quad \mathbb{P}(-3 \leq Y \leq 3) = 0.9973.$$

Let $X \sim \mathcal{N}(m, \sigma^2)$. Then:

$$\mathbb{P}(m-\sigma \leq X \leq m+\sigma) = 0.6826, \quad \mathbb{P}(m-2\sigma \leq X \leq m+2\sigma) = 0.9544,$$
$$\mathbb{P}(m-3\sigma \leq X \leq m+3\sigma) = 0.9973.$$

2.6.3 Exponential Distribution

Let $\lambda > 0$ be a strictly positive real number.

Definition 2.12 (Exponential Distribution)**.** *A random variable X follows an exponential distribution $E(\lambda)$ if its probability density function is given by*

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

The random variable X only takes positive values. It is widely used in applications such as:

- the operating time of a computer system before the first failure,
- radioactive decay,
- waiting times between arrivals of "clients" in queuing systems (service counters, access to a server, work accidents, etc.).

Proposition 2.11. *The cumulative distribution function of $X \sim E(\lambda)$ is*

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

Moreover,

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

Proposition 2.12 (Memoryless Property)**.** *The exponential distribution satisfies the* memoryless property*. Let $X \sim E(\lambda)$. Then, for all $s, t > 0$,*

$$\mathbb{P}[X > t + s \mid X > t] = \mathbb{P}[X > s].$$

Proof.

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

□

2.7 Classical Probability Laws

Table 2.2: Discrete laws

Law of X	**Parameters**	$P(X=k)$	$E(X)$	$V(X)$	$\varphi_X(t)$
Uniform $\{a,\dots,b\}$	$a \le b$	$\frac{1}{b-a+1}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$	$\frac{e^{itb}-e^{ita-1}}{(b-a+1)(e^{it}-1)}$
Bernoulli $\{0,1\}$	$p \in [0,1]$	$p^k(1-p)^{1-k}$	p	$p(1-p)$	$pe^{it}+(1-p)$
Binomial $\{0,\dots,n\}$	$n \in \mathbb{N}, p \in [0,1]$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$	$(pe^{it}+1-p)^n$
Poisson $\mathbb{N}$	$\lambda > 0$	$\frac{e^{-\lambda}\lambda^k}{k!}$	λ	λ	$e^{\lambda(e^{it}-1)}$
Geometric $\mathbb{N}^*$	$p \in (0,1)$	$p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^{it}}{1-(1-p)e^{it}}$

Table 2.3: Continuous laws

Law of X	**Parameters**	$f_X(x)$	$E(X)$	$V(X)$	$\varphi_X(t)$
Uniform $[a,b]$	$a < b$	$\frac{1}{b-a}\mathbf{1}_{[a,b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{itb}-e^{ita}}{it(b-a)}$
Exponential	$l > 0$	$le^{-lx}\mathbf{1}_{[0,\infty)}(x)$	$\frac{1}{l}$	$\frac{1}{l^2}$	$\frac{l}{l-it}$
Normal	$m \in \mathbb{R}, \sigma > 0$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-m)^2/(2\sigma^2)}$	m	σ^2	$e^{itm-\sigma^2t^2/2}$

2.8 Function of a Continuous Random Variable

Let X be a continuous random variable with density f, and let ϕ be a function on $\mathbb{R}$. Define

$$Y = \phi(X).$$

We seek the density of Y in the case where Y is continuous. To do this, we write

$$\mathbb{P}[a \le Y \le b] = \int_a^b g(y)\,dy, \quad \text{for all } a < b,$$

where g is the density of Y.

Suppose ϕ is bijective, of class C^1, and increasing. Otherwise, we divide $\mathbb{R}$ into intervals where ϕ is bijective, before performing the change of variable below. Let $a < b$ be arbitrary:

$$\mathbb{P}[a \le Y \le b] = \mathbb{P}[a \le \phi(X) \le b] = \mathbb{P}[\phi^{-1}(a) \le X \le \phi^{-1}(b)] = \int_{\phi^{-1}(a)}^{\phi^{-1}(b)} f(x)\,dx.$$

Using the change of variable $y = \phi(x)$, we obtain

$$\mathbb{P}[a \le Y \le b] = \int_a^b f(\phi^{-1}(y))\left|(\phi^{-1})'(y)\right|dy.$$

Thus, the density of Y is

$$h(y) = f(\phi^{-1}(y))\ \left|(\phi^{-1})'(y)\right|.$$

Example 2.12. *Let X have density $f = \frac{1}{2}\mathbf{1}_{[0,2]}$. Let $c \neq 0$ and d be real numbers. Find the law of $Y = cX + d$.*

For arbitrary $a < b$, For example, we assume that c is positive.

$$\mathbb{P}[a \leq Y \leq b] = \mathbb{P}[a \leq cX + d \leq b] = \mathbb{P}\left[\frac{a-d}{c} \leq X \leq \frac{b-d}{c}\right] = \frac{1}{2}\int_{\frac{a-d}{c}}^{\frac{b-d}{c}} \mathbf{1}_{[0,2]}(x)\,dx.$$

With the change of variable $y = cx + d$, $dy = c\,dx$:

$$\mathbb{P}[a \leq Y \leq b] = \frac{1}{2c}\int_a^b \mathbf{1}_{[0,2]}\left(\frac{y-d}{c}\right)\,dy.$$

Since $0 \leq \frac{y-d}{c} \leq 2$ iff $d \leq y \leq 2c + d$, we get

$$\mathbb{P}[a \leq Y \leq b] = \frac{1}{2c}\int_a^b \mathbf{1}_{[d,2c+d]}(y)\,dy.$$

Hence, the density of Y is

$$g(y) = \frac{1}{2c}\mathbf{1}_{[d,2c+d]}(y).$$

Theorem 2.3 (A Useful Transformation)**.** *Let X be a continuous random variable with cumulative distribution function F. Then $F(X)$ follows the uniform distribution on $[0, 1]$.*

Proof. Even if F is not strictly increasing, we can define a generalized inverse by

$$F^{-1}(y) = \min\{x : F(x) \geq y\}.$$

Then, by continuity of F,

$$\mathbb{P}[F(X) < y] = \mathbb{P}[X < F^{-1}(y)] = F(F^{-1}(y)) = y,$$

which is exactly the cumulative distribution function of the uniform distribution on $[0, 1]$. □

This result allows us to simulate many distributions using random numbers that are uniformly distributed between 0 and 1.

Theorem 2.4 (Inverse Transform Method)**.** *Let U be a random variable with uniform distribution on $[0, 1]$. Then*

$$Y = F^{-1}(U)$$

has cumulative distribution function F.

Proof. By the definition of F^{-1}, for all $0 < y < 1$,

$$F^{-1}(y) \leq x \iff F(x) \geq y.$$

Replacing y with U, we obtain

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

□

Example 2.13. *To simulate a realization of a random variable with exponential distribution $E(\lambda)$, it suffices to compute*

$$-\frac{\log(1-u)}{\lambda},$$

where u is a random number uniformly distributed between 0 and 1.

2.9 Characteristic function and Fourier transform

To study convergence in distribution in more detail, we will use the notion of the *characteristic function* of a random variable and the *Fourier transform* of a probability measure.

Definition 2.13. *The characteristic function of a random variable $Y : \Omega \to \mathbb{R}$ is defined by*

$$\varphi_Y(t) = \mathbb{E}(e^{itY}) = \int_\Omega e^{itY}\, dP = \int_{\mathbb{R}} e^{ity}\, dP_Y(y).$$

The Fourier transform of a probability measure μ defined on the Borel σ-algebra of $\mathbb{R}$ is defined by

$$\hat{\mu}(t) = \int_{\mathbb{R}} e^{itx}\, d\mu(x).$$

Hence, we have the equality $\varphi_Y(t) = \hat{P}_Y(t)$.

Properties:

- $|\varphi_Y(t)| \leq 1$ for all $t \in \mathbb{R}$,
- $\varphi_Y(0) = 1$,
- $t \mapsto \varphi_Y(t)$ is continuous on $\mathbb{R}$,
- if Y is integrable, then $t \mapsto \varphi_Y(t)$ is differentiable and $\varphi_Y'(0) = i\,\mathbb{E}(Y)$,
- if Y is square-integrable, then $t \mapsto \varphi_Y(t)$ is of class C^2 and $\varphi_Y''(0) = -\mathbb{E}(Y^2)$.

The continuity and differentiability follow from the theorems on continuity and differentiability under the integral sign. For example, if Y is integrable, we have

$$\left|\frac{\partial}{\partial t} e^{itY}\right| = |iY e^{itY}| \leq |Y|,$$

which implies

$$\varphi_Y'(t) = \frac{d}{dt}\int_\Omega e^{itY}\, dP = \int_\Omega \frac{\partial}{\partial t} e^{itY}\, dP = \int_\Omega iY e^{itY}\, dP.$$

The distribution of a random variable is completely determined by its characteristic function.

Proposition 2.13. *Two random variables with the same characteristic function have the same distribution:*

$$\varphi_X = \varphi_Y \implies P_X = P_Y.$$

We now proceed to some explicit calculations of characteristic functions.

Discrete case: The random variable Y takes a finite or countable set of values y_k, $k \in I$, with $I = \{1, \ldots, n\}$ or $I = \mathbb{N}$. Then

$$\varphi_Y(t) = \mathbb{E}(e^{itY}) = \sum_{k \in I} e^{ity_k} P(Y = y_k).$$

- **Bernoulli law with parameter** $p \in [0,1]$
  If Y follows a Bernoulli distribution, then $P(Y=0) = 1-p$ and $P(Y=1) = p$. The characteristic function is

$$\varphi_Y(t) = e^{it\cdot 0}P(Y=0) + e^{it\cdot 1}P(Y=1) = 1 - p + pe^{it}.$$

- **Uniform law on** $\{1,\dots,n\}$**,** $n \in \mathbb{N}^*$
  If Y is uniformly distributed on $\{1,\dots,n\}$, then $P(Y=k) = 1/n$ for $k = 1,\dots,n$, which implies

$$\varphi_Y(t) = \sum_{k=1}^{n} e^{itk}P(Y=k) = \frac{1}{n}\sum_{k=1}^{n} e^{itk} = \frac{e^{it}\left(1-e^{itn}\right)}{n(1-e^{it})}, \quad t \notin 2\pi\mathbb{Z}.$$

- **Continuous case**
  A random variable Y with density $f_Y : \mathbb{R} \to \mathbb{R}^+$ satisfies $P(Y \in A) = \int_A f_Y(y)\,dy$, so

$$\varphi_Y(t) = \mathbb{E}(e^{itY}) = \int_{\mathbb{R}} e^{ity}\,dP_Y(y) = \int_{\mathbb{R}} e^{ity} f_Y(y)\,dy.$$

- **Uniform law on** $[a,b]$**,** $a < b$
  For Y uniform on $[a,b]$, we have

$$\varphi_Y(t) = \int_{\mathbb{R}} e^{ity}\frac{1}{b-a}\mathbf{1}_{[a,b]}(y)\,dy = \frac{1}{b-a}\int_a^b e^{ity}\,dy = \frac{e^{itb}-e^{ita}}{it(b-a)}, \quad t \neq 0.$$

- **Exponential law with parameter** $\lambda > 0$
  For $Y \sim \mathrm{Exp}(\lambda)$, we have

$$\varphi_Y(t) = \int_0^{+\infty} e^{ity}\lambda e^{-\lambda y}\,dy = \int_0^{+\infty} \lambda e^{(it-\lambda)y}\,dy = \frac{\lambda}{\lambda - it}.$$

- **Gauss law with parameters** μ **and** σ^2 For $Y \sim \mathcal{N}(\mu,\sigma^2)$, we have

$$\varphi_Y(t) = e^{it\mu - \frac{1}{2}t^2\sigma^2}.$$

Remark 2.11. *Sometimes, instead of the characteristic function, one uses the notion of the generating function.*

2.10 Exercises

Exercise 2.8. *Show that if X is a random variable with a standard normal distribution $N(0,1)$, then the random variable*

$$Z = \sigma X + m$$

follows a normal distribution $N(m,\sigma^2)$.

Solution 2.8. *Let X be a random variable with distribution $N(0,1)$, and define*

$$Z = \sigma X + m,$$

where $\sigma > 0$ and $m \in \mathbb{R}$.

The cumulative distribution function of Z is, for any $z \in \mathbb{R}$,

$$\begin{aligned} P(Z \le z) &= P(\sigma X + m \le z) \\ &= P\Big(X \le \frac{z-m}{\sigma}\Big). \end{aligned}$$

Since $X \sim N(0,1)$, we have

$$P\Big(X \le \frac{z-m}{\sigma}\Big) = \Phi\Big(\frac{z-m}{\sigma}\Big),$$

where Φ denotes the distribution function of the standard normal law.

This is exactly the distribution function of a normal random variable with mean m and variance σ^2. Therefore,

$$Z \sim N(m, \sigma^2).$$

A linear transformation of a standard normal random variable remains normally distributed, with mean multiplied and shifted by the same transformation.

Exercise 2.9. *Let X be a standard normal random variable, $X \sim N(0,1)$. Calculate the following probabilities:*

$$P[X \le 1.62], \quad P[X \ge -0.52], \quad P[-1 < X < 1].$$

Find u such that

$$P[X \le u] = 0.334.$$

Exercise 2.10. *The width (in cm) of a slot in an aluminum part is normally distributed with mean $\mu = 2$ and standard deviation $\sigma = 0.012$. The tolerance limits are given as 2.000 ± 0.012. Determine the percentage of defective parts.*

Exercise 2.11. *Let X be a Gaussian random variable. It is known that*

$$P[X \le 3] = 0.5517 \quad \textit{and} \quad P[X \ge 7] = 0.0166.$$

Determine the mean and standard deviation of X.

Solution 2.9. *Let $X \sim \mathcal{N}(\mu, \sigma^2)$.*

We are given

$$P(X \le 3) = 0.5517 \quad \textit{and} \quad P(X \ge 7) = 0.0166.$$

From the standard normal distribution table,

$$P(Z \le 0.13) = 0.5517, \qquad P(Z \ge 2.13) = 0.0166,$$

where $Z \sim \mathcal{N}(0,1)$.

Thus,

$$\frac{3-\mu}{\sigma} = 0.13, \qquad \frac{7-\mu}{\sigma} = 2.13.$$

Subtracting the first equation from the second gives

$$\frac{4}{\sigma} = 2.00 \quad \Longrightarrow \quad \sigma = 2.$$

Substituting $\sigma = 2$ *into*

$$\frac{3-\mu}{2} = 0.13 \quad \Longrightarrow \quad \mu = 3 - 0.26 = 2.74.$$

Exercise 2.12. *1. In a packaging factory, a machine fills coffee packages of 250g. The amount of coffee actually poured is variable, following a normal distribution with adjustable mean and standard deviation 3. What theoretical mean should be chosen so that 90% of the customers receive at least 250g of coffee?*

2. Let X *be a continuous random variable with density*

$$f(x) = \begin{cases} a(9x - 3x^2), & 0 < x < 3, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Compute the constant a*.*

(b) Determine $\mathbb{P}[X > 1]$ *and* $\mathbb{P}[1/2 < X < 3/2]$*.*

(c) Find the cumulative distribution function of X*.*

(d) Compute the expectation and variance of X*.*

3. Let X *be a continuous uniform random variable on* $[a, b]$*, with* $a < b$ *positive. Let* c *and* d *be positive numbers such that* $a < pc < pd < b$*. Express*

$$\mathbb{P}[c < X^2 < d]$$

as an integral from c *to* d *and deduce the density of* X^2*.*

4. Let X *be an exponential random variable* $E(1)$*. Compute* $\mathbb{P}[X \leq 2]$ *and* $\mathbb{P}[X > 0.5]$*. Determine the density of* $Y = 3X$*. What distribution does* Y *follow?*

Solution 2.10. *1.* ***Coffee package problem***

Let $X \sim N(\mu, \sigma = 3)$*. We want* $\mathbb{P}[X \geq 250] = 0.9$*, i.e.,* $\mathbb{P}[X < 250] = 0.1$*.*

$$z = \frac{250-\mu}{3} \implies z = -1.28 \implies 250 - \mu = -3.84 \implies \mu \approx 253.84.$$

2. ***Continuous random variable with density*** $f(x) = a(9x - 3x^2)$**,** $0 < x < 3$

(a) Compute a*:*

$$\int_0^3 a(9x - 3x^2)dx = 1 \implies a = \frac{2}{27}.$$

(b) Probabilities:

$$\mathbb{P}[X > 1] = 1 - \int_0^1 f(x)dx = \frac{7}{9}, \quad \mathbb{P}[1/2 < X < 3/2] = \int_{1/2}^{3/2} f(x)dx \approx 0.37.$$

(c) CDF:

$$F(x) = \begin{cases} 0, & x \le 0, \\ \frac{2}{27}(4.5x^2 - x^3), & 0 < x < 3, \\ 1, & x \ge 3. \end{cases}$$

(d) Expectation:

$$\mathbb{E}[X] = \int_0^3 xf(x)dx = \frac{2}{27}\int_0^3 (9x^2 - 3x^3)dx = 1.5.$$

Variance:

$$\mathrm{Var}(X) = \int_0^3 x^2 f(x)dx - (\mathbb{E}[X])^2 = 0.36.$$

3. ***Transformation of uniform variable***

Let $X \sim U[a, b]$, $Y = X^2$, $a < \sqrt{c} < \sqrt{d} < b$:

$$\mathbb{P}[c < X^2 < d] = \mathbb{P}[\sqrt{c} < X < \sqrt{d}] = \int_{\sqrt{c}}^{\sqrt{d}} \frac{1}{b-a}dx = \frac{\sqrt{d} - \sqrt{c}}{b-a}.$$

Density of Y:

$$f_Y(y) = \frac{1}{2\sqrt{y}(b-a)}, \quad y \in [a^2, b^2].$$

4. ***Exponential distribution***

Let $X \sim E(1)$:

$$\mathbb{P}[X \le 2] = 1 - e^{-2} \approx 0.8647, \quad \mathbb{P}[X > 0.5] = e^{-0.5} \approx 0.6065.$$

Let $Y = 3X$. Then

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X \le y/3) = 1 - e^{-y/3}, \quad y > 0.$$

Hence $Y \sim E(\lambda = 1/3)$.

2.11 Random pairs of variables

2.11.1 Discrete Random Pairs and Vectors

Joint Distribution

Let X and Y be two discrete random variables with

$$X(\Omega) = \{x_i\,;\, i \in \mathbb{N}\} \quad \text{and} \quad Y(\Omega) = \{y_j\,;\, j \in \mathbb{N}\}.$$

The joint distribution of the pair (X, Y) is defined by the image set

$$(X, Y)(\Omega) \subset X(\Omega) \times Y(\Omega),$$

together with the probabilities

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}\Big(\{\omega \in \Omega\,;\, X(\omega) = x \text{ and } Y(\omega) = y\}\Big),$$

for every pair $(x, y) \in (X, Y)(\Omega)$.

Remark. We must of course have

$$\sum_{x,y} \mathbb{P}(X = x, Y = y) = 1.$$

More generally, let $X_1, \ldots, X_n$ be n discrete random variables taking values in $\mathbb{N}$. The joint distribution of the random vector $(X_1, \ldots, X_n)$ is defined by the image set

$$(X_1, \ldots, X_n)(\Omega) \subset \mathbb{N}^n,$$

together with the probabilities

$$\mathbb{P}(X_1 = i_1, \ldots, X_n = i_n),$$

for every n-tuple $(i_1, \ldots, i_n) \in \mathbb{N}^n$.

Example 2.14. *Fix $p \in (0, 1)$ and $\lambda > 0$, and consider the pair of random variables (X, Y) taking values in $\{0, 1\} \times \mathbb{N}$, whose joint distribution is given by*

$$\begin{cases} \mathbb{P}(X = 0, Y = 0) = 1 - p, \\ \mathbb{P}(X = 1, Y = k) = p\, e^{-\lambda} \dfrac{\lambda^k}{k!}, \quad \textit{for all } k \in \mathbb{N}, \\ \mathbb{P}(X = j, Y = k) = 0, \quad \textit{otherwise.} \end{cases}$$

We clearly have

$$\sum_{i,j} \mathbb{P}(X = i, Y = j) = 1,$$

and thus the joint distribution of the discrete random pair (X, Y) is well defined.

Example 2.15. *Consider an urn containing four tokens numbered from 1 to 4. Two tokens are drawn successively at random* without replacement*. Let (X, Y) denote the outcomes of the first and second draws, respectively.*

We have

$$\mathbb{P}(X = i, Y = i) = 0 \quad \textit{for all } i \in \{1, 2, 3, 4\},$$

and

$$\mathbb{P}(X = i, Y = j) = \frac{1}{12}, \quad \textit{for } 1 \leq i, j \leq 4 \textit{ with } i \neq j.$$

The joint probabilities can be represented in the following table (for example, the entry in the second column of the first row corresponds to $\mathbb{P}(X = 1, Y = 2)$):

$X \backslash Y$	1	2	3	4
1	0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
2	$\frac{1}{12}$	0	$\frac{1}{12}$	$\frac{1}{12}$
3	$\frac{1}{12}$	$\frac{1}{12}$	0	$\frac{1}{12}$
4	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	0

Example 2.16. *(Trinomial distribution). Let n be a strictly positive integer and let $p_x, p_y > 0$ be real parameters such that*

$$p_x + p_y \leq 1.$$

The trinomial distribution with parameters (n, p_x, p_y) is defined as the joint distribution of the pair (X, Y) taking values in $\mathbb{N}^2$ and given, for all $(i, j) \in \mathbb{N}^2$ such that $i + j \leq n$, by

$$\mathbb{P}(X = i, Y = j) = \frac{n!}{i!\, j!\, (n-i-j)!}\, p_x^i\, p_y^j\, (1 - p_x - p_y)^{n-i-j}.$$

$$\mathbb{P}(X = i, Y = j) = 0 \quad \textit{otherwise.}$$

Exercise. Show that this indeed defines the joint distribution of a discrete random pair.

The trinomial distribution is an extension of the binomial distribution. Indeed, consider an experiment with three possible outcomes, denoted by x, y, and z, occurring with probabilities p_x, p_y, and

$$p_z = 1 - p_x - p_y.$$

Repeat this experiment independently n times (where n is fixed), and count the number of occurrences of outcome x (denoted by X) and of outcome y (denoted by Y) among these n trials. It is then a counting exercise to show that the pair (X, Y) follows a trinomial distribution with parameters (n, p_x, p_y).

Marginal Distributions

Definition 2.14 (Marginal distributions)**.** *The two marginal distributions of the pair (X, Y) are the distributions of the random variables X and Y. They are obtained as follows:*

$$\mathbb{P}(X = x) = \sum_{y \in Y(\Omega)} \mathbb{P}(X = x, Y = y), \qquad \mathbb{P}(Y = y) = \sum_{x \in X(\Omega)} \mathbb{P}(X = x, Y = y).$$

Proof. We have

$$\{X = x\} = \{X = x,\, Y \in Y(\Omega)\} = \bigcup_{y \in Y(\Omega)} \{X = x,\, Y = y\}.$$

Since the union is countable and consists of disjoint events, we obtain

$$\mathbb{P}(X = x) = \sum_{y \in Y(\Omega)} \mathbb{P}(X = x, Y = y).$$

More generally, a random vector $(X_1, \ldots, X_n)$ taking values in $\mathbb{Z}^n$ has n one-dimensional marginal distributions, but also $n(n-1)$ two-dimensional marginal distributions, and so on.

For example, we have

$$\mathbb{P}(X_1 = x) = \sum_{(x_2, \ldots, x_n) \in \mathbb{Z}^{n-1}} \mathbb{P}(X_1 = x, X_2 = x_2, \ldots, X_n = x_n).$$

Returning to the previous examples 2.14-2.15

Example 2.17. *Let us determine the distribution of X. The random variable X takes values in $\{0,1\}$, and we have*

$$\mathbb{P}(X=0)=\sum_{j\in\mathbb{N}}\mathbb{P}(X=0,Y=j)=1-p.$$

Similarly,

$$\mathbb{P}(X=1)=\sum_{j\in\mathbb{N}}\mathbb{P}(X=1,Y=j)=\sum_{j\geq 0}pe^{-\lambda}\frac{\lambda^j}{j!}=p.$$

Therefore, the random variable X follows a Bernoulli distribution with parameter p. Let us also compute the distribution of Y.

We have

$$\mathbb{P}(Y=0)=\mathbb{P}(X=0,Y=0)+\mathbb{P}(X=1,Y=0)=1-p+pe^{-\lambda}.$$

For every $j\geq 1$,

$$\mathbb{P}(Y=j)=\mathbb{P}(X=0,Y=j)+\mathbb{P}(X=1,Y=j)=pe^{-\lambda}\frac{\lambda^j}{j!}.$$

Example 2.18. *It is sufficient to sum the entries of the table by columns to obtain the distribution of X, and by rows to obtain the distribution of Y. In practice, one may add an extra column and an extra row to the table in order to write the marginal distributions of X and Y. Before concluding, one should check that the sum of the entries in this column (and in this row) is equal to 1.*

In this example, both X and Y follow the uniform distribution on

$$\{1,2,3,4\}.$$

2.11.2 Distribution of $f(X,Y)$

Problem. Let (X,Y) be a pair of discrete random variables whose joint distribution is known. We wish to determine the distribution of the random variable

$$Z=f(X,Y),$$

where

$$f: X(\Omega)\times Y(\Omega)\longrightarrow\mathbb{R}$$

is a given function. For instance, one often needs to find the distribution of $X+Y$, $X-Y$, or XY. Determining the distribution of X from that of (X,Y) corresponds to choosing the function $f(x,y)=x$.

Proposition 1.2. We have

$$Z(\Omega)=f\Big((X,Y)(\Omega)\Big),$$

and for every $z\in f\Big((X,Y)(\Omega)\Big)$,

$$\mathbb{P}(Z=z)=\sum_{\substack{(x,y)\in(X,Y)(\Omega)\\ f(x,y)=z}}\mathbb{P}(X=x,Y=y).$$

Example. Let us once again consider Example 2.14 and take the function $f(x,y) = xy$. The random variable XY takes values in $\mathbb{N}$, and we have

$$\mathbb{P}(XY = 0) = \mathbb{P}(X = 0, Y = 0) + \mathbb{P}(X = 1, Y = 0) = 1 - p + pe^{-\lambda}.$$

For every $k \in \mathbb{N}^*$,

$$\mathbb{P}(XY = k) = \mathbb{P}(X = 1, Y = k) = pe^{-\lambda}\frac{\lambda^k}{k!}.$$

An important special case. We now consider the function $f(x,y) = x + y$. We obtain

$$\mathbb{P}(X + Y = z) = \sum_{\substack{(x,y)\in(X,Y)(\Omega)\\ x+y=z}} \mathbb{P}(X = x, Y = y) = \sum_{x\in X(\Omega)} \mathbb{P}(X = x, Y = z - x)$$

$$= \sum_{y\in Y(\Omega)} \mathbb{P}(X = z - y, Y = y).$$

More generally, if $X = (X_1, \ldots, X_n)$ is a discrete random vector and

$$f : X(\Omega) \longrightarrow \mathbb{R}$$

is a given function, then

$$\mathbb{P}\Big(f(X_1, \ldots, X_n) = z\Big) = \sum_{\substack{x_1,\ldots,x_n\\ f(x_1,\ldots,x_n)=z}} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n).$$

From this, we deduce the following fundamental corollary:

Proposition 2.14. *Let X and Y be two discrete and integrable random variables. Then the random variable*

$$Z = X + Y$$

is integrable, and we have

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Proof. We have already seen that Z is a discrete random variable and that its distribution is given, for every $z \in (X + Y)(\Omega)$, by

$$\mathbb{P}(X + Y = z) = \sum_{x\in X(\Omega)} \mathbb{P}(X = x, Y = z - x).$$

Therefore,

$$\mathbb{E}(|X + Y|) = \sum_{z\in(X+Y)(\Omega)} \left[|z| \sum_{x\in X(\Omega)} \mathbb{P}(X = x, Y = z - x)\right].$$

Moreover,

$$\mathbb{E}(|X + Y|) = \sum_{z\in(X+Y)(\Omega)} \left[\sum_{x\in X(\Omega)} |x + (z - x)|\, \mathbb{P}(X = x, Y = z - x)\right].$$

Using the triangle inequality, we obtain

$$\mathbb{E}(|X+Y|) \leq \sum_{z\in(X+Y)(\Omega)} \left[\sum_{x\in X(\Omega)} \Big(|x| + |z-x| \Big) \mathbb{P}(X=x, Y=z-x) \right].$$

Hence,

$$\mathbb{E}(|X+Y|) \leq \sum_{z\in(X+Y)(\Omega)} \left[\sum_{x\in X(\Omega)} |x| \, \mathbb{P}(X=x, Y=z-x) \right]$$
$$+ \sum_{z\in(X+Y)(\Omega)} \left[\sum_{x\in X(\Omega)} |z-x| \, \mathbb{P}(X=x, Y=z-x) \right].$$

Let us first study the first sum:

$$\sum_{z\in(X+Y)(\Omega)} \left[\sum_{x\in X(\Omega)} |x| \, \mathbb{P}(X=x, Y=z-x) \right].$$

By changing the order of summation, we obtain

$$\sum_{x\in X(\Omega)} \left[|x| \sum_{z\in(X+Y)(\Omega)} \mathbb{P}(X=x, Y=z-x) \right].$$

Since

$$\sum_{z\in(X+Y)(\Omega)} \mathbb{P}(X=x, Y=z-x) = \mathbb{P}(X=x),$$

it follows that

$$\sum_{x\in X(\Omega)} |x| \, \mathbb{P}(X=x) = \mathbb{E}(|X|).$$

Let us now turn to the second sum. Summing over $x \in X(\Omega)$ or over those x such that $z - x \in Y(\Omega)$ does not change the value of the sum. Hence,

$$\sum_{z\in(X+Y)(\Omega)} \left[\sum_{x\in X(\Omega)} |z-x| \, \mathbb{P}(X=x, Y=z-x) \right] = \sum_{z\in(X+Y)(\Omega)} \left[\sum_{z-x\in Y(\Omega)} |z-x| \, \mathbb{P}(X=x, Y=z-x) \right].$$

Setting $y = z - x$, we obtain

$$\sum_{z\in(X+Y)(\Omega)} \left[\sum_{y\in Y(\Omega)} |y| \, \mathbb{P}(X=z-y, Y=y) \right].$$

By changing the order of summation, this becomes

$$\sum_{y\in Y(\Omega)} \left[|y| \sum_{z\in(X+Y)(\Omega)} \mathbb{P}(X=z-y, Y=y) \right].$$

Since

$$\sum_{z\in(X+Y)(\Omega)} \mathbb{P}(X=z-y, Y=y) = \mathbb{P}(Y=y),$$

we finally obtain

$$\sum_{y\in Y(\Omega)} |y| \, \mathbb{P}(Y=y) = \mathbb{E}(|Y|).$$

We therefore observe that if X and Y are integrable, then $X + Y$ is also integrable. Moreover, by carrying out the same computation without absolute values, we conclude that

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Example 2.19. *Suppose that the pair (X, Y) follows a trinomial distribution with parameters (n, p_x, p_y), and let us compute the distribution of $X + Y$.*

This random variable takes values in $\mathbb{N}$, and for every integer k we have

$$\mathbb{P}(X + Y = k) = \sum_{j=0}^{n} \mathbb{P}(X = j, Y = k - j).$$

For $k > n$, each term in the sum is equal to zero, hence

$$\mathbb{P}(X + Y = k) = 0.$$

Now fix an integer $k \in \{0, 1, \ldots, n\}$. We obtain

$$\begin{aligned}\mathbb{P}(X + Y = k) &= \sum_{j=0}^{k} \mathbb{P}(X = j, Y = k - j) \\ &= \sum_{j=0}^{k} \frac{n!}{j!(k-j)!(n-k)!} \, p_x^j \, p_y^{k-j} \, (1 - p_x - p_y)^{n-k}.\end{aligned}$$

Factoring out the terms independent of j, we get

$$\mathbb{P}(X + Y = k) = \frac{n!}{(n-k)!} \, (1 - p_x - p_y)^{n-k} \sum_{j=0}^{k} \frac{1}{j!(k-j)!} \, p_x^j p_y^{k-j}.$$

Using the binomial identity, this yields

$$\mathbb{P}(X + Y = k) = \frac{n!}{k!(n-k)!} \, (1 - p_x - p_y)^{n-k} \, (p_x + p_y)^k.$$

Therefore, the random variable $X + Y$ follows a binomial distribution with parameters

$$\mathrm{Bin}(n,\, p_x + p_y).$$

2.12 Conditional Law

Consider a pair (X, Y) of discrete random variables, whose joint distribution is known, and fix y such that $P(Y = y) > 0$.

The conditional distribution of X given the event $\{Y = y\}$ is defined as a distribution over $X(\Omega)$ with conditional probabilities $P(X = x \mid Y = y)$ for all $x \in X(\Omega)$.

It is easy to verify that

$$\sum_{x \in X(\Omega)} P(X = x \mid Y = y) = 1,$$

which implies that the conditional distribution of X given $\{Y = y\}$ is indeed the law of a random variable.

Moreover, the law of total probability implies that

$$P(X = x) = \sum_{y \in Y(\Omega)} P(X = x \mid Y = y) P(Y = y).$$

This definition of the conditional distribution extends to random vectors: for example, for a random triplet (X, Y, Z), one can study the conditional distribution of X given $\{Y = y\}$, the conditional distribution of X given $\{Y = y \text{ and } Z = z\}$, or the conditional distribution of the pair (X, Y) given $\{Z = z\}$, and so on.

Example 2.20. *The conditional distribution of Y given $\{X = 1\}$ is defined over $Y(\Omega) \subset \mathbb{N}$ and for any positive integer k, we have*

$$P(Y = k \mid X = 1) = \frac{P(Y = k \text{ and } X = 1)}{P(X = 1)} = \frac{p e^{-\lambda} \frac{\lambda^k}{k!}}{p} = e^{-\lambda} \frac{\lambda^k}{k!}.$$

Hence, the conditional distribution of Y given $\{X = 1\}$ is a Poisson distribution with parameter λ.

Example 2.21. *The conditional distribution of X given $\{Y = 1\}$ is the uniform distribution over $\{2, 3, 4\}$.*

2.12.1 Independence of Discrete Random Variables

Definition 2.15. *Two discrete random variables X and Y are said to be independent if for all $x \in X(\Omega)$ and all $y \in Y(\Omega)$, the events $\{X = x\}$ and $\{Y = y\}$ are independent, that is,*

$$P(X = x, Y = y) = P(X = x) P(Y = y).$$

If X and Y are two independent discrete random variables, then for all $y \in Y(\Omega)$ and all $x \in X(\Omega)$ such that $P(Y = y) > 0$ and $P(X = x) > 0$, we have

$$P(X = x \mid Y = y) = P(X = x) \quad \text{and} \quad P(Y = y \mid X = x) = P(Y = y).$$

More generally, n discrete random variables $X_1, \ldots, X_n$ are (mutually or n-wise) independent if, for any choice of $x_1 \in X_1(\Omega), \ldots, x_n \in X_n(\Omega)$, we have

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n).$$

Remark 2.12.

- *n-wise independence implies pairwise independence, but the converse is not true. Write a proof of this result for $n = 3$.*
- *Events $A_1, \ldots, A_n$ are independent if and only if the indicator random variables $1_{A_1}, \ldots, 1_{A_n}$ are independent.*

Remark 2.13. *When the random variables X and Y are independent, knowing the marginal distributions is enough to determine the joint distribution of the pair (X, Y), whereas for arbitrary random variables, this is not sufficient.*

2.13 Expectation and Covariance Matrix

For clarity, all results in this section and the next are stated for discrete random pairs, and they extend straightforwardly to discrete random vectors.

Consider a discrete random pair (X, Y).

Definition 2.16.
- *The **expectation** of the pair (X, Y) is defined if X and Y are integrable, and in that case we have*

$$E(X, Y) = (E(X), E(Y)).$$

- *If X and Y are square-integrable random variables, the **covariance** of X and Y, or of the pair (X, Y), is given by*

$$\operatorname{cov}(X, Y) = E(XY) - E(X)E(Y) = E\Big[(X - E(X))(Y - E(Y))\Big].$$

If X and Y are square-integrable random variables, the **covariance matrix** of the pair (X, Y) is the matrix

$$C = \begin{pmatrix} \operatorname{var}(X) & \operatorname{cov}(X, Y) \\ \operatorname{cov}(X, Y) & \operatorname{var}(Y) \end{pmatrix}.$$

More generally, the covariance matrix of a vector $(X_1, \ldots, X_n)$, each component of which is square-integrable, is an $n \times n$ matrix whose diagonal entries are the variances of X_i and whose (i, j)-th entry is the covariance $\operatorname{cov}(X_i, X_j)$ for all $i \neq j$.

Remark 2.14. *The calculation of the expectation of X or Y involves only the marginal distributions, but as we will see, it is not necessary to explicitly determine these marginal distributions.*

Proposition 2.15. *Let (X, Y) be a pair of discrete random variables. For any function $h : \mathbb{R}^2 \to \mathbb{R}$ such that*

$$\sum_{x \in X(\Omega),\, y \in Y(\Omega)} |h(x, y)| P(X = x, Y = y) < \infty,$$

the random variable $h(X, Y)$ is integrable and

$$E[h(X, Y)] = \sum_{x \in X(\Omega),\, y \in Y(\Omega)} h(x, y) P(X = x, Y = y).$$

Proof. Let $Z = h(X, Y)$. We have an explicit expression for the law of Z: this random variable is discrete, and for all $z \in (h(X, Y))(\Omega)$,

$$P(Z = z) = \sum_{\substack{(x,y) \in (X,Y)(\Omega) \\ h(x,y) = z}} P(X = x, Y = y).$$

Hence, Z is integrable if the sum

$$S = \sum_{z \in (h(X,Y))(\Omega)} |z| \sum_{\substack{(x,y) \in (X,Y)(\Omega) \\ h(x,y) = z}} P(X = x, Y = y)$$

is finite converges. Now, this sum can be written as

$$S = \sum_{z \in (h(X,Y))(\Omega)} \sum_{\substack{(x,y) \in (X,Y)(\Omega) \\ h(x,y) = z}} |h(x, y)| P(X = x, Y = y) = \sum_{(x,y) \in (X,Y)(\Omega)} |h(x, y)| P(X = x, Y = y).$$

Applying the same procedure to the sum without absolute values yields the desired result for $E[h(X, Y)]$. □

Application

This proposition allows us to express the expectations of X, Y, or XY without explicitly determining the law of these random variables. If the pair (X, Y) is discrete, we have

$$E(X) = \sum_{(x,y)\in(X,Y)(\Omega)} x\, P(X = x, Y = y),$$

$$E(Y) = \sum_{(x,y)\in(X,Y)(\Omega)} y\, P(X = x, Y = y),$$

$$E(XY) = \sum_{(x,y)\in(X,Y)(\Omega)} xy\, P(X = x, Y = y),$$

provided that the series converge.

Proposition 2.16. *1. A covariance matrix C is always symmetric and positive semi-definite, i.e., for all $v \in \mathbb{R}^n$, $\langle v, Cv\rangle \geq 0$.*

2. If X and Y are two independent and integrable random variables, then $E(XY) = E(X)E(Y)$ and hence $\operatorname{cov}(X, Y) = 0$. The converse of this result is false.

3. If X and Y are two independent random variables and f and g are functions such that $f(X)$ and $g(Y)$ are integrable, then $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$. The converse of this result is false.

4. If the random variables $X_1, \ldots, X_n$ are independent and square-integrable, then the covariance matrix of $(X_1, \ldots, X_n)$ is diagonal. The converse of this result is false.

Proof. 1. Let $X = (X_1, \ldots, X_n)$ be a random vector whose components are square-integrable, and let C denote its covariance matrix. Fix $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$.

Without loss of generality, by replacing each X_i with $X_i - E(X_i)$, we may assume that the random variables are centered (i.e., have zero mean). We have

$$\langle v, Cv\rangle = \sum_{i,j} C_{ij} v_i v_j = \sum_{i,j} E(X_i X_j) v_i v_j = E\left[\sum_{i,j} v_i X_i v_j X_j\right] = E\left[\left(\sum_i v_i X_i\right)\left(\sum_j v_j X_j\right)\right]$$

$$= E\left[\left(\sum_i v_i X_i\right)^2\right].$$

Hence, $\langle v, Cv\rangle \geq 0$.

2. Consider two discrete independent and integrable random variables X and Y. We show that XY is integrable and calculate its expectation. We have

$$E(|XY|) = \sum_{x\in X(\Omega),\, y\in Y(\Omega)} |xy| P(X = x, Y = y).$$

Since X and Y are independent, $P(X = x, Y = y) = P(X = x)P(Y = y)$, so

$$E(|XY|) = \sum_{x\in X(\Omega),\, y\in Y(\Omega)} |xy| P(X = x)P(Y = y) = \sum_{x\in X(\Omega)} |x| P(X = x) \sum_{y\in Y(\Omega)} |y| P(Y = y)$$

$$= E(|X|E|Y|),$$

which is finite because X and Y are integrable. We then show by a similar calculation that $E(XY) = E(X)E(Y)$, and hence we deduce that $\operatorname{cov}(X, Y) = 0$.

Points 3 and 4 follow easily from 2. □ □

Theorem 2.5 (Characterization of independence via measurable functions)**.** *Let X and Y be two random variables defined on the same probability space. The following assertions are equivalent:*

1. *X and Y are independent.*
2. *For all bounded (or integrable) measurable functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$,*
$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)].$$
3. *For all Borel sets $A, B \subset \mathbb{R}$,*
$$\mathbb{P}(X \in A,\ Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B).$$

Example 2.22. *Let X be a random variable uniformly distributed on $[-1, 1]$, and define*
$$Y = X^2.$$

- *Since the distribution of X is symmetric around 0, we have*
$$\mathbb{E}[X] = 0.$$
- *Moreover,*
$$\mathbb{E}[XY] = \mathbb{E}[X^3] = 0,$$
because X^3 is an odd function on $[-1, 1]$.
- *Hence,*
$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0.$$

However, X and Y are not independent*, since Y is a deterministic function of X. For example,*
$$\mathbb{P}(Y \le 1/4 \mid X > 0) = \mathbb{P}(X^2 \le 1/4 \mid X > 0) = \mathbb{P}(0 < X \le 1/2 \mid X > 0) = \tfrac{1}{2},$$
whereas
$$\mathbb{P}(Y \le 1/4) = \mathbb{P}(|X| \le 1/2) = \tfrac{1}{2}.$$
More generally, knowing X determines Y, which contradicts independence.

The following proposition, although very simple to prove, is very useful:

Proposition 2.17. *If X and Y are square-integrable random variables, then*
$$\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\,\mathrm{cov}(X, Y).$$
Moreover, if X and Y are independent, we have
$$\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y).$$

Furthermore, for two square-integrable random variables, it is possible to obtain an upper bound on $\mathrm{cov}(X, Y)$ using the variances of X and Y:

Proposition 2.18 (Cauchy-Schwarz Inequality)**.** *Let X and Y be square-integrable random variables. Then*

$$|E(XY)| \leq E(|XY|) \leq \sqrt{E(X^2)E(Y^2)},$$

and

$$|\operatorname{cov}(X,Y)| \leq \sqrt{\operatorname{var}(X)\operatorname{var}(Y)}.$$

Proof. The second point is obtained by applying the first to $X - E(X)$ and $Y - E(Y)$.

The first point is shown in the same way as in the classical case (scalar product of two vectors): we consider the polynomial

$$R(\lambda) = E(X^2)\lambda^2 - 2E|XY|\,\lambda + E(Y^2).$$

This polynomial can be factorized as

$$R(\lambda) = E\Big[(\lambda|X| - |Y|)^2\Big].$$

Hence, it cannot have two distinct real roots, which means that its (reduced) discriminant is non-positive:

$$(E|XY|)^2 - E(X^2)E(Y^2) \leq 0.$$

The equality case is treated similarly: the discriminant is zero only if the polynomial has a double real root. In that case, there exists λ_0 such that

$$E\Big[(\lambda_0|X| + |Y|)^2\Big] = 0.$$

Since the expectation of a non-negative random variable can be zero only if the variable itself is (almost surely) zero, we then have

$$|Y| = -\lambda_0|X|, \quad \text{almost surely.}$$

The proofs without absolute values are similar. □

2.14 Random Pairs with Density

2.14.1 Density of a Pair

Definition 2.17. *A pair (X,Y) of real random variables is said to* have a density with respect to the Lebesgue measure on $\mathbb{R}^2$ *if there exists a function $f : \mathbb{R}^2 \to \mathbb{R}^+$ such that, for all intervals I and J and for any bounded or positive continuous function h, we have*

$$P(X \in I \text{ and } Y \in J) = \int_{I\times J} f(x,y)\,dx\,dy \quad \text{and} \quad E[h(X,Y)] = \int_{\mathbb{R}^2} h(x,y)f(x,y)\,dx\,dy.$$

Remark 2.15.

- *The density f is always a positive, integrable function whose integral over $\mathbb{R}^2$ with respect to the Lebesgue measure equals 1.*
- *One can show that the definition of a density is equivalent to: for any open set $U \subset \mathbb{R}^2$,*

$$P((X,Y) \in U) = \int_U f(x,y)\,dx\,dy.$$

Examples

1. Let $D = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$ be the unit disk in $\mathbb{R}^2$, and define f on $\mathbb{R}^2$ by
$$f(x, y) = \frac{1_D(x, y)}{\pi}.$$
Then f is a probability density on $\mathbb{R}^2$: it is the density of the uniform distribution on D, i.e., the law obtained by choosing a point uniformly at random in D.

2. Let X be an exponential random variable. The pair (X, X) does not admit a density with respect to the Lebesgue measure on $\mathbb{R}^2$. Indeed, if such a density f existed, it would have to be (almost everywhere) zero outside the line $\{x = y\}$ since $P(X \neq Y) = 0$. We would then have
$$1 = \int_{\mathbb{R}^2} f(x, y)\, dx\, dy = \int_{\{x=y\}\subset\mathbb{R}^2} f(x, y)\, dx\, dy.$$
But a line has zero Lebesgue measure in $\mathbb{R}^2$, so the integral of any function over this domain (with respect to the Lebesgue measure on $\mathbb{R}^2$) is zero, leading to a contradiction.

2.14.2 Marginal Distributions

The marginal distributions of the pair (X, Y) are the distributions of the random variables X and Y. We can easily verify the following proposition:

Proposition 2.19. *Let (X, Y) be a pair of random variables with a (joint) density $f : \mathbb{R}^2 \to \mathbb{R}^+$. Then each of the random variables X and Y has a density, denoted respectively by f_X and f_Y, given by*
$$f_X(x) = \int_{\mathbb{R}} f(x, y)\, dy \quad \text{and} \quad f_Y(y) = \int_{\mathbb{R}} f(x, y)\, dx.$$

Proof. Indeed, let $t \in \mathbb{R}$ and compute the cumulative distribution function of X:
$$P(X \leq t) = P(X \in (-\infty, t] \text{ and } Y \in \mathbb{R}) = \int_{(-\infty,t]\times\mathbb{R}} f(x, y)\, dx\, dy = \int_{-\infty}^{t} \left(\int_{\mathbb{R}} f(x, y)\, dy \right) dx.$$
□

This principle applies to random vectors with densities, with the caveat that for a triple, the marginal distributions include the distributions of each individual variable as well as the distributions of the pairs of variables.

Exercise 2.13. *Compute the marginal distributions of the pair (X, Y) that is uniformly distributed over the disk D centered at 0 with radius 1.*

As mentioned previously, random variables X and Y are independent if, for all intervals I and J, we have
$$P(X \in I \text{ and } Y \in J) = P(X \in I)P(Y \in J).$$
More generally, n random variables $X_1, \ldots, X_n$ are (mutually) independent if, for all intervals $I_1, \ldots, I_n$, we have
$$P\left(\bigcap_{k\leq n} \{X_k \in I_k\} \right) = \prod_{k\leq n} P(X_k \in I_k).$$

2.14.3 Distribution of a Tuple of Independent Random Variables

The link between densities and independence is clear:

Proposition 2.20. *Two random variables (X, Y) with densities f and g are independent if and only if the law of the pair admits a density, and this density is the function*

$$(x, y) \mapsto f(x)g(y).$$

The distribution of a pair of random variables is the product of the distributions of each variable when they are independent.

Proposition 2.21. *Let X, Y be two independent random variables. Then*

$$P(X, Y) = P_X \otimes P_Y.$$

For any Borel function $h : \mathbb{R}^2 \to \mathbb{R}$, positive or $P(X,Y)$-integrable,

$$E(h(X, Y)) = \int_{\mathbb{R}^2} h(x, y)\, dP_{(X,Y)}(x, y) = \int_{\mathbb{R}} \int_{\mathbb{R}} h(x, y)\, dP_X(x)\, dP_Y(y).$$

Proposition 2.22. *Let $(\Omega, \mathcal{T}, P)$ be a probability space, and let $X, Y : \Omega \to \mathbb{R}$ be random variables. Let $f, g : \mathbb{R} \to \mathbb{R}$ be Borel functions such that $f(X)$ and $g(Y)$ are P-integrable. Assume that X and Y are independent. Then*

$$E(f(X)g(Y)) = E(f(X))E(g(Y)).$$

Proof. Using Fubini's theorem:

$$\begin{aligned} E(f(X)g(Y)) &= \int_{\mathbb{R}^2} f(x)g(y)\, dP_{(X,Y)}(x, y) \\ &= \int_{\mathbb{R}^2} f(x)g(y)\, dP_X(x)\, dP_Y(y) \\ &= \int_{\mathbb{R}} f(x)\, dP_X(x) \int_{\mathbb{R}} g(y)\, dP_Y(y) \\ &= E(f(X))E(g(Y)). \end{aligned}$$

□

The case where f and g are the identity functions gives the important formula:

$$E(XY) = E(X)E(Y),$$

when X and Y are independent and P-integrable.

These results generalize to tuples of random variables.

Proposition 2.23. *Let $X_1, \ldots, X_n$ be independent random variables, $h : \mathbb{R}^n \to \mathbb{R}$ a Borel function, positive or $P(X_1, \ldots, X_n)$-integrable, and $f_i : \mathbb{R} \to \mathbb{R}$ Borel functions, positive or P_{X_i}-integrable. Then*

$$P(X_1, \ldots, X_n) = P_{X_1} \otimes P_{X_2} \otimes \cdots \otimes P_{X_n},$$

$$E\Big(h(X_1, \ldots, X_n)\Big) = \int_{\mathbb{R}^n} h(x_1, \ldots, x_n)\, dP_{X_1}(x_1) \ldots dP_{X_n}(x_n),$$

$$E\left(\prod_{i=1}^{n} f_i(X_i)\right) = \prod_{i=1}^{n} E(f_i(X_i)).$$

Recall that the covariance of two random variables is defined as the expectation of the product minus the product of expectations. By taking f_i and f_j as the identity function, and the other f_k as the constant function equal to 1, we obtain the following corollary:

Corollary 2.1. *Let $X_1, \ldots, X_n$ be square-integrable, independent random variables. Then*

$$E(X_i X_j) = E(X_i)E(X_j), \quad \text{if } i \neq j,$$

$$\operatorname{Cov}(X_i, X_j) = 0, \quad \text{if } i \neq j,$$

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

2.14.4 Application: Law of the Sum of Two Independent Random Variables

To compute the law of a sum of two independent random variables, we introduce the convolution of two probability laws.

Definition 2.18. *Let X, Y be independent random variables. The* convolution *of the laws of X and Y is a probability measure defined for any Borel set $A \subset \mathbb{R}$ by*

$$P_X * P_Y(A) = \int_{\mathbb{R}} P_X(A - y)\, dP_Y(y),$$

where $A - y = \{x - y \mid x \in A\}$. Equivalently, for any bounded measurable function f,

$$\int_{\mathbb{R}} f\, d(P_X * P_Y) = \int_{\mathbb{R}^2} f(x + y)\, dP_X(x)\, dP_Y(y).$$

which show that the law of a sum of independent random variables is equal to the convolution of their laws.

Corollary 2.2. *The distribution of the sum of two independent random variables is equal to the convolution of their distributions:*

$$P_{X+Y} = P_X * P_Y.$$

It can be verified in an exercise that the convolution of the distributions of two continuous random variables with densities is a distribution whose density is the convolution of the densities:

$$f_{X+Y}(z) = f_X * f_Y(z) = \int_{\mathbb{R}} f_X(z - y) f_Y(y)\, dy.$$

2.14.5 Expectation and Covariance

Definition 2.19. *As for discrete pairs, the expectation of a pair of integrable random variables is defined as the pair of expectations:*

$$E(X, Y) = (E(X), E(Y)).$$

It is easy to verify that

$$E(X) = \int_{\mathbb{R}^2} x f(x, y)\, dx\, dy \quad \text{and} \quad E(Y) = \int_{\mathbb{R}^2} y f(x, y)\, dx\, dy,$$

whenever these integrals converge absolutely.

If the random variables are square-integrable, we also define the covariance matrix of the pair (X, Y) *by*

$$C = \begin{pmatrix} \operatorname{var}(X) & \operatorname{cov}(X,Y) \\ \operatorname{cov}(X,Y) & \operatorname{var}(Y) \end{pmatrix}.$$

The off-diagonal terms of this matrix are

$$\operatorname{cov}(X,Y) = E\Big((X - E(X))(Y - E(Y))\Big) = E(XY) - E(X)E(Y),$$

where

$$E(XY) = \int_{\mathbb{R}^2} xy\, f(x,y)\, dx\, dy.$$

The covariance matrix is, as in the discrete case, a symmetric and positive (in the sense of bilinear forms) matrix.

These definitions naturally extend to the case of random vectors with a density, by replacing double integrals with multiple integrals.

Remark 2.16. *To identify the density of a pair, one usually uses a test function* $h : \mathbb{R}^2 \to \mathbb{R}$*, continuous and bounded, and tries to write*

$$E(h(X,Y)) = \int_{\mathbb{R}^2} h(x,y) f(x,y)\, dx\, dy.$$

The function $f : \mathbb{R}^2 \to \mathbb{R}^+$ *above, if it exists, will be the density of the pair* (X, Y)*.*

All properties seen in the discrete case remain true for pairs with a density: notably, if the pair consists of independent square-integrable random variables, $\operatorname{cov}(X,Y) = 0$ *and* $\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y)$*.*

As in the discrete case, the covariance of two random variables whose pair admits a density can be zero without the variables being independent; for example, consider the uniform distribution on the disk centered at 0 *with radius* 1*.*

2.14.6 Distribution of the Sum

It can be verified that if the pair (X, Y) has a density, the random variable $X + Y$ also has a density, and this density can be explicitly written:

Proposition 2.24. *1. If the pair* (X, Y) *admits the density function* f*, then the density of the random variable* $X + Y$ *is the function* g *defined by*

$$g(z) = \int_{\mathbb{R}} f(x, z-x)\, dx = \int_{\mathbb{R}} f(z-y, y)\, dy.$$

2. If X *and* Y *are independent random variables with densities* f_X *and* f_Y*, the density* g *of* $X + Y$ *is the convolution of* f_X *and* f_Y*:*

$$g(z) = \int_{\mathbb{R}} f_X(x) f_Y(z-x)\, dx = \int_{\mathbb{R}} f_X(z-y) f_Y(y)\, dy.$$

Proof. Let $h : \mathbb{R} \to \mathbb{R}$ be a piecewise continuous and bounded function. Then

$$E[h(X+Y)] = \int_{\mathbb{R}^2} h(x+y) f(x,y)\, dx\, dy = \int_{\mathbb{R}} h(z) \left(\int_{\mathbb{R}} f(x, z-x)\, dx \right) dz,$$

by setting $z = x+y$ and keeping x as the other variable. If instead one had used the change of variables $z = x+y$ and y, one would have obtained the other expression announced. □

7.1 Vector-valued Random Variables

Definition 2.20 (Random Vector)**.** *Let $(\Omega, \mathcal{T}, \mathbb{P})$ be a probability space and let $X_1, \dots, X_d$ be real-valued random variables defined on Ω. The mapping from Ω into $\mathbb{R}^d$ defined by*

$$\omega \longmapsto \begin{pmatrix} X_1(\omega) \\ \vdots \\ X_d(\omega) \end{pmatrix}$$

is called a random vector*.*

The notions of expectation, covariance, and characteristic function extend naturally to random vectors. The expectation of the random vector $(X_1, \dots, X_d)$ is now the vector

$$\mathbb{E}(X) = \Big(\mathbb{E}(X_1), \dots, \mathbb{E}(X_d)\Big).$$

Its covariance matrix, sometimes denoted by Σ, is a $d \times d$ matrix defined by

$$V(X) = \Big(\mathrm{Cov}(X_i, X_j)\Big)_{1 \le i,j \le d}.$$

The characteristic function of X is defined on $\mathbb{R}^d$ by

$$\forall u \in \mathbb{R}^d, \qquad \varphi_X(u) = \mathbb{E}\Big(e^{iu \cdot X}\Big) = \mathbb{E}\left(\prod_{k=1}^{d} e^{iu_k X_k}\right).$$

In what follows, we use the notation

$$u \cdot X = \sum_{i=1}^{d} u_i X_i, \qquad u \in \mathbb{R}^d.$$

The vectors u and X are considered as column vectors.

Proposition 2.25. *Let X be a random vector with values in $\mathbb{R}^d$ and let $u = (u_1, \dots, u_d) \in \mathbb{R}^d$. Then*

$$\mathbb{E}(u \cdot X) = u \cdot \mathbb{E}(X), \qquad V(u \cdot X) = u^{\mathsf{T}} V(X)\, u.$$

Proof. These formulas follow directly from the properties of expectation and variance. Indeed,

$$\mathbb{E}(u \cdot X) = \mathbb{E}\left(\sum_{i=1}^{d} u_i X_i\right) = \sum_{i=1}^{d} u_i \mathbb{E}(X_i) = u \cdot \mathbb{E}(X).$$

Moreover,

$$V(u \cdot X) = V\left(\sum_{i=1}^{d} u_i X_i\right) = \sum_{i,j=1}^{d} u_i u_j \mathrm{Cov}(X_i, X_j) = u^{\mathsf{T}} V(X)\, u.$$

□

Note that the mapping $u \mapsto V(u \cdot X)$ defines a quadratic form on $\mathbb{R}^d$. This quadratic form is positive, since for all $u \in \mathbb{R}^d$,

$$u^{\mathsf{T}} V(X)\, u = V(u \cdot X) \ge 0.$$

Most notions concerning random variables admit a natural analogue for random vectors. In a previous chapter, we explained how to define the distribution of a tuple of random variables. Since a random vector is a tuple, these definitions apply directly here.

The distribution (law) of a random vector $X = (X_1, \dots, X_n)$ is therefore a probability measure defined on the Borel σ-algebra of $\mathbb{R}^d$ by

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A), \qquad \text{for all Borel sets } A \subset \mathbb{R}^d.$$

One speaks of *discrete* random vectors or of random vectors *with a density* according to whether this distribution is discrete or absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$.

We also say that two random vectors $X = (X_1, \dots, X_d)$ and $Y = (Y_1, \dots, Y_d)$ are independent if

$$\mathbb{P}_{(X,Y)} = \mathbb{P}_X \otimes \mathbb{P}_Y,$$

that is,

$$\mathbb{P}(X_1, \dots, X_d, Y_1, \dots, Y_d) = \mathbb{P}(X_1, \dots, X_d) \otimes \mathbb{P}(Y_1, \dots, Y_d).$$

The notion of integrability also extends naturally to random vectors. A random vector X is said to be integrable if

$$\mathbb{E}(\|X\|) < \infty,$$

where $\|\cdot\|$ denotes a norm on $\mathbb{R}^d$, for example the Euclidean norm. It is said to be square-integrable if

$$\mathbb{E}(\|X\|^2) < \infty,$$

and similarly for higher moments.

2.15 Gaussian Vectors

We now consider the generalization of the normal distribution to the multidimensional case.

Definition 2.21 (Gaussian Vector)**.** *A random vector $(X_1, \dots, X_d)$ is said to be* Gaussian *if, for all $u_1, \dots, u_d \in \mathbb{R}$, the linear combination*

$$u_1 X_1 + \cdots + u_d X_d$$

follows a normal distribution or is almost surely constant.

We adopt the convention that the Dirac measure δ_m is regarded as a normal distribution with zero variance and mean equal to $m \in \mathbb{R}$. With this convention, a constant random variable is considered to follow a normal distribution with zero standard deviation.

Note that the components X_i of a Gaussian vector are normally distributed. Indeed, it suffices to take all coefficients u_i equal to zero except one.

Let us give a first example of a Gaussian vector.

Proposition 2.26. *Let $X_1, \dots, X_d$ be independent random variables, each following a normal distribution. Then the vector $(X_1, \dots, X_d)$ is Gaussian and its covariance matrix is diagonal.*

Lemma 2.1. *Let $a, b, c \in \mathbb{R}$ and let Y_1, Y_2 be two independent random variables following the standard normal distribution, that is,*

$$\mathbb{E}(Y_1) = \mathbb{E}(Y_2) = 0, \qquad V(Y_1) = V(Y_2) = 1.$$

Then the random variable

$$aY_1 + bY_2 + c$$

follows a normal distribution with mean c and variance $a^2 + b^2$.

Proof. Let us define

$$Z = aY_1 + bY_2 + c,$$

and let $g : \mathbb{R} \to \mathbb{R}$ be a bounded measurable function. Then

$$\mathbb{E}\Big(g(Z)\Big) = \iint g(ay_1 + by_2 + c)\, d\mathbb{P}_{Y_1}(y_1)\, d\mathbb{P}_{Y_2}(y_2).$$

Since Y_1 and Y_2 are independent standard normal random variables, we have

$$\mathbb{E}\Big(g(Z)\Big) = \frac{1}{2\pi} \iint g(ay_1 + by_2 + c)\, e^{-\frac{1}{2}(y_1^2+y_2^2)}\, dy_1 dy_2.$$

We now perform the change of variables

$$\begin{cases} z_1 = ay_1 + by_2 + c, \\ z_2 = -by_1 + ay_2. \end{cases}$$

A direct computation yields

$$(z_1 - c)^2 + z_2^2 = (a^2 + b^2)(y_1^2 + y_2^2), \qquad dz_1\, dz_2 = (a^2 + b^2)\, dy_1 dy_2.$$

Consequently,

$$\mathbb{E}\Big(g(Z)\Big) = \frac{1}{2\pi(a^2 + b^2)} \iint g(z_1)\, \exp\left(-\frac{(z_1 - c)^2 + z_2^2}{2(a^2 + b^2)}\right) dz_2 dz_1.$$

Integrating with respect to z_2, we obtain

$$\mathbb{E}\Big(g(Z)\Big) = \frac{1}{\sqrt{2\pi(a^2 + b^2)}} \int g(z_1)\, \exp\left(-\frac{(z_1 - c)^2}{2(a^2 + b^2)}\right) dz_1.$$

This is precisely the expectation with respect to a normal distribution with mean c and variance $a^2 + b^2$. Hence, Z follows a normal distribution with parameters $(c, a^2 + b^2)$. □

Proof. (**Proof of the Proposition 2.26**)

We first consider the case of two random variables X_1 and X_2 with means m_1, m_2 and variances σ_1^2, σ_2^2, assumed to be nonzero. We normalize these variables by setting

$$Y_i = \frac{X_i - m_i}{\sigma_i}, \qquad i = 1, 2.$$

By the lemma, any linear combination of Y_1 and Y_2 follows a normal distribution. Therefore, the same holds for

$$u_1 X_1 + u_2 X_2 = u_1 m_1 + u_2 m_2 + u_1 \sigma_1 Y_1 + u_2 \sigma_2 Y_2.$$

The case of n random variables follows immediately by induction on n. This completes the proof of the proposition. □

2.15.1 Affine Transformations of Gaussian Vectors

New Gaussian vectors can be constructed by applying affine transformations to Gaussian vectors.

Proposition 2.27. *Let $X = (X_1, \dots, X_d)$ be a Gaussian vector, let A be a $d' \times d$ matrix, and let B be a vector in $\mathbb{R}^{d'}$. Then the vector $AX + B$ is Gaussian with values in $\mathbb{R}^{d'}$, and*

$$\mathbb{E}(AX + B) = A\,\mathbb{E}(X) + B, \qquad V(AX + B) = A\,V(X)\,A^{\mathsf{T}}.$$

Proof. Any linear combination of the coordinates of the vector $AX + B$ is a linear combination of the coordinates of X and of the constant random variable equal to 1. Hence, it follows a normal distribution.

Let $a_{i,j}$ denote the coefficients of the matrix A and b_i those of the vector B. Then

$$\mathbb{E}\Big((AX + B)_i\Big) = \mathbb{E}\left(\sum_j a_{i,j}X_j + b_i\right) = \sum_j a_{i,j}\mathbb{E}(X_j) + b_i = (A\,\mathbb{E}(X) + B)_i,$$

and

$$\mathrm{Cov}\Big((AX + B)_i, (AX + B)_j\Big) = \sum_{k,l} a_{i,k}a_{j,l}\mathrm{Cov}(X_k, X_l).$$

This proves the stated formulas for the expectation and covariance matrix. □

This quantity is precisely the (i, j)-entry of the matrix $A\,V(X)\,A^{\mathsf{T}}$. The proposition is therefore proved.

Distribution of Gaussian Vectors

We shall show that the distribution of a Gaussian vector depends only on its mean vector and its covariance matrix, and we shall explicitly determine its density. We begin by computing the characteristic function of a Gaussian vector.

Recall that the characteristic function of a normal distribution with parameters m and σ^2 is given by

$$\varphi(t) = e^{itm - \frac{1}{2}\sigma^2 t^2}.$$

Proposition 2.28. *Let $X = (X_1, \dots, X_d)$ be a Gaussian vector with mean vector m and covariance matrix Σ. Then*

$$\varphi_X(u) = e^{iu\cdot m - \frac{1}{2}u^{\mathsf{T}}\Sigma u}, \qquad u \in \mathbb{R}^d.$$

Proof. We know that the random variable $u \cdot X$ follows a normal distribution, and we have already computed its expectation and variance:

$$\mathbb{E}(u \cdot X) = u \cdot \mathbb{E}(X) = u \cdot m, \qquad V(u \cdot X) = u^{\mathsf{T}}V(X)u = u^{\mathsf{T}}\Sigma u.$$

It follows that, for all $t \in \mathbb{R}$,

$$\mathbb{E}\Big(e^{it\,u\cdot X}\Big) = e^{it\,u\cdot m - \frac{1}{2}t^2\,u^{\mathsf{T}}\Sigma u}.$$

The result is obtained by taking $t = 1$. □

As in the case of a real-valued random variable, one can show that the characteristic function of a random vector uniquely determines its distribution. Hence, the distribution of a Gaussian vector is completely determined by its mean vector m and its covariance matrix Σ.

In order to compute the density of a Gaussian vector, we shall need some properties of symmetric matrices. Recall that a symmetric matrix Σ is said to be *positive semidefinite* if

$$u^\mathsf{T}\Sigma u \geq 0 \quad \text{for all } u \in \mathbb{R}^d,$$

and *positive definite* if

$$u^\mathsf{T}\Sigma u > 0 \quad \text{for all } u \in \mathbb{R}^d \setminus \{0\}.$$

A symmetric positive matrix is positive definite if and only if it is invertible, that is, if and only if its determinant is nonzero.

Theorem 2.6. *Every Gaussian vector X has the same distribution as a Gaussian vector of the form $AY + B$, where Y is a Gaussian vector whose components are independent, identically distributed, and follow the standard normal distribution $\mathcal{N}(0, 1)$.*

Proof. The proof relies on the following result: any symmetric positive matrix S can be written in the form

$$S = TDT^\mathsf{T},$$

where T is an invertible matrix and D is a diagonal matrix whose entries are equal to 0 or 1. The matrix T can be obtained by applying the Gaussian reduction algorithm to the quadratic form $u \mapsto u^\mathsf{T}Su$. Alternatively, it can be constructed by diagonalizing S in an orthonormal basis.

We take as symmetric matrix the covariance matrix $V(X)$, which is indeed positive, since

$$u^\mathsf{T}V(X)u = V(u \cdot X) \geq 0.$$

Let $Y = (Y_1, \ldots, Y_d)$ be a Gaussian vector whose components are independent, centered, and whose covariance matrix is the identity. Define

$$Z = TDY + \mathbb{E}(X),$$

and compute its expectation and covariance matrix.

We have

$$\mathbb{E}(Z) = TD\,\mathbb{E}(Y) + \mathbb{E}(X) = \mathbb{E}(X),$$

and

$$V(Z) = V(TDY) = TD\,V(Y)\,D^\mathsf{T}T^\mathsf{T} = TDD^\mathsf{T}T^\mathsf{T} = TDT^\mathsf{T} = V(X).$$

The vectors Z and X are Gaussian and have the same expectation and the same covariance matrix. Hence, they have the same characteristic function and therefore the same distribution. This completes the proof. □

Note that the previous proof shows that any symmetric positive matrix is the covariance matrix of a Gaussian vector.

We are now in a position to determine the density of Gaussian vectors whose covariance matrix is positive definite, or equivalently, invertible, or, in other words, has nonzero determinant. Such Gaussian vectors are said to be *non-degenerate*.

Theorem 2.7. *Let Σ be a symmetric positive definite $d \times d$ matrix and let $m \in \mathbb{R}^d$. The random vector $X = (X_1, \dots, X_d)$ with density*

$$f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \exp\Big(-\frac{1}{2}(x-m)^{\mathsf{T}}\Sigma^{-1}(x-m)\Big), \qquad x \in \mathbb{R}^d,$$

is a Gaussian vector with mean m and covariance matrix Σ.

Conversely, any Gaussian vector whose covariance matrix Σ has nonzero determinant admits the above function as its density.

Proof. We use the decompositions

$$X = TY + \mathbb{E}(X), \qquad \Sigma = TDT^{\mathsf{T}},$$

introduced in the previous proposition. Since Σ is positive definite, the matrix D is equal to the identity matrix.

Let $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, denote $dx = dx_1 \cdots dx_d$, and set $m = \mathbb{E}(X)$. Then, for any bounded measurable function g,

$$\mathbb{E}\Big(g(X)\Big) = \mathbb{E}\Big(g(TY + \mathbb{E}(X))\Big) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} g(Ty+m) \exp\Big(-\frac{1}{2}y^{\mathsf{T}}y\Big)\, dy.$$

□

$$= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} g(x) \exp\Big(-\frac{1}{2}(x-m)^{\mathsf{T}}(TT^{\mathsf{T}})^{-1}(x-m)\Big) \Big|\det(dx_T)\Big|.$$

This follows from the change of variables $x = Ty + m$, for which

$$dx = |\det(T)|\, dy.$$

It suffices to note that

$$\det(T)^2 = \det(\Sigma)$$

in order to conclude.

As an application of the previous results, we obtain the following criterion concerning the independence of the components of a Gaussian vector.

Corollary 2.3. *Let X be a Gaussian vector. Its components $(X_1, \dots, X_d)$ are independent if and only if the covariance matrix is diagonal, that is,*

$$\mathrm{Cov}(X_i, X_j) = 0 \qquad \textit{for all } i \neq j.$$

Indeed, if the covariance matrix is diagonal, the vector X has the same distribution as a vector whose coordinates are functions of the corresponding coordinates of a vector Y whose components are independent.

Chapter 3

Limit theorems (Inequalities and convergence types in probability)

3.1 Probability Inequalities

We have already used several types of inequalities. In this chapter, we provide a more systematic description of the inequalities and bounds commonly used in probability and statistics [9–11, 11, 14, 17].

3.1.1 Boole's Inequality and Bonferroni Inequalities

Boole's inequality (also known as the union bound) states that for any countable collection of events, the probability that at least one of the events occurs is no greater than the sum of the probabilities of the individual events.

Proposition 3.1 (Boole's Inequality)**.** *Suppose* $(\Omega, \mathcal{F}, P)$ *is a probability space, and let* $E_1, E_2, \cdots \in \mathcal{F}$ *be events. Then*

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} P(E_i).$$

Proof. We give a proof for a finite collection of events using mathematical induction.

Base case: For $n = 1$, we have

$$P(E_1) \leq P(E_1),$$

which is trivially true.

Inductive step: Assume Boole's inequality holds for n events, i.e.,

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} P(E_i).$$

Consider $n+1$ events $E_1, \ldots, E_n, E_{n+1}$. Then

$$\begin{aligned} P\left(\bigcup_{i=1}^{n+1} E_i\right) &= P\left(\left(\bigcup_{i=1}^{n} E_i\right) \cup E_{n+1}\right) \\ &= P\Big(\bigcup_{i=1}^{n} E_i\Big) + P(E_{n+1}) - P\Big(\big(\bigcup_{i=1}^{n} E_i\big) \cap E_{n+1}\Big) \\ &\leq P\Big(\bigcup_{i=1}^{n} E_i\Big) + P(E_{n+1}) \\ &\leq \sum_{i=1}^{n} P(E_i) + P(E_{n+1}) \\ &= \sum_{i=1}^{n+1} P(E_i), \end{aligned}$$

where we used the induction hypothesis in the second inequality. □

One of the interpretations of Boole's inequality is what is known as σ-sub-additivity in measure theory applied here to the probability measure P.

Boole's inequality can be extended to get lower and upper bounds on the probability of unions of events, known as Bonferroni inequalities. As before, suppose $(\Omega, \mathcal{F}, P)$ is a probability space, and $E_1, E_2, \ldots, E_n \in \mathcal{F}$ are events. Define

$$S_1 := \sum_{i=1}^{n} P(E_i), \quad S_2 := \sum_{1\leq i<j\leq n} P(E_i \cap E_j), \quad S_k := \sum_{1\leq i_1<\cdots<i_k\leq n} P(E_{i_1} \cap \cdots \cap E_{i_k}), \quad k = 3, \ldots, n.$$

Proposition 3.2 (Bonferroni inequalities)**.** *For odd k in $1, \ldots, n$,*

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{j=1}^{k} (-1)^{j-1} S_j,$$

and for even k in $2, \ldots, n$,

$$P\left(\bigcup_{i=1}^{n} E_i\right) \geq \sum_{j=1}^{k} (-1)^{j-1} S_j.$$

We omit the proof, which starts with considering the case $k = 1$, for which we need to show

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{j=1}^{1} (-1)^{j-1} S_j = S_1 = \sum_{i=1}^{n} P(E_i),$$

which is Boole's inequality. When $k = 2$,

$$P\left(\bigcup_{i=1}^{n} E_i\right) \geq \sum_{j=1}^{2} (-1)^{j-1} S_j = S_1 - S_2 = \sum_{i=1}^{n} P(E_i) - \sum_{1\leq i<j\leq n} P(E_i \cap E_j),$$

which for $n = 2$ is the inclusion-exclusion identity

Example 3.1. *Suppose we place n distinguishable balls into m distinguishable boxes at random ($n > m$). Let E be the event that a box is empty. The sample space can be described as*

$$\Omega = \{\omega = (\omega_1, \ldots, \omega_n) : 1 \leq \omega_i \leq m\}$$

with $P(\omega) = \frac{1}{m^n}$. *Denote*

$$E_l := \{\omega : \omega_i \neq l \text{ for all } i = 1, \ldots, n\}, \quad l = 1, 2, \ldots, m.$$

Then,

$$E = E_1 \cup \cdots \cup E_{m-1},$$

since E_m *is empty and including it or not does not change the result.*

We can see that for any k*,*

$$P(E_{i_1} \cup \cdots \cup E_{i_k}) = \left(\frac{m-k}{m}\right)^n.$$

Then we can use the inclusion-exclusion principle to get

$$P(E) = m\left(1 - \frac{1}{m}\right)^n - \binom{m}{2}\left(1 - \frac{2}{m}\right)^n + \cdots + (-1)^{m-2}\binom{m}{m-1}\left(1 - \frac{m-1}{m}\right)^n.$$

The last term is zero, since all boxes cannot be empty. The expression is quite complicated. But if we use Bonferroni inequalities we see that

$$m\left(1 - \frac{1}{m}\right)^n - \binom{m}{2}\left(1 - \frac{2}{m}\right)^n \leq P(E) \leq m\left(1 - \frac{1}{m}\right)^n.$$

This gives a good estimate when n *is large compared to* m*. For example, if* $m = 10$ *then*

$$10 \cdot (0.9)^n - 45 \cdot (0.8)^n \leq P(E) \leq 10 \cdot (0.9)^n.$$

In particular, for $n = 50$*, then* $45 \cdot (0.8)^{50} = 0.00064226146$*, which is the difference between the left and right sides of the estimates. This gives a rather good estimate.*

3.1.2 Markov's Inequality

Proposition 3.3 (Markov's inequality)**.** *Suppose* X *is a nonnegative random variable. Then, for any* $a > 0$*, we have*

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

Proof. We only give the proof for a continuous random variable; the case of a discrete random variable is similar. Suppose X is a positive continuous random variable, then we can write

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx\, \mathbf{1}_{\{X>0\}} = \int_0^{\infty} x f_X(x)\, dx.$$

For any $a > 0$, we have

$$\int_0^{\infty} x f_X(x)\, dx \geq \int_a^{\infty} x f_X(x)\, dx \geq \int_a^{\infty} a f_X(x)\, dx = a \int_a^{\infty} f_X(x)\, dx = aP(X > a).$$

Therefore,

$$aP(X > a) \leq \mathbb{E}[X],$$

which is what we wanted to prove. □

Example 3.2. *First, we observe that Boole's inequality can be interpreted as expectations of the number of occurred events. Suppose* $(S, \mathcal{F}, P)$ *is a probability space, and* $E_1, E_2, \cdots \in \mathcal{F}$ *are events. Define ...*

$$X_i := \begin{cases} 1, & \text{if } E_i \text{ occurs,} \\ 0, & \text{otherwise.} \end{cases}$$

Then

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

is the number of events that occur. Therefore,

$$\mathbb{E}[X] = P(E_1) + \cdots + P(E_n).$$

Now we would like to prove Boole's inequality using Markov's inequality. Note that X is a nonnegative random variable, so we can apply Markov's inequality. For $a = 1$ we get

$$P(X > 1) \leq \mathbb{E}[X] = P(E_1) + \cdots + P(E_n).$$

Finally, the event $X > 1$ means that at least one of the events $E_1, E_2, \ldots, E_n$ occurs, so

$$P\Big(\bigcup_{i=1}^{n} E_i\Big) = P(X > 0) \leq \mathbb{E}[X] = P(E_1) + \cdots + P(E_n),$$

which completes the proof.

Example 3.3. *Suppose $X \sim Binom(n, p)$. We would like to use Markov's inequality to find an upper bound on $P(X > qn)$ for $p < q < 1$. Note that X is a nonnegative random variable and $\mathbb{E}[X] = np$. By Markov's inequality, we have*

$$P(X > qn) \leq \frac{\mathbb{E}[X]}{qn} = \frac{np}{qn} = \frac{p}{q}.$$

3.1.3 Chebyshev's Inequality

This inequality shows that the difference between a random variable and its expectation is controlled by its variance. Informally, it shows how far the random variable is from its mean on average.

Proposition 3.4 (Chebyshev's inequality)**.** *Suppose X is a random variable. Then for any $b > 0$ we have*

$$P\Big(|X - \mathbb{E}[X]| > b\Big) \leq \frac{\mathrm{Var}(X)}{b^2}.$$

Proof. Define

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

Then Y is a nonnegative random variable, and we can apply Markov's inequalityin proposition 3.3 to Y. Then for $b > 0$ we have

$$P\Big(|X - \mathbb{E}[X]| > b\Big) = P(Y > b^2) \leq \frac{\mathbb{E}[Y]}{b^2} = \frac{\mathrm{Var}(X)}{b^2}.$$

□

$$P\left(Y > b^2\right) \leq \frac{\mathbb{E}[Y]}{b^2}.$$

Note that

$$\mathbb{E}[Y] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathrm{Var}(X),$$
$$P\left(Y > b^2\right) = P\left((X - \mathbb{E}[X])^2 > b^2\right) = P(|X - \mathbb{E}[X]| > b),$$

which completes the proof.

Example 3.4. *Consider again $X \sim Binom(n, p)$. We now use Chebyshev's inequality to find an upper bound on $P(X > qn)$ for $p < q < 1$.*

Recall that $\mathbb{E}[X] = np$. By Chebyshev's inequality with $b = (q - p)n > 0$ we have

$$\begin{aligned} P(X > qn) &= P(X - np > (q-p)n) \leq P(|X - np| > (q-p)n) \qquad (3.1) \\ &\leq \frac{\mathrm{Var}(X)}{((q-p)n)^2} = \frac{np(1-p)}{((q-p)n)^2} = \frac{p(1-p)}{(q-p)^2 n}. \end{aligned}$$

3.1.4 Chernoff bounds

Proposition 3.5 (Chernoff bounds)**.** *Suppose X is a random variable and denote by $m_X(t)$ its moment generating function. Then for any $a \in \mathbb{R}$,*

$$P(X > a) \leq \min_{t>0} e^{-ta} m_X(t), \quad P(X \leq a) \leq \min_{t<0} e^{-ta} m_X(t).$$

Proof.

$$P(X > a) = P\left(e^{tX} > e^{ta}\right), \quad t > 0,$$
$$P(X \leq a) = P\left(e^{tX} > e^{ta}\right), \quad t < 0.$$

Note that e^{tX} is a positive random variable for any $t \in \mathbb{R}$. Therefore, we can apply Markov's inequality to see that

$$P(X > a) \leq e^{-ta}\mathbb{E}[e^{tX}], \quad t > 0,$$
$$P(X \leq a) \leq e^{-ta}\mathbb{E}[e^{tX}], \quad t < 0.$$

Recall that $\mathbb{E}[e^{tX}]$ is the moment generating function $m_X(t)$, and so we have

$$P(X > a) \leq \frac{m_X(t)}{e^{ta}}, \quad t > 0,$$
$$P(X \leq a) \leq \frac{m_X(t)}{e^{ta}}, \quad t < 0.$$

Taking the minimum over appropriate t gives the result. □

Example 3.5. *Consider again $X \sim \mathrm{Binom}(n, p)$. We now use Chernoff bounds for $P(X > qn)$ for $p < q < 1$. Recall that the moment generating function for X is as follows:*

$$m_X(t) = (pe^t + (1-p))^n.$$

Thus a Chernoff bound gives

$$P(X > qn) \leq \min_{t>0} e^{-tqn}(pe^t + (1-p))^n.$$

To find the minimum of $g(t) = e^{-tqn}(pe^t + (1-p))^n$, *we can take its derivative. Using the only critical point of this function, we can see that the minimum on* $(0, \infty)$ *is achieved at* t^* *such that*

$$t^* = \frac{q(1-p)}{(1-q)p},$$

and so

$$\begin{aligned} g(t^*) &= \left(\frac{q(1-p)}{(1-q)p}\right)^{-qn} \left(\frac{p}{q(1-p)/(1-q)p} + (1-p)\right)^n \\ &= \left(\frac{q(1-p)}{(1-q)p}\right)^{-qn} \left(\frac{1-p}{1-q}\right)^n \\ &= \left(\frac{p}{q}\right)^{qn} \left(\frac{1-p}{1-q}\right)^{-qn} \left(\frac{1-p}{1-q}\right)^n \\ &= \left(\frac{p}{q}\right)^{qn} \left(\frac{1-p}{1-q}\right)^{(1-q)n}. \end{aligned} \tag{3.2}$$

Thus the Chernoff bound gives

$$P(X > qn) \leq \left(\frac{p}{q}\right)^{qn} \left(\frac{1-p}{1-q}\right)^{(1-q)n}.$$

Example 3.6. *(Comparison of Markov's, Chebyshev's inequalities and Chernoff bounds). These three inequalities for the binomial random variable* $X \sim \text{Binom}(n, p)$ *give:*

- *Markov's inequality:*
$$P(X > qn) \leq \frac{p}{q},$$
- *Chebyshev's inequality:*
$$P(X > qn) \leq \frac{p(1-p)}{(q-p)^2 n},$$
- *Chernoff bound:*
$$P(X > qn) \leq \left(\frac{p}{q}\right)^{qn} \left(\frac{1-p}{1-q}\right)^{(1-q)n}.$$

Clearly the right-hand sides are very different: Markov's inequality gives a bound independent of n, *and the Chernoff bound is the strongest with exponential convergence to* 0 *as* $n \to \infty$.

For example, for $p = 1/2$ *and* $q = 3/4$ *we have*

$$\textit{Markov's inequality: } P\left(X > \frac{3n}{4}\right) \leq \frac{2}{3}, \quad \textit{Chebyshev's inequality: } P\left(X > \frac{3n}{4}\right) \leq \frac{4}{n},$$

$$\textit{Chernoff bound: } P\left(X > \frac{3n}{4}\right) \leq \left(\frac{16}{27}\right)^{n/4}.$$

For example, for $p = 1/3$ *and* $q = 2/3$ *we have... Markov's inequality:*

$$P\left(X > \frac{3n}{4}\right) \leq \frac{1}{2},$$

Chebyshev's inequality:

$$P\left(X > \frac{3n}{4}\right) \leq \frac{2}{n},$$

Chernoff bound:

$$P\left(X > \frac{3n}{4}\right) \leq 2^{-n/2}.$$

3.1.5 Cauchy-Schwarz inequality

This inequality appears in a number of areas of mathematics including linear algebra. We will apply it to give a different proof for the bound for correlation coefficients. Note that the Cauchy-Schwarz inequality can be easily generalized to random vectors X and Y.

Proposition 3.6 (Cauchy-Schwarz inequality)**.** *Suppose X and Y are two random variables, then*

$$(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2] \cdot \mathbb{E}[Y^2],$$

and the equality holds if and only if $X = aY$ for some constant $a \in \mathbb{R}$.

Proof. Define the random variable

$$U := (X - sY)^2$$

which is a nonnegative random variable for any $s \in \mathbb{R}$. Then

$$0 \leq EU = \mathbb{E}[(X - sY)^2] = \mathbb{E}[X^2] - 2s\mathbb{E}[XY] + s^2\mathbb{E}[Y^2].$$

Define

$$g(s) := \mathbb{E}[X^2] - 2s\mathbb{E}[XY] + s^2\mathbb{E}[Y^2],$$

which is a quadratic polynomial in s. What we know is that $g(s)$ is nonnegative for all s. Completing the square we see that

$$g(s) = \mathbb{E}[Y^2]s^2 - 2\mathbb{E}[XY]s + \mathbb{E}[X^2] = \left(\sqrt{\mathbb{E}[Y^2]}s - \frac{\mathbb{E}[XY]}{\sqrt{\mathbb{E}[Y^2]}}\right)^2 + \mathbb{E}[X^2] - \frac{(\mathbb{E}[XY])^2}{\mathbb{E}[Y^2]},$$

so $g(s) \geq 0$ for all s if and only if

$$\mathbb{E}[X^2] - \frac{(\mathbb{E}[XY])^2}{\mathbb{E}[Y^2]} \geq 0,$$

which is what we needed to show.

To deal with the last claim, observe that if $U = 0$ with probability one, then $g(s) = EU = 0$. This happens only if

$$\mathbb{E}[X^2] - \frac{(\mathbb{E}[XY])^2}{\mathbb{E}[Y^2]} = 0.$$

And if

$$\mathbb{E}[X^2] - \frac{(\mathbb{E}[XY])^2}{\mathbb{E}[Y^2]} = 0,$$

then

$$X - \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}Y = 0,$$

that is, X is a scalar multiple of Y. □

Example 3.7. *We can use the Cauchy-Schwarz inequality to prove one of the properties of the correlation coefficient. Namely, suppose X and Y are random variables, then $|\rho(X,Y)| \leq 1$. Moreover, $|\rho(X,Y)| = 1$ if and only if there are constants $a, b \in \mathbb{R}$ such that $X = a + bY$.*

We will use normalized random variables as before, namely,

$$U := \frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}}, \quad V := \frac{Y - \mathbb{E}[Y]}{\sqrt{\mathrm{Var}(Y)}}.$$

Then $\mathbb{E}[U] = \mathbb{E}[V] = 0$, $\mathbb{E}[U^2] = \mathbb{E}[V^2] = 1$. We can use the Cauchy-Schwarz inequality for U and V to see that

$$|\mathbb{E}[UV]| \leq \sqrt{\mathbb{E}[U^2] \cdot \mathbb{E}[V^2]} = 1,$$

and the equality holds if and only if $U = aV$ for some $a \in \mathbb{R}$.

We have

$$\rho(X,Y) = \mathbb{E}[UV],$$

which gives the bound we need. Note that if $U = aV$, then

$$\frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}} = a\frac{Y - \mathbb{E}[Y]}{\sqrt{\mathrm{Var}(Y)}},$$

therefore

$$X = a\frac{\sqrt{\mathrm{Var}(X)}}{\sqrt{\mathrm{Var}(Y)}}(Y - \mathbb{E}[Y]) + \mathbb{E}[X] = a\frac{\sqrt{\mathrm{Var}(X)}}{\sqrt{\mathrm{Var}(Y)}}Y - a\frac{\sqrt{\mathrm{Var}(X)}}{\sqrt{\mathrm{Var}(Y)}}\mathbb{E}[Y] + \mathbb{E}[X],$$

which completes the proof.

3.1.6 Jensen's Inequality

Recall that a function $g : \mathbb{R} \to \mathbb{R}$ is *convex* on $[a,b]$ if for each $x, y \in [a,b]$ and each $\lambda \in [0,1]$ we have

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y).$$

Note that for a convex function g, this property holds for any convex linear combination of points in $[a,b]$, that is,

$$g(\lambda_1 x_1 + \cdots + \lambda_n x_n) \leq \lambda_1 g(x_1) + \cdots + \lambda_n g(x_n),$$

where $\lambda_1 + \cdots + \lambda_n = 1$, $0 \leq \lambda_1, \ldots, \lambda_n \leq 1$, and $x_1, \ldots, x_n \in [a,b]$.

If g is twice differentiable, then we have a simple test to see if a function is convex, namely, g is convex if $g''(x) > 0$ for all $x \in [a,b]$. Geometrically, one can show that if g is convex, then if we draw a line segment between any two points on the graph of the function, the entire segment lies above the graph of g. A function g is *concave* if $-g$ is convex. Typical examples of convex functions are $g(x) = x^2$ and $g(x) = e^x$. Examples of concave functions are $g(x) = -x^2$ and $g(x) = \log x$. Convex and concave functions are always continuous.

Convex functions lie above tangents

Suppose $a < c < b$ and $g : [a, b] \to \mathbb{R}$ is convex. Then there exist $A, B \in \mathbb{R}$ such that

$$g(c) = Ac + B \quad \text{and for all } x \in [a, b], \quad g(x) \geq Ax + B.$$

Proof. For $a \leq x < c < y \leq b$, we can write c as a convex combination of x and y, namely,

$$c = \lambda x + (1 - \lambda)y \quad \text{with} \quad \lambda = \frac{y - c}{y - x} \in [0, 1].$$

Therefore,

$$g(c) \leq \lambda g(x) + (1 - \lambda)g(y),$$

which implies that

$$\frac{g(c) - g(x)}{c - x} \leq \frac{g(y) - g(c)}{y - c}.$$

Thus we can take

$$\sup_{x<c} \frac{g(c) - g(x)}{c - x} \leq A \leq \inf_{y>c} \frac{g(y) - g(c)}{y - c}.$$

so that we have for all $x < y$ in $[a, b]$ that

$$g(x) \geq A(x - c) + g(c) = Ax + (g(c) - Ac).$$

□

Proposition 3.7 (Jensen's inequality)**.** *Suppose X is a random variable such that $P(a \leq X \leq b) = 1$. If $g : \mathbb{R} \to \mathbb{R}$ is convex on $[a, b]$, then*

$$E[g(X)] \geq g(EX).$$

If g is concave, then

$$E[g(X)] \leq g(EX).$$

Proof. If X is constant, then there is nothing to prove, so assume X is not constant. Then we have

$$a < EX < b.$$

Denote $c := EX$. Then

$$g(x) \geq Ax + B \quad \text{and} \quad g(EX) = AEX + B$$

for some $A, B \in \mathbb{R}$. Also note that

$$|g(X)| \leq |A||X| + |B| \leq |A| \max\{|a|, |b|\}|X| + |B|,$$

so $E|g(X)| < \infty$ and therefore $E[g(X)]$ is well defined. Now we can use $AX + B \leq g(X)$ to see that

$$g(EX) = AEX + B = E[AX + B] \leq E[g(X)].$$

Example 3.8 (Arithmetic-geometric mean inequality)**.** *Suppose $a_1, \dots, a_n$ are positive numbers, and X is a discrete random variable with the mass function*

$$f_X(a_k) = \frac{1}{n}, \quad k = 1, \dots, n.$$

Note that the function $g(x) = -\log x$ is convex on $(0, \infty)$. Jensen's inequality gives that

$$-\log\left(\frac{1}{n}\sum_{k=1}^{n} a_k\right) = -\log(EX) \le E[-\log X] = -\frac{1}{n}\sum_{k=1}^{n}\log a_k.$$

Exponentiating this we get

$$\frac{1}{n}\sum_{k=1}^{n} a_k \ge \left(\prod_{k=1}^{n} a_k\right)^{1/n}.$$

Example 3.9. *Suppose $p > 1$, then the function $g(x) = |x|^p$ is convex. Then*

$$E[|X|^p] \ge |E[X]|^p$$

for any random variable X such that $E[X]$ is defined. In particular,

$$E[X^2] \ge (E[X])^2,$$

and therefore

$$E[X^2] - (E[X])^2 \ge 0.$$

3.1.7 Independence of σ-algebras and the Borel-Cantelli Lemma

Definition 3.1. *Two σ-algebras $\mathcal{T}_1 \subset \mathcal{T}$, $\mathcal{T}_2 \subset \mathcal{T}$ are said to be independent if for every $A \in \mathcal{T}_1$ and $B \in \mathcal{T}_2$, A and B are independent.*

Let $(\mathcal{T}_i)_{i\in I}$ be a family of σ-algebras included in $\mathcal{T}$. They are said to be independent if for every finite subset $S \subset I$ and any family $(A_i)_{i\in S}$ with $A_i \in \mathcal{T}_i$ for all $i \in S$, the events $(A_i)_{i\in S}$ are mutually independent:

$$P\left(\bigcap_{i\in S} A_i\right) = \prod_{i\in S} P(A_i).$$

Borel-Cantelli Lemma

This is a first application of the notion of independent events.

Lemma 3.1 (Borel-Cantelli)**.** *Let $(\Omega, \mathcal{T}, P)$ be a probability space and $(A_i)_{i\in\mathbb{N}}$ a sequence of events.*

1. *If $\sum_{i\in\mathbb{N}} P(A_i) < +\infty$, then almost every $\omega \in \Omega$ belongs to only finitely many A_i.*
2. *If $\sum_{i\in\mathbb{N}} P(A_i) = +\infty$ and the A_i are mutually independent, then almost every $\omega \in \Omega$ belongs to infinitely many A_i.*

The limit superior of the sequence (A_i) is defined as:

$$\limsup_{i\in\mathbb{N}} A_i = \bigcap_{N\in\mathbb{N}} \bigcup_{i\ge N} A_i = \{\omega \in \Omega \mid \omega \text{ belongs to infinitely many } A_i\}.$$

The lemma can then be restated as:

$$\sum_{i\in\mathbb{N}} P(A_i) < +\infty \implies P\left(\limsup_{i\in\mathbb{N}} A_i\right) = 0,$$

$$\sum_{i\in\mathbb{N}} P(A_i) = +\infty \text{ and } (A_i)_{i\in\mathbb{N}} \text{ independent} \implies P\left(\limsup_{i\in\mathbb{N}} A_i\right) = 1.$$

Proof of the Lemma

We have the relation

$$\#\{i \in \mathbb{N} \mid \omega \in A_i\} = \sum_{i \in \mathbb{N}} \mathbf{1}_{A_i}(\omega).$$

Integrating this equality gives:

$$\int \#\{i \in \mathbb{N} \mid \omega \in A_i\}\, dP(\omega) = \int \sum_{i \in \mathbb{N}} \mathbf{1}_{A_i}\, dP = \sum_{i \in \mathbb{N}} P(A_i) < +\infty.$$

The function $\omega \mapsto \#\{i \in \mathbb{N} \mid \omega \in A_i\}$ is integrable, hence finite almost everywhere; for almost every $\omega \in \Omega$, $\#\{i \in \mathbb{N} \mid \omega \in A_i\} < +\infty$.

Now assume that the (A_i) are independent and let $M, N \in \mathbb{N}$ with $N \leq M$:

$$P\left(\bigcap_{i=N}^{M} A_i^c\right) = \prod_{i=N}^{M} P(A_i^c) = \prod_{i=N}^{M} (1 - P(A_i)) \leq \exp\left(-\sum_{i=N}^{M} P(A_i)\right),$$

using the inequality $1 - x \leq e^{-x}$, valid for all $x \in \mathbb{R}$.

Hence,

$$P\left(\bigcap_{i \geq N} A_i^c\right) \leq P\left(\bigcap_{i=N}^{M} A_i^c\right) \leq \exp\left(-\sum_{i=N}^{M} P(A_i)\right),$$

for all $M \geq N$. Taking the limit $M \to \infty$, we obtain:

$$P\left(\bigcap_{i \geq N} A_i^c\right) = 0.$$

This implies:

$$P\left(\bigcup_{N \in \mathbb{N}} \bigcap_{i \geq N} A_i\right) = 1,$$

i.e.,

$$P(\limsup A_i) = 1.$$

As a corollary, consider a sequence (X_i) of independent random variables such that $P(X_i = 1) = P(X_i = -1) = 1/2$ for all i. From the above, almost surely there exist arbitrarily long consecutive sequences of 1 in (X_i), whose sum is therefore unbounded.

3.2 Exercises

Exercise 3.1. *Suppose $X \sim Exp(\lambda)$. Using Markov's inequality estimate $P(X > a)$ for $a > 0$ and compare it with the exact value of this probability.*

Solution Markov's inequality gives

$$P(X > a) \leq \frac{EX}{a} = \frac{1}{a\lambda},$$

while the exact value is

$$P(X > a) = \int_a^{\infty} \lambda e^{-\lambda x} dx = e^{-\lambda a} \leq \frac{1}{a\lambda}.$$

Exercise 3.2. *Suppose $X \sim Exp(\lambda)$. Using Chebyshev's inequality estimate $P(|X - E[X]| > b)$ for $b > 0$.*

Solution We have $EX = 1/\lambda$ and $\mathrm{Var}(X) = 1/\lambda^2$. By Chebyshev's inequality we have

$$P(|X - EX| > b) \leq \frac{\mathrm{Var}(X)}{b^2} = \frac{1}{b^2\lambda^2}.$$

Exercise 3.3. *Suppose $X \sim Exp(\lambda)$. Using Chernoff bounds estimate $P(X > a)$ for $a > E[X]$ and compare it with the exact value of this probability.*

Solution Recall first that the moment generating function is

$$m_X(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

Using Chernoff bounds, we have

$$P(X > a) \leq \min_{t>0} e^{-ta} m_X(t) = \min_{t>0} e^{-ta} \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

To find the minimum, the critical point is $t = \lambda - 1/a > 0$ since $a > EX = 1/\lambda$. Using this t, we get

$$e^{-\lambda a} = P(X > a) \leq a\lambda e^{1-\lambda a} = (a\lambda e)\, e^{-\lambda a} = (a\lambda e) P(X > a),$$

where $a\lambda e > 1$.

Exercise 3.4. *Suppose $X > 0$ is a random variable such that $Var(X) > 0$. Decide which of the two quantities is larger.*

Solution

(A) $EX^3 > (EX)^3$ since $(x^3)'' = 6x > 0$ for $x > 0$.

(B) $EX^{3/2} > (EX)^{3/2}$ since $(x^{3/2})'' = \frac{3}{4\sqrt{x}} > 0$ for $x > 0$.

(C) $EX^{2/3} < (EX)^{2/3}$ since $(x^{2/3})'' = -\frac{2}{9x^{4/3}} < 0$ for $x > 0$.

(D) $E[\log(X + 1)] < \log(EX + 1)$ since $(\log(x))'' = -1/x^2 < 0$ for $x > 0$.

(E) $E[e^X] > e^{EX}$ since $(e^x)'' = e^x > 0$ for any x.

(F) $E[e^{-X}] > e^{-EX}$ since $(e^{-x})'' = e^{-x} > 0$ for any x.

3.3 Different Types of Convergence

The previous results involve various notions of convergence. We will clarify these notions and study the relationships between them. Let us first recall the definition of L^p norms for $p \geq 1$.

Let $(\Omega, \mathcal{T}, P)$ be a probability space. For $p \in [1, +\infty)$, the L^p norm of a random variable $Y : \Omega \to \mathbb{R}$ is defined by

$$\|Y\|_p = E(|Y|^p)^{1/p} = \left(\int |Y|^p \, dP \right)^{1/p}.$$

The L^∞ norm of Y is defined by

$$\|Y\|_\infty = \inf\Big\{ C > 0 \mid \exists \Omega' \subset \Omega,\ P(\Omega') = 1,\ \text{and } |Y(\omega)| \leq C \text{ for all } \omega \in \Omega' \Big\}.$$

Definition 3.2. *Let (Y_n) and Y be random variables defined on $(\Omega, \mathcal{T}, P)$ and let $p \in [1, +\infty]$.*

- *The sequence (Y_n) converges in L^p norm to Y if*
$$\|Y_n - Y\|_p \xrightarrow[n\to\infty]{} 0.$$
- *The sequence (Y_n) converges in probability to Y if*
$$\forall \varepsilon > 0, \quad P(|Y_n - Y| > \varepsilon) \xrightarrow[n\to\infty]{} 0.$$
- *The sequence (Y_n) converges almost surely to Y if*
$$Y_n(\omega) \xrightarrow[n\to\infty]{} Y(\omega) \quad \text{for almost all } \omega \in \Omega.$$
- *The sequence (Y_n) converges in distribution (in law) to Y if*
$$\forall f : \mathbb{R} \to \mathbb{R} \text{ continuous and bounded}, \quad \int f \, dP_{Y_n} \xrightarrow[n\to\infty]{} \int f \, dP_Y.$$
or
$$\lim_{n\longrightarrow\infty} F_{Y_n}(y) = F_Y(y),$$
where F_{Y_n} and F_Y are respectively the distribution functions of Y_n and Y.

Proposition 3.8. *Let $p, q \in \mathbb{R}$ such that $1 \le p \le q \le \infty$. The following implications hold:*

$$CV \text{ in } L^\infty \implies CV \text{ in } L^q \implies CV \text{ in } L^p \implies CV \text{ in } L^1 \implies CV \text{ in probability} \implies CV \text{ in law},$$
$$CV \text{ in } L^\infty \implies CV \text{ almost surely} \implies CV \text{ in probability},$$
$$CV \text{ in } L^\infty \implies CV \text{ in probability} \implies CV \text{ almost surely for a subsequence.}$$

Remark 3.1. *Convergence in L^2 implies convergence in probability. This is how we proved the weak law of large numbers. It states that S_n/n converges to $E(X_0)$ in probability if the (X_i) are independent and identically distributed. We obtained this result by showing that*
$$V\Big(S_n/n\Big) \xrightarrow[n\to\infty]{} 0.$$
According to the following relation, this is equivalent to L^2 convergence:
$$V(S_n/n) = E\Big[(S_n/n - E[S_n/n])^2\Big] = E\Big[(S_n/n - E(X_0))^2\Big] = \|S_n/n - E(X_0)\|_2^2.$$

Proof of the proposition

- **L^q convergence implies L^p convergence for $p \le q$:** We prove the inequality $\|Y\|_p \le \|Y\|_q$ using Hölder's inequality: for all $p, q \ge 1$ such that $1/p + 1/q = 1$,
$$\int |YZ| \, dP \le \|Y\|_p \, \|Z\|_q.$$
Take Y constant equal to 1, then $\|Y\|_p = 1$ and $\|Z\|_1 \le \|Z\|_q$. This proves the result for $p = 1$. For general p, replace q by q/p and Z by $|Y|^p$, which gives
$$\int |Y|^p \, dP \le \left(\int |Y|^q \, dP \right)^{p/q}, \quad \|Y\|_p \le \|Y\|_q.$$

- **L^∞ convergence implies L^p convergence:** For almost every $\omega \in \Omega$, $|Y(\omega)| \leq \|Y\|_\infty$. Integrating, we get

$$\|Y\|_p^p = \int |Y(\omega)|^p \, dP(\omega) \leq \int \|Y\|_\infty^p \, dP = \|Y\|_\infty^p.$$

- **L^1 convergence implies convergence in probability:** This is a consequence of Markov's inequality. If $Y_n \xrightarrow[n\to\infty]{L^1} Y$, then

$$P(|Y_n - Y| > \varepsilon) \leq \frac{E(|Y_n - Y|)}{\varepsilon} = \frac{\|Y_n - Y\|_1}{\varepsilon} \xrightarrow[n\to\infty]{} 0.$$

- **L^∞ convergence implies almost sure convergence:** If $Y_n \xrightarrow[n\to\infty]{L^\infty} Y$, there exists $\Omega' \subset \Omega$ with probability 1 such that

$$\sup_{\omega\in\Omega'} |Y_n(\omega) - Y(\omega)| \xrightarrow[n\to\infty]{} 0.$$

  It follows that for every $\omega \in \Omega'$,

$$Y_n(\omega) \xrightarrow[n\to\infty]{} Y(\omega).$$

- **Convergence in probability implies almost sure convergence along a subsequence:** We know that for every $\varepsilon > 0$,

$$P(|Y_n - Y| > \varepsilon) \xrightarrow[n\to\infty]{} 0.$$

  For each $k \in \mathbb{N}$, one can choose $n_k \in \mathbb{N}$ large enough such that

$$P(|Y_{n_k} - Y| > 1/k) \leq \frac{1}{2^k}.$$

  Then,

$$\sum_{k=0}^{\infty} P(|Y_{n_k} - Y| > 1/k) < \infty.$$

  By the Borel-Cantelli lemma, for almost every $\omega \in \Omega$, except for finitely many indices k, we have

$$|Y_{n_k}(\omega) - Y(\omega)| < 1/k.$$

  Hence, the subsequence (Y_{n_k}) converges to Y almost surely.

- **Almost sure convergence implies convergence in probability:** We have the following two conditions:
  - $1_{\{|Y_n - Y|>\varepsilon\}}(\omega) \xrightarrow[n\to\infty]{} 0$ for almost every $\omega \in \Omega$, since $|Y_n(\omega) - Y(\omega)| \xrightarrow[n\to\infty]{} 0$.
  - $|1_{\{|Y_n - Y|>\varepsilon\}}| \leq 1_\Omega$, and 1_Ω is integrable and independent of n.

  Applying the dominated convergence theorem, we get

$$\lim_{n\to\infty} P(|Y_n - Y| > \varepsilon) = \lim_{n\to\infty} \int 1_{\{|Y_n - Y|>\varepsilon\}} \, dP = \int \lim_{n\to\infty} 1_{\{|Y_n - Y|>\varepsilon\}} \, dP = 0.$$

  The implication *convergence in probability* $\implies$ *convergence in law* will be proved later.

3.4 Convergence in Law and Characteristic Functions

Recall that Y_n converges in law to Y if for every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$\int f \, dP_{Y_n} \longrightarrow_{n\to\infty} \int f \, dP_Y.$$

Definition 3.3. *Let μ_n and μ be probability measures defined on the Borel σ-algebra of $\mathbb{R}$. We say that μ_n converges weakly to μ if for every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,*

$$\int f \, d\mu_n \longrightarrow_{n\to\infty} \int f \, d\mu.$$

The sequence Y_n converges in law to Y if and only if P_{Y_n} converges weakly to P_Y.

We now relate convergence in law to pointwise convergence of characteristic functions, which will be useful in proving the Central Limit Theorem.

Theorem 3.1. *Let μ, μ_n, $n \in \mathbb{N}$, be probability measures defined on the Borel σ-algebra of $\mathbb{R}$. The following properties are equivalent:*

- *$\int f \, d\mu_n \longrightarrow_{n\to\infty} \int f \, d\mu$ for every function f of the form $f(x) = e^{itx}$, $t \in \mathbb{R}$,*
- *$\int f \, d\mu_n \longrightarrow_{n\to\infty} \int f \, d\mu$ for every C^∞ function f with compact support,*
- *$\int f \, d\mu_n \longrightarrow_{n\to\infty} \int f \, d\mu$ for every bounded continuous function f.*

The first point corresponds to the convergence of the Fourier transforms of μ_n, while the last point corresponds to the weak convergence of μ_n to μ.

From this, we deduce the following result, called *Lévy's convergence theorem*:

Theorem 3.2 (Lévy)**.** *Let μ, μ_n be probability measures defined on the Borel σ-algebra of $\mathbb{R}$. If for every $t \in \mathbb{R}$,*

$$\hat{\mu}_n(t) \longrightarrow_{n\to\infty} \hat{\mu}(t),$$

then μ_n converges weakly to μ.

Let Y and Y_n, $n \in \mathbb{N}$, be random variables defined on a probability space $(\Omega, \mathcal{T}, P)$. If for every $t \in \mathbb{R}$,

$$\varphi_{Y_n}(t) \longrightarrow_{n\to\infty} \varphi_Y(t),$$

then Y_n converges in law to Y.

3.5 Reminder on Compact Support and Fourier Inversion

Recall that a function $f : \mathbb{R} \to \mathbb{R}$ has *compact support* if there exists $A > 0$ such that f is zero outside $[-A, A]$. An example of a C^∞ function with compact support is

$$f(x) = e^{-\frac{1}{1-x^2}} \mathbf{1}_{[-1,1]}(x).$$

To prove the previous theorems, we need the Fourier inversion formula. Let $f : \mathbb{R} \to \mathbb{R}$ be a function integrable with respect to the Lebesgue measure. Its Fourier transform is defined by

$$\hat{f}(t) = \int_{\mathbb{R}} e^{-itx} f(x) \, dx.$$

One can show that this function is continuous by applying the theorem on continuity under the integral sign.

Theorem 3.3. *(Fourier Inversion Formula)*
Let f be a C^∞ function with compact support. Then $\hat{f}$ is integrable and

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{itx}\hat{f}(t)\,dt \quad \text{for all } x \in \mathbb{R}.$$

Corollary 3.1. *Let μ be a probability measure defined on the Borel σ-algebra of $\mathbb{R}$, and let $f : \mathbb{R} \to \mathbb{C}$ be a C^∞ function with compact support. Then*

$$\int_{\mathbb{R}} f(x)\,d\mu(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(t)\hat{\mu}(t)\,dt.$$

Proof.

$$\begin{aligned}\int_{\mathbb{R}} f(x)\,d\mu(x) &= \int_{\mathbb{R}} \frac{1}{2\pi}\int_{\mathbb{R}} e^{itx}\hat{f}(t)\,dt\,d\mu(x)\\ &= \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{itx}\hat{f}(t)\,d\mu(x)\,dt\\ &= \frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(t)\left(\int_{\mathbb{R}} e^{itx}\,d\mu(x)\right)dt\\ &= \frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(t)\hat{\mu}(t)\,dt.\end{aligned}$$

Here we used Fubini's theorem to interchange the two integrals. This is justified since

$$\int_{\mathbb{R}}\int_{\mathbb{R}} \left|e^{itx}\hat{f}(t)\right| d\mu(x)\,dt = \int_{\mathbb{R}} d\mu(x)\int_{\mathbb{R}} |\hat{f}(t)|\,dt = \int_{\mathbb{R}} |\hat{f}(t)|\,dt < \infty.$$

Thus the proof of the corollary is complete. □

Proof. (**Proof of Theorem 3.1**)
Let us show that

$$\int f\,d\mu_n \to \int f\,d\mu$$

for every C^∞ function f with compact support if $\hat{\mu}_n(t) \to \hat{\mu}(t)$ for all $t \in \mathbb{R}$. By the previous corollary,

$$\int f(x)\,d\mu_n(x) = \frac{1}{2\pi}\int \hat{f}(t)\hat{\mu}_n(t)\,dt, \quad \int f(x)\,d\mu(x) = \frac{1}{2\pi}\int \hat{f}(t)\hat{\mu}(t)\,dt.$$

Applying the dominated convergence theorem gives

$$\int \hat{f}(t)\hat{\mu}_n(t)\,dt \longrightarrow \int \hat{f}(t)\hat{\mu}(t)\,dt.$$

The use of the dominated convergence theorem is justified here because for all $t \in \mathbb{R}$, $\hat{\mu}_n(t) \to \hat{\mu}(t)$ by hypothesis, and $\hat{f}\hat{\mu}_n$ is bounded by $\hat{f}$ which is integrable. The first part of Theorem 3.1 is thus proved.

We now aim to show that if

$$\int f\,d\mu_n \to \int f\,d\mu$$

for every C^∞ function with compact support, then the same holds for every bounded continuous function. We can reduce to the case where f is real-valued by decomposing it into its real and imaginary parts. Consider the set C of bounded Borel functions for which

$$\int f\,d\mu_n \to \int f\,d\mu :$$

$$C = \{f : \mathbb{R} \to \mathbb{R} \text{ Borel bounded } \mid \int_{\mathbb{R}} f\,d\mu_n \to \int_{\mathbb{R}} f\,d\mu\}.$$

□

Lemma 3.2. *Let f be a bounded Borel function such that for every $\varepsilon > 0$, there exist $g, h \in C$ satisfying $g \le f \le h$ and*

$$\int (h - g)\, d\mu < \varepsilon.$$

Then $f \in C$.

Proof. Let $\varepsilon > 0$. We have

$$\int g\, d\mu \le \int h\, d\mu \quad \text{and} \quad \int h\, d\mu - \int g\, d\mu < \varepsilon.$$

Since $\int g\, d\mu_n \to \int g\, d\mu$, there exists $N \in \mathbb{N}$ such that for all $n \ge N$,

$$\int g\, d\mu_n > \int f\, d\mu - \varepsilon.$$

Similarly, there exists $N' \in \mathbb{N}$ such that for all $n \ge N'$,

$$\int h\, d\mu_n < \int f\, d\mu + \varepsilon.$$

Hence, for $n \ge \max(N, N')$,

$$\int f\, d\mu - \varepsilon < \int g\, d\mu_n \le \int f\, d\mu_n \le \int h\, d\mu_n < \int f\, d\mu + \varepsilon.$$

This proves the lemma. □

Proof. **(End of the Proof of Theorem 3.1**) Let us show that every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$ belongs to C. Let $\varepsilon > 0$. By monotone convergence,

$$\mu([-A, A]) \to \mu\Big(\bigcup_{A \in \mathbb{N}} [-A, A] \Big) = \mu(\mathbb{R}) = 1 \quad \text{as } A \to \infty.$$

Choose $A \in \mathbb{N}$ such that $\mu([-A, A]) > 1 - \varepsilon$. By the Stone-Weierstrass theorem, we can find a polynomial P uniformly close to f on $[-A - 1, A + 1]$:

$$\forall x \in [-A - 1, A + 1], \quad P(x) - \varepsilon \le f(x) \le P(x) + \varepsilon.$$

Consider a C^∞ function φ that vanishes outside $[-A-1, A+1]$, equals 1 on $[-A, A]$, and satisfies $0 \le \varphi \le 1$. We show that $f \in C$ by applying Lemma 4 with the functions

$$g = \varphi(P - \varepsilon) + (1 - \varphi)(-\|f\|_\infty), \quad h = \varphi(P + \varepsilon) + (1 - \varphi)\|f\|_\infty.$$

Let us check the assumptions of Lemma 3.2. Since C is a vector space that contains constants and C^∞ functions with compact support, g and h indeed belong to C. Moreover,

$$\int (h - g)\, d\mu \le 2 \int (1 - \varphi)\|f\|_\infty + \varepsilon\, d\mu \le 2\|f\|_\infty\, \mu([-A, A]^c) + 2\varepsilon \le 2(\|f\|_\infty + 1)\varepsilon.$$

Let us show the inequality $g \le f$. For all $x \in \mathbb{R}$,

$$\varphi(P - \varepsilon) \le \varphi f, \quad -\|f\|_\infty \le f,$$

hence

$$g = \varphi(P - \varepsilon) + (1 - \varphi)(-\|f\|_\infty) \le \varphi f + (1 - \varphi) f = f.$$

A similar calculation shows that $f \le h$. Theorem 3.1 is thus proved. □

Corollary 3.2. *Let $(X_n)_{n\in\mathbb{N}}$ and X be random variables defined on a probability space $(\Omega, \mathcal{T}, P)$ such that X_n converges in law to X. Then for all $a, b \in \mathbb{R}$ such that $P(X = a) = P(X = b) = 0$,*

$$P(a \le X_n \le b) \longrightarrow P(a \le X \le b) \quad \text{as } n \to \infty.$$

Moreover, the distribution functions of X_n converge to the distribution function of X at all points $x \in \mathbb{R}$ such that $P(X = x) = 0$:

$$F_{X_n}(x) \longrightarrow F_X(x) \quad \text{if } P(X = x) = 0.$$

Remark 3.2. *It can be shown that the convergence of distribution functions at all points x such that $P(X = x) = 0$ is actually equivalent to the convergence in law of the sequence X_n to X, by sandwiching any continuous function between step functions whose discontinuities form a negligible set.*

Proposition 3.9. *Let X_n and X be random variables. If X_n converges to X in probability, then X_n converges to X in law.*

Proof. Let $f : \mathbb{R} \to \mathbb{R}$ be a C^1 function with bounded derivative. By the mean value theorem, for all $x, y \in \mathbb{R}$,

$$|f(x) - f(y)| \le \sup_{\mathbb{R}} |f'| \, |x - y|.$$

We show that the difference $\int f \, dP_{X_n} - \int f \, dP_X$ tends to 0 as $n \to \infty$:

$$\begin{aligned}
\left| \int f \, dP_{X_n} - \int f \, dP_X \right| &= \left| \int f(X_n) \, dP - \int f(X) \, dP \right| \\
&\le \int |f(X_n) - f(X)| \, dP \\
&\le \int_{|X_n - X| > \delta} |f(X_n) - f(X)| \, dP + \int_{|X_n - X| \le \delta} |f(X_n) - f(X)| \, dP \\
&\le 2 \sup_{\mathbb{R}} |f| \, P(|X_n - X| > \delta) + \sup_{\mathbb{R}} |f'| \, \delta.
\end{aligned}$$

Since X_n converges to X in probability, $P(|X_n - X| > \delta) \to 0$ as $n \to \infty$. For any $\varepsilon > 0$, choose δ such that $\sup |f'|\delta < \varepsilon/2$. Then there exists $N \in \mathbb{N}$ such that for all $n \ge N$,

$$P(|X_n - X| > \delta) \le \frac{\varepsilon}{4 \sup_{\mathbb{R}} |f|},$$

which implies

$$\left| \int f \, dP_{X_n} - \int f \, dP_X \right| < \varepsilon.$$

This proves the proposition. □

Proposition 3.10. *Two random variables that have the same characteristic function have the same law.*

Proof. Let μ and ν be the laws of the random variables, and consider the constant sequence $\mu_n = \nu$. The sequence $\int f \, d\mu_n$ is constant and equal to $\int f \, d\nu$, so the convergences in Theorem 3.1 become equalities.

Thus, there is an equivalence between the equality

$$\int f \, d\nu = \int f \, d\mu$$

for every function of the form $f(x) = e^{itx}$ and the same equality for every bounded continuous function f.

It follows that the two measures are equal as soon as they have the same characteristic function. □

3.6 Weak and Strong Laws of Large Numbers

We are interested in the asymptotic behavior of sequences of random variables. Let $(\Omega, \mathcal{T}, P)$ be a probability space, and for each integer $n \in \mathbb{N}$, let $X_n : \Omega \to \mathbb{R}$ be a random variable.

Definition 3.4. *A sequence of random variables $(X_i)_{i\in\mathbb{N}}$ is said to be identically distributed if all X_i have the same distribution:*

$$\forall i, j \in \mathbb{N}, \quad P_{X_i} = P_{X_j}.$$

In other words, for any Borel set $A \subset \mathbb{R}$,

$$P(X_i \in A) = P(X_j \in A),$$

and for any Borel function $f : \mathbb{R} \to \mathbb{R}$, positive or integrable with respect to P_{X_0},

$$E(f(X_i)) = E(f(X_j)).$$

In particular, if the X_i are integrable, then $E(X_i) = E(X_j)$; if they are square-integrable, then $E(X_i^2) = E(X_j^2)$ and $V(X_i) = V(X_j)$.

3.6.1 Weak Law of Large Numbers

Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of independent and identically distributed (i.i.d.) random variables. Define

$$S_n = \sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n.$$

For $\omega \in \Omega$, the quantity

$$\frac{S_n(\omega)}{n} = \frac{X_1(\omega) + X_2(\omega) + \cdots + X_n(\omega)}{n}$$

is the empirical mean calculated from the sample corresponding to the outcome $\omega \in \Omega$.

We aim to study the asymptotic behavior of the sample mean $\frac{S_n}{n}$.

Theorem 3.4 (Weak Law of Large Numbers)**.** *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of independent and identically distributed (i.i.d.) random variables with finite second moments. Then, for any $\varepsilon > 0$,*

$$P\left(\left|\frac{S_n}{n} - E(X_0)\right| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0,$$

where $S_n = \sum_{i=1}^{n} X_i$.

The proof of this theorem relies on the following lemma:

Lemma 3.3. *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables. Then*

$$E(S_n) = nE(X_0), \quad V(S_n) = nV(X_0).$$

Proof of the Lemma. By linearity of expectation,

$$E(S_n) = E(X_1 + X_2 + \cdots + X_n) = E(X_1) + \cdots + E(X_n) = nE(X_0).$$

For the variance, we have

$$V(S_n) = V(X_1 + \cdots + X_n) = V(X_1) + \cdots + V(X_n) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j),$$

where $\mathrm{Cov}(X_i, X_j) = E(X_iX_j) - E(X_i)E(X_j)$. Since the variables are independent, $\mathrm{Cov}(X_i, X_j) = 0$, and therefore

$$V(S_n) = V(X_1) + \cdots + V(X_n) = nV(X_0).$$

□

Proof of the Theorem. From the lemma,

$$E\left(\frac{S_n}{n}\right) = E(X_0), \quad V\left(\frac{S_n}{n}\right) = \frac{V(X_0)}{n}, \quad \sigma\left(\frac{S_n}{n}\right) = \frac{\sigma(X_0)}{\sqrt{n}}.$$

Applying Chebyshev's inequality, we obtain

$$P\left(\left|\frac{S_n}{n} - E(X_0)\right| > \varepsilon\right) \leq \frac{V(X_0)}{n\varepsilon^2} \xrightarrow[n\to\infty]{} 0.$$

□

Remark 3.3. *The weak law of large numbers also holds for integrable, independent, identically distributed random variables.*

3.6.2 Strong Law of Large Numbers

Theorem 3.5 (Strong Law of Large Numbers)**.** *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of integrable, independent, identically distributed random variables, and define $S_n = \sum_{i=1}^n X_i$. Then, for almost every $\omega \in \Omega$,*

$$\frac{S_n(\omega)}{n} \xrightarrow[n\to\infty]{} E(X_0).$$

In other words, the set

$$\left\{\omega \in \Omega \,\middle|\, \frac{S_n(\omega)}{n} \to E(X_0) \text{ as } n \to \infty\right\}$$

has probability 1.

We will prove the strong law of large numbers using the following lemma.

Lemma 3.4. *Let (Y_i) be a sequence of random variables. If for all $\varepsilon > 0$,*

$$\sum_{i=1}^{\infty} P(|Y_i| > \varepsilon) < \infty,$$

then, the sequence $(Y_i)_{i\in\mathbb{N}}$ converges almost surely to 0:

$$\text{for almost every } \omega \in \Omega, \quad Y_i(\omega) \xrightarrow[i\to\infty]{} 0.$$

The lemma states that

$$P\Big(\{\omega \in \Omega \mid Y_i(\omega) \xrightarrow[i\to\infty]{} 0\}\Big) = 1.$$

Proof. We apply the Borel-Cantelli lemma. Let $\varepsilon > 0$ be fixed and define

$$A_i = \{|Y_i| > \varepsilon\}.$$

Since $\sum P(A_i) < \infty$, almost every $\omega \in \Omega$ belongs to only finitely many A_i. Denote this set by C_ε. Then $P(C_\varepsilon) = 1$.

For every $\omega \in C_\varepsilon$, there exists $N \in \mathbb{N}$ such that for all $n \geq N$, $\omega \notin A_n$ and $|Y_n(\omega)| < \varepsilon$.

Now take $\varepsilon = 1/k$, $k \in \mathbb{N}^*$, and consider the intersection

$$C = \bigcap_{k\in\mathbb{N}^*} C_{1/k}, \quad P(C) = 1.$$

For every $\omega \in C$ and every $k \in \mathbb{N}^*$, $\omega \in C_{1/k}$, so there exists $N \in \mathbb{N}$ such that for all $n \geq N$, $|Y_n(\omega)| < 1/k$. This shows that

$$\lim_{n\to\infty} Y_n(\omega) = 0,$$

as desired. □

Proof. (**Proof of the Strong Law of Large Numbers)** For simplicity, we assume that the X_i are square-integrable in this proof. We have $E(X_i) = E(X_1)$ for all i. By replacing X_i with $X_i - E(X_i)$, we can assume $E(X_i) = 0$. We say that the random variables are centered. We want to show that

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

Let us try to apply the previous lemma. Recall that $E(S_i) = E(X_1) = 0$. By Chebyshev's inequality,

$$P\left(\frac{S_i}{i} > \varepsilon\right) \leq \frac{V(S_i)}{i^2} = \frac{V(X_1)}{i}.$$

Then

$$\sum_{i=1}^{\infty} P\left(\frac{S_i}{i} > \varepsilon\right) \leq V(X_1) \sum_{i=1}^{\infty} \frac{1}{i} = +\infty.$$

Hence, the condition of the lemma, with $Y_i = S_i/i$, is not satisfied. We replace i by i^2 and set $Y_i = S_{i^2}/i^2$. Then we have

$$\sum_{i=1}^{\infty} P\left(\frac{S_{i^2}}{i^2} > \varepsilon\right) \leq V(X_1) \sum_{i=1}^{\infty} \frac{1}{i^2} < \infty.$$

The series $\sum_{i=1}^{\infty} \frac{1}{i^2}$ is convergent (its sum equals $\pi^2/6$). The previous lemma implies the convergence of the sequence (S_{i^2}/i^2):

$$\frac{S_{i^2}}{i^2} \xrightarrow[i\to\infty]{} 0 \quad \text{a.s.}$$

For each $n \in \mathbb{N}^*$, let $i \in \mathbb{N}$ be the largest integer such that

$$i^2 \leq n.$$

Then $i = \lfloor \sqrt{n} \rfloor$, and we have the bounds

$$i^2 \leq n \leq (i+1)^2 - 1, \qquad i^2 \leq n \leq i^2 + 2i, \qquad 0 \leq n - i^2 \leq 2i \leq 2\sqrt{n}.$$

We decompose

$$S_n = \sum_{k=1}^{n} X_k = \sum_{k=1}^{i^2} X_k + \sum_{k=i^2+1}^{n} X_k.$$

Hence,

$$\frac{S_n}{n} \leq \frac{S_{i^2}}{i^2} + \frac{1}{n} \sum_{k=i^2+1}^{n} X_k.$$

To bound the last term, we argue as before. For any $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n} \sum_{k=i^2+1}^{n} X_k > \varepsilon \right) \leq \frac{1}{n^2\varepsilon^2} \operatorname{Var}\left(\sum_{k=i^2+1}^{n} X_k \right) \leq \frac{n - i^2}{n^2\varepsilon^2} \operatorname{Var}(X_1) \leq \frac{2\sqrt{n}}{n^2\varepsilon^2} \operatorname{Var}(X_1).$$

Thus,

$$\mathbb{P}\left(\frac{1}{n} \sum_{k=i^2+1}^{n} X_k > \varepsilon \right) \leq \frac{C}{n^{3/2}},$$

for some constant $C > 0$. Since the series $\sum_{n=1}^{\infty} n^{-3/2}$ is convergent, the lemma implies that

$$\frac{1}{n} \sum_{k=i^2+1}^{n} X_k \xrightarrow[n\to\infty]{} 0 \quad \text{a.s.}$$

The result follows.

□

3.7 Central Limit Theorem

To prove the Central Limit Theorem, we will use the characterization of convergence in law via characteristic functions. We start by computing the characteristic function of the normal distribution.

3.7.1 Characteristic Function of the Normal Distribution

Theorem 3.6. *Let Y be a random variable following the standard normal distribution (mean $m = 0$, standard deviation $\sigma = 1$). Its density is given by*

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

and its characteristic function is

$$\varphi_Y(t) = e^{-t^2/2}.$$

Proof. By definition,

$$\varphi_Y(t) = \int_{\mathbb{R}} e^{ity} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy.$$

We know that

$$e^{ity} = \sum_{k=0}^{\infty} \frac{(ity)^k}{k!} \quad \text{for } y \in \mathbb{R}.$$

Substitute this series into the integral:

$$\int_{\mathbb{R}} e^{ity} e^{-y^2/2}\, dy = \int_{\mathbb{R}} \sum_{k=0}^{\infty} \frac{(it)^k y^k}{k!} e^{-y^2/2}\, dy = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \int_{\mathbb{R}} y^k e^{-y^2/2}\, dy.$$

To justify the interchange of the sum and the integral, we need to verify that

$$\int_{\mathbb{R}} \sum_{k=0}^{\infty} \frac{|(it)^k y^k|}{k!} e^{-y^2/2}\, dy < \infty,$$

which holds since

$$\int_{-\infty}^{+\infty} e^{-y^2/2} \sum_{k=0}^{\infty} \frac{|ty|^k}{k!}\, dy = \int_{-\infty}^{+\infty} e^{-y^2/2} e^{|ty|}\, dy < \infty.$$

We now need to compute

$$I_k = \int_{\mathbb{R}} y^k e^{-y^2/2}\, dy.$$

When k is odd, the function $y \mapsto y^k e^{-y^2/2}$ is odd, so its integral is zero:

$$I_{2l+1} = 0 \quad \text{for all } l \in \mathbb{N}.$$

For even k, $k = 2l$, we perform integration by parts to obtain the relation

$$I_{2l+2} = \int_{\mathbb{R}} y^{2l+1} y e^{-y^2/2}\, dy = \left[y^{2l+1}(-e^{-y^2/2}) \right]_{-\infty}^{+\infty} + \int_{\mathbb{R}} (2l+1) y^{2l} e^{-y^2/2}\, dy = (2l+1) I_{2l}.$$

We know that

$$I_0 = \int_{\mathbb{R}} e^{-y^2/2}\, dy = \sqrt{2\pi},$$

so that

$$I_{2l} = (2l-1)(2l-3)\cdots 3 \cdot 1 \cdot \sqrt{2\pi}.$$

Then,

$$\frac{I_{2l}}{(2l)!} = \frac{(2l-1)(2l-3)\cdots 1}{(2l)(2l-1)(2l-2)\cdots 1} \sqrt{2\pi} = \frac{1}{2 \cdot 4 \cdot \cdots \cdot 2l} \sqrt{2\pi} = \frac{1}{2^l l!} \sqrt{2\pi}.$$

We can now compute the characteristic function φ_Y:

$$\varphi_Y(t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{ity} e^{-y^2/2}\, dy = \sum_{l=0}^{\infty} \frac{(it)^{2l}}{(2l)!} I_{2l} = \sum_{l=0}^{\infty} \frac{(-1)^l t^{2l}}{2^l l!} = e^{-t^2/2}.$$

□

3.7.2 Central Limit Theorem

The following result is called the central limit theorem (CLT), a fundamental result in probability theory.

Theorem 3.7. *Let $(\Omega, \mathcal{T}, P)$ be a probability space, and let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed random variables with finite second moments and non-zero variance. Let*

$$S_n = \sum_{i=1}^{n} X_i.$$

Then the random variable

$$\frac{S_n - \mathbb{E}(X_0)}{\sigma(X_0)/\sqrt{n}} = \frac{\sqrt{n}}{\sigma(X_0)}\left(\frac{S_n}{n} - \mathbb{E}(X_0)\right)$$

converges in distribution to a standard normal variable with mean 0 and variance 1. In particular, for any interval $[a, b] \subset \mathbb{R}$,

$$P\left(a \leq \frac{\sqrt{n}}{\sigma(X_0)}\left(\frac{S_n}{n} - \mathbb{E}(X_0)\right) \leq b\right) \longrightarrow_{n\to\infty} \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx.$$

Remark 3.4. *The event above can be written as follows:*

$$\{a \leq \sigma(X_0)\sqrt{n}(S_n/n - \mathbb{E}(X_0)) \leq b\} = \{\sigma(X_0)\sqrt{n}(S_n/n - \mathbb{E}(X_0)) \in [a, b]\}$$

$$= \{\mathbb{E}(X_0) + \frac{a\,\sigma(X_0)}{\sqrt{n}} \leq S_n/n \leq \mathbb{E}(X_0) + \frac{b\,\sigma(X_0)}{\sqrt{n}}\}.$$

When n is large, the probability that S_n/n lies in the interval

$$\left[\mathbb{E}(X_0) - t\frac{\sigma(X_0)}{\sqrt{n}},\ \mathbb{E}(X_0) + t\frac{\sigma(X_0)}{\sqrt{n}}\right]$$

is approximately

$$\frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-x^2/2}\, dx.$$

- *For $t = 1.96$, $\frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-x^2/2}\, dx \simeq 0.95$.*
- *For $t = 2.58$, $\frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-x^2/2}\, dx \simeq 0.99$.*

Hence, there is approximately a 99% chance, when n is large, that the empirical mean S_n/n lies in the interval

$$\left[\mathbb{E}(X_0) - 2.58\frac{\sigma(X_0)}{\sqrt{n}},\ \mathbb{E}(X_0) + 2.58\frac{\sigma(X_0)}{\sqrt{n}}\right].$$

It is customary to denote the convergence in distribution of a sequence of random variables (Y_n) to the normal distribution with parameters m, σ by

$$Y_n \xrightarrow{law} \mathcal{N}(m, \sigma^2), \quad n \to \infty.$$

In the case where the X_i are independent and identically distributed with zero mean and unit standard deviation, the central limit theorem can be summarized as:

$$\frac{S_n}{\sqrt{n}} \xrightarrow{law} \mathcal{N}(0, 1), \quad n \to \infty.$$

Proof. (**Proof of Central Limit Theorem**) Without loss of generality, we can replace X_i by $X_i - \mathbb{E}(X_i)$, so we may assume that the X_i are centered: $\mathbb{E}(X_i) = 0$. By dividing by $\sigma(X_i)$, we can also assume that $\sigma(X_i) = 1$. We want to show that the law of $\frac{S_n}{\sqrt{n}}$ converges to the normal distribution. It is sufficient to show that

$$\varphi_{\frac{S_n}{\sqrt{n}}}(t) \xrightarrow[n\to\infty]{} e^{-t^2/2} \quad \text{for all } t \in \mathbb{R}.$$

We have

$$\begin{aligned}
\varphi_{\frac{S_n}{\sqrt{n}}}(t) &= \mathbb{E}\left[e^{it\frac{S_n}{\sqrt{n}}}\right] \\
&= \mathbb{E}\left[e^{i\frac{t}{\sqrt{n}}\sum_{k=1}^n X_k}\right] \\
&= \mathbb{E}\left[\prod_{k=1}^n e^{i\frac{t}{\sqrt{n}}X_k}\right] \\
&= \prod_{k=1}^n \mathbb{E}\left[e^{i\frac{t}{\sqrt{n}}X_k}\right] \quad \text{by independence} \\
&= \left(\mathbb{E}[e^{i\frac{t}{\sqrt{n}}X_0}]\right)^n \quad \text{since the } X_i \text{ are identically distributed} \\
&= \varphi_{X_0}\left(\frac{t}{\sqrt{n}}\right)^n.
\end{aligned}$$

To compute the limit as $n \to \infty$, we use a Taylor expansion. Since X_0 is square-integrable, φ_{X_0} is C^2, and we have

$$\varphi_{X_0}(t) = \mathbb{E}[e^{itX_0}], \quad \varphi'_{X_0}(t) = \mathbb{E}[iX_0e^{itX_0}], \quad \varphi''_{X_0}(t) = \mathbb{E}[-X_0^2e^{itX_0}],$$

with

$$\varphi_{X_0}(0) = 1, \quad \varphi'_{X_0}(0) = i\mathbb{E}[X_0] = 0, \quad \varphi''_{X_0}(0) = -\mathbb{E}[X_0^2] = -1.$$

By Taylor's formula,

$$\varphi_{X_0}(x) = 1 - \frac{x^2}{2} + x^2\varepsilon_0(x), \quad \text{with } \varepsilon_0(x) \to 0 \text{ as } x \to 0.$$

This implies

$$\varphi_{\frac{S_n}{\sqrt{n}}}(t) = \varphi_{X_0}\left(\frac{t}{\sqrt{n}}\right)^n = \left(1 - \frac{t^2}{2n} + \frac{t^2}{n}\varepsilon_0\left(\frac{t}{\sqrt{n}}\right)\right)^n.$$

Taking logarithms:

$$n\ln\left(1 - \frac{t^2}{2n} + \frac{t^2}{n}\varepsilon_0\left(\frac{t}{\sqrt{n}}\right)\right) = -\frac{t^2}{2} + \varepsilon_2\left(\frac{1}{\sqrt{n}}\right),$$

so that

$$\varphi_{\frac{S_n}{\sqrt{n}}}(t) = \left(1 - \frac{t^2}{2n} + \frac{t^2}{n}\varepsilon_0\left(\frac{t}{\sqrt{n}}\right)\right)^n \xrightarrow[n\to\infty]{} e^{-t^2/2}.$$

□

Numerical Illustration:

Consider the throw of a six-sided die, modeled by a random variable X_0 that follows a uniform distribution on the set $\{1, 2, 3, 4, 5, 6\}$:

$$\mathbb{P}(X_0 = k) = \frac{1}{6}, \quad k \in \{1, 2, 3, 4, 5, 6\}.$$

We repeat the throw n times, $n \in \mathbb{N}^*$, which is described by a sequence of independent random variables $X_1, \ldots, X_n$ having the same distribution as X_0. Let

$$S_n = \sum_{k=1}^{n} X_k.$$

We compute the frequency graph of S_n for any n recursively using the formula

$$\mathbb{P}(S_{n+1} = k) = \sum_{l} \mathbb{P}(X_{n+1} = l)\, \mathbb{P}(S_n = k - l),$$

where the sum runs over all possible values l of X_{n+1}.

If n is sufficiently large, the frequency graph should approach a Gaussian distribution once appropriately normalized.

3.7.3 Limit Theorems for Random Vectors

The classical limit theorems for real-valued random variables admit natural analogues in the case of random vectors.

When we speak of a limit theorem, we are interested in the asymptotic behavior of a sequence $\{X_k\}_{k\in\mathbb{N}}$ of random vectors in $\mathbb{R}^d$. To avoid any ambiguity, we denote in this section the components of the vector X_k by

$$X_k = \left(X_k^{(1)}, \ldots, X_k^{(d)}\right).$$

The notation X_k refers to a term of a sequence of random vectors in $\mathbb{R}^d$, and not to a component of a single random vector.

Law of Large Numbers

We work on a probability space $(\Omega, \mathcal{T}, \mathbb{P})$. Recall that two random vectors $X = (X^{(1)}, \ldots, X^{(d)})$ and $Y = (Y^{(1)}, \ldots, Y^{(d)})$ are independent if

$$\mathbb{P}_{(X,Y)} = \mathbb{P}_X \otimes \mathbb{P}_Y,$$

that is,

$$\mathbb{P}\left(X^{(1)}, \ldots, X^{(d)}, Y^{(1)}, \ldots, Y^{(d)}\right) = \mathbb{P}\left(X^{(1)}, \ldots, X^{(d)}\right) \otimes \mathbb{P}\left(Y^{(1)}, \ldots, Y^{(d)}\right).$$

Similarly, a sequence (X_k) of random vectors is said to be independent if, for every $n \in \mathbb{N}$,

$$\mathbb{P}(X_1, \ldots, X_n) = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}.$$

The weak and strong laws of large numbers extend without difficulty to the case of random vectors; it suffices to work componentwise. Let

$$S_n = X_1 + \cdots + X_n,$$

which is a random vector that assigns to each outcome $\omega \in \Omega$ a vector in $\mathbb{R}^d$. We denote by $S_n^{(i)}$ its i-th component:

$$S_n^{(i)} = X_1^{(i)} + \cdots + X_n^{(i)}.$$

Theorem 3.8 (Strong Law of Large Numbers for Random Vectors)**.** *Let $(\Omega, \mathcal{T}, \mathbb{P})$ be a probability space and let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random vectors with values in $\mathbb{R}^d$, which are integrable. Then*

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}(X_0),$$

where $S_n = X_1 + \cdots + X_n$.

The above convergence is clearly equivalent to the componentwise convergences

$$\frac{S_n^{(i)}}{n} \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}\left(X_0^{(i)}\right), \qquad i = 1, \ldots, d,$$

which follow from the one-dimensional case applied to the sequences $\left(X_n^{(i)}\right)_{n\in\mathbb{N}}$.

3.7.4 Multidimensional Central Limit Theorem

The limit in the central limit theorem for random vectors involves a multidimensional normal distribution. We denote by $\mathcal{N}(0, \Sigma)$ the distribution of a Gaussian vector in $\mathbb{R}^d$ with mean equal to the zero vector and covariance matrix equal to Σ.

Theorem 3.9 (Multidimensional Central Limit Theorem)**.** *Let $(\Omega, \mathcal{T}, \mathbb{P})$ be a probability space and let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random vectors with values in $\mathbb{R}^d$, which are square-integrable. Let m denote the mean vector of each X_n and let Σ be their covariance matrix, assumed to be invertible. Then*

$$\frac{S_n - nm}{\sqrt{n}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

This implies the convergence

$$\mathbb{P}\Big(m_i n + a_i\sqrt{n} \le S_n^{(i)} \le m_i n + b_i\sqrt{n} \text{ for all } i \in \{1, \ldots, d\}\Big) \xrightarrow[n\to\infty]{}$$

$$\frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} \exp\Big(-\frac{1}{2} x^{\mathsf{T}} \Sigma^{-1} x\Big)\, dx.$$

The proof proceeds as in the one-dimensional case. Convergence in distribution is characterized using characteristic functions.

Proposition 3.11. *Let (Y_n) be a sequence of random vectors defined on a probability space $(\Omega, \mathcal{T}, \mathbb{P})$. If, for all $u \in \mathbb{R}^d$,*

$$\varphi_{Y_n}(u) \xrightarrow[n\to\infty]{} \varphi_Y(u),$$

then Y_n converges in distribution to Y.

We then perform a Taylor expansion of the characteristic function of

$$\frac{S_n - m}{\sqrt{n}}.$$

The calculations carried out in $\mathbb{R}$ extend to $\mathbb{R}^d$ without any difficulty.

3.8 Exercises

Exercise 3.5. *Let $(U_n)_{n\geq 1}$ be a sequence of independent random variables, all uniformly distributed on $[0,1]$. Define*

$$M_n = \max(U_1, \dots, U_n) \quad \textit{and} \quad X_n = n(1 - M_n).$$

- *Determine the cumulative distribution function of X_n.*
- *Study the convergence in distribution of the sequence (X_n).*

Solution. We first determine the distribution function of M_n. Since M_n takes values in $[0,1]$, it is clear that

$$\mathbb{P}(M_n \leq x) = 0 \quad \text{if } x \leq 0, \qquad \mathbb{P}(M_n \leq x) = 1 \quad \text{if } x \geq 1.$$

Now let $x \in (0,1)$. Then

$$M_n \leq x \iff \forall i \in \{1, \dots, n\},\ U_i \leq x.$$

Since the random variables $(U_i)_{1\leq i\leq n}$ are independent, we obtain

$$\mathbb{P}(M_n \leq x) = \prod_{i=1}^{n} \mathbb{P}(U_i \leq x) = x^n.$$

To obtain the distribution function of X_n, we observe that

$$X_n \leq x \iff M_n \geq 1 - \frac{x}{n},$$

hence

$$\mathbb{P}(X_n \leq x) = 1 - \mathbb{P}\left(M_n \leq 1 - \frac{x}{n}\right).$$

Moreover,

$$1 - \frac{x}{n} \in [0,1] \iff x \in [0,n].$$

Therefore, the distribution function of X_n is given by

$$F_{X_n}(x) = \begin{cases} 0, & x \leq 0, \\ 1 - \left(1 - \dfrac{x}{n}\right)^n, & 0 \leq x \leq n, \\ 1, & x \geq n. \end{cases}$$

We now study the pointwise limit of $F_{X_n}(x)$. For $x \leq 0$, we clearly have

$$\lim_{n\to\infty} F_{X_n}(x) = 0.$$

For $x \geq 0$, when n is large enough we have $x \leq n$, and thus

$$F_{X_n}(x) = 1 - \left(1 - \frac{x}{n}\right)^n.$$

Using the classical limit

$$\lim_{n\to\infty}\left(1-\frac{x}{n}\right)^n = e^{-x},$$

we obtain

$$\lim_{n\to\infty} F_{X_n}(x) = \begin{cases} 0, & x \le 0, \\ 1-e^{-x}, & x \ge 0. \end{cases}$$

This is the distribution function of an exponential random variable with parameter 1. Therefore,

$$X_n \xrightarrow{\mathcal{L}} \text{Exp}(1).$$

Exercise 3.6. *Let $(X_n)_{n\ge1}$ be a sequence of random variables that converges in distribution to a constant random variable*

$$X = a.$$

Prove that the sequence (X_n) also converges in probability to X.

Solution. Let $\varepsilon > 0$. We must prove that

$$\mathbb{P}\Big(|X_n - a| > \varepsilon\Big) \longrightarrow 0 \quad \text{as } n \to \infty.$$

We write

$$\mathbb{P}\Big(|X_n - a| > \varepsilon\Big) = 1 - \mathbb{P}\Big(|X_n - a| \le \varepsilon\Big).$$

Moreover,

$$\mathbb{P}\Big(|X_n - a| \le \varepsilon\Big) = \mathbb{P}\Big(X_n \in [a-\varepsilon,\, a+\varepsilon]\Big) \ge \mathbb{P}\Big(X_n \in (a-\tfrac{\varepsilon}{2},\, a+\varepsilon]\Big).$$

Using the distribution function F_{X_n} of X_n, we obtain

$$\mathbb{P}\Big(X_n \in (a-\tfrac{\varepsilon}{2},\, a+\varepsilon]\Big) = F_{X_n}(a+\varepsilon) - F_{X_n}(a-\tfrac{\varepsilon}{2}).$$

Hence,

$$\mathbb{P}\Big(|X_n - a| > \varepsilon\Big) \le 1 - F_{X_n}(a+\varepsilon) + F_{X_n}(a-\tfrac{\varepsilon}{2}).$$

Since X is almost surely equal to a, its distribution function F_X satisfies

$$F_X(x) = \begin{cases} 0, & x < a, \\ 1, & x \ge a. \end{cases}$$

In particular, F_X is continuous at the points $a+\varepsilon$ and $a-\frac{\varepsilon}{2}$. By the convergence in distribution of X_n to X, we therefore have

$$F_{X_n}(a+\varepsilon) \longrightarrow F_X(a+\varepsilon) = 1, \qquad F_{X_n}(a-\tfrac{\varepsilon}{2}) \longrightarrow F_X(a-\tfrac{\varepsilon}{2}) = 0.$$

Consequently,

$$\mathbb{P}\Big(|X_n - a| > \varepsilon\Big) \longrightarrow 0,$$

which proves that

$$X_n \xrightarrow{\mathbb{P}} a.$$

Exercise 3.7. *A perfectly fair coin is tossed infinitely many times. Show that, with probability one, two consecutive heads occur infinitely often.*

Solution. For each integer $n \geq 1$, define the event

$$B_n = \{\text{the } 2n\text{-th toss is heads and the } (2n+1)\text{-th toss is heads}\}.$$

Since the coin tosses are independent, the events $(B_n)_{n\geq 1}$ are independent. Moreover, for each $n \geq 1$, we have

$$\mathbb{P}(B_n) = \frac{1}{4}.$$

Therefore,

$$\sum_{n=1}^{\infty} \mathbb{P}(B_n) = \sum_{n=1}^{\infty} \frac{1}{4} = +\infty.$$

By the second Borel–Cantelli lemma, it follows that, with probability one, infinitely many of the events B_n occur. Hence, with probability one, two consecutive heads appear infinitely often.

Exercise 3.8. *For every nonzero natural integer n, consider the function f_n defined by*

$$f_n(x) = \mathbf{1}_{\mathbb{R}_+}(x)\, n^2 x \exp\left(-\frac{n^2 x^2}{2}\right),$$

where $\mathbf{1}_{\mathbb{R}_+}$ denotes the indicator function of $\mathbb{R}_+ = [0, +\infty)$.

- *Show that f_n is the probability density function of a random variable.*
- *Let (X_n) be a sequence of random variables such that, for every $n \geq 1$, X_n has density f_n. Prove that the sequence (X_n) converges in probability to a random variable X, which you must specify.*

Solution. We first note that the function f_n is nonnegative and continuous on $\mathbb{R}$. Moreover, for any $x > 0$, we have

$$\int_{-\infty}^{x} f_n(t)\, dt = \int_0^x n^2 t \exp\left(-\frac{n^2 t^2}{2}\right) dt.$$

By a direct computation,

$$\int_0^x n^2 t \exp\left(-\frac{n^2 t^2}{2}\right) dt = \left[-\exp\left(-\frac{n^2 t^2}{2}\right)\right]_0^x = 1 - \exp\left(-\frac{n^2 x^2}{2}\right).$$

Letting $x \to +\infty$, we obtain

$$\int_{-\infty}^{+\infty} f_n(t)\, dt = 1.$$

Hence, f_n is a probability density function.

Let $\varepsilon > 0$. We compute

$$\mathbb{P}\big(|X_n| \geq \varepsilon\big) = \int_{\varepsilon}^{+\infty} n^2 t \exp\left(-\frac{n^2 t^2}{2}\right) dt = \left[-\exp\left(-\frac{n^2 t^2}{2}\right)\right]_{\varepsilon}^{+\infty} = \exp\left(-\frac{n^2 \varepsilon^2}{2}\right).$$

As $n \to +\infty$, we have

$$\exp\left(-\frac{n^2 \varepsilon^2}{2}\right) \longrightarrow 0.$$

Therefore,

$$\mathbb{P}\Big(|X_n| \geq \varepsilon\Big) \longrightarrow 0,$$

which proves that

$$X_n \xrightarrow{\mathbb{P}} 0.$$

Thus, the sequence (X_n) converges in probability to the constant random variable

$$X = 0.$$

Exercise 3.9. *Let n be a nonzero natural integer and let $a \in \mathbb{R}$. Consider the function f_n defined on $\mathbb{R}$ by*

$$f_n(x) = \frac{an}{\pi\Big(1 + n^2x^2\Big)}.$$

- *Determine the value of a for which f_n is the probability density function of a random variable.*
- *Let (X_n) be a sequence of random variables such that each X_n has density f_n. Study the existence of moments of X_n.*
- *Study the convergence in distribution of the sequence (X_n).*
- *Study the convergence in probability of the sequence (X_n).*

Solution. 1. Determination of a.

Since f_n is continuous and nonnegative, it is a probability density function if and only if

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

Using the change of variables $u = nx$, we obtain

$$\int_{-\infty}^{+\infty} \frac{an}{\pi(1+n^2x^2)}\,dx = \int_{-\infty}^{+\infty} \frac{a}{\pi(1+u^2)}\,du = \frac{a}{\pi}\Big[\arctan(u)\Big]_{-\infty}^{+\infty} = \frac{a}{\pi}\,\pi = a.$$

Hence, f_n is a probability density function if and only if

$$a = 1.$$

2. Existence of moments.

For large $|x|$, we have the asymptotic behavior

$$x f_n(x) \sim \frac{1}{\pi n x}, \qquad |x| \to +\infty.$$

Since the function $x \mapsto \frac{1}{|x|}$ is not integrable at infinity, the integral

$$\int_{\mathbb{R}} |x|\, f_n(x)\,dx$$

diverges. Therefore, the random variable X_n has no finite expectation, nor does it admit any finite moment of positive order.

3. Convergence in distribution.

Let F_n denote the distribution function of X_n. For any $x \in \mathbb{R}$,

$$F_n(x) = \int_{-\infty}^{x} f_n(t)\,dt = \frac{1}{\pi}\left(\arctan(nx) + \frac{\pi}{2}\right).$$

If $x < 0$, then $\arctan(nx) \to -\frac{\pi}{2}$ as $n \to +\infty$, and thus

$$F_n(x) \longrightarrow 0.$$

If $x > 0$, then $\arctan(nx) \to \frac{\pi}{2}$ as $n \to +\infty$, and hence

$$F_n(x) \longrightarrow 1.$$

Let X be the random variable identically equal to 0. Its distribution function F_X satisfies

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0. \end{cases}$$

At every continuity point of F_X, we have $F_n(x) \to F_X(x)$. Therefore,

$$X_n \xrightarrow{\mathcal{L}} 0.$$

4. Convergence in probability.

We now prove that (X_n) converges in probability to 0. Let $\varepsilon > 0$. We must show that

$$\lim_{n \to +\infty} \mathbb{P}(|X_n| \geq \varepsilon) = 0.$$

Since the density f_n is an even function, we have

$$\mathbb{P}(|X_n| \geq \varepsilon) = 2\int_{\varepsilon}^{+\infty} \frac{n}{\pi(1 + n^2x^2)}\,dx.$$

Using the change of variables $u = nx$, we obtain

$$\mathbb{P}(|X_n| \geq \varepsilon) = 2\int_{n\varepsilon}^{+\infty} \frac{1}{\pi(1 + u^2)}\,du.$$

Since $\int_0^{+\infty} \frac{1}{1+u^2}\,du$ is convergent, the tail integral

$$\int_{n\varepsilon}^{+\infty} \frac{1}{1 + u^2}\,du$$

tends to 0 as $n \to +\infty$. Hence,

$$\mathbb{P}(|X_n| \geq \varepsilon) \longrightarrow 0,$$

which proves that

$$X_n \xrightarrow{\mathbb{P}} 0.$$

Exercise 3.10. ***Linear transformation of a Gaussian vector.***

Let

$$X = (X_1, X_2)^{\mathsf{T}}$$

be a centered real-valued two-dimensional Gaussian random vector with covariance matrix

$$K_X = \begin{pmatrix} 3 & \rho\sqrt{3} \\ \rho\sqrt{3} & 1 \end{pmatrix}, \qquad |\rho| < 1.$$

Define a new random vector

$$Y = (Y_1, Y_2)^{\mathsf{T}}$$

by

$$\begin{cases} Y_1 = \dfrac{X_1}{\sqrt{3}} - X_2, \\ Y_2 = \dfrac{X_1}{\sqrt{3}} + X_2. \end{cases}$$

1. *Compute the variance* $\mathrm{Var}(X_1)$ *and the covariance* $\mathrm{Cov}(X_1, X_2)$.
2. *In which case are the random variables* X_1 *and* X_2 *independent?*
3. *Compute the covariance* $\mathrm{Cov}(Y_1, Y_2)$ *and the variances of* Y_1 *and* Y_2.
4. *Are the random variables* Y_1 *and* Y_2 *independent? Justify your answer.*

Solution. 1. Variance and covariance of (X_1, X_2)**.**
From the covariance matrix K_X, we directly obtain

$$\mathrm{Var}(X_1) = 3, \qquad \mathrm{Cov}(X_1, X_2) = \mathbb{E}\Big[(X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)\Big] = \rho\sqrt{3}.$$

2. Independence of X_1 **and** X_2**.**
Since the vector X is Gaussian, the components X_1 and X_2 are independent if and only if their covariance is zero. Thus, X_1 and X_2 are independent if and only if

$$\mathrm{Cov}(X_1, X_2) = 0 \quad \Longleftrightarrow \quad \rho = 0.$$

3. Covariance and variances of (Y_1, Y_2)**.**
Since X is centered, we have

$$\mathbb{E}(Y_1) = \mathbb{E}(Y_2) = 0.$$

We now compute the covariance:

$$\begin{aligned} \mathrm{Cov}(Y_1, Y_2) &= \mathbb{E}[Y_1 Y_2] \\ &= \mathbb{E}\left[\left(\frac{X_1}{\sqrt{3}} - X_2\right)\left(\frac{X_1}{\sqrt{3}} + X_2\right)\right] \\ &= \frac{1}{3}\mathbb{E}[X_1^2] + \frac{1}{\sqrt{3}}\mathbb{E}[X_1 X_2] - \frac{1}{\sqrt{3}}\mathbb{E}[X_1 X_2] - \mathbb{E}[X_2^2] \\ &= \frac{1}{3}\mathbb{E}[X_1^2] - \mathbb{E}[X_2^2]. \end{aligned}$$

Since $\mathbb{E}[X_1^2] = \mathrm{Var}(X_1) = 3$ and $\mathbb{E}[X_2^2] = \mathrm{Var}(X_2) = 1$, we obtain

$$\mathrm{Cov}(Y_1, Y_2) = 1 - 1 = 0.$$

Next, the variances of Y_1 and Y_2 are

$$\mathrm{Var}(Y_1) = \mathrm{Var}\left(\frac{X_1}{\sqrt{3}} - X_2\right) = \frac{1}{3}\,\mathrm{Var}(X_1) + \mathrm{Var}(X_2) - \frac{2}{\sqrt{3}}\,\mathrm{Cov}(X_1, X_2),$$

$$\mathrm{Var}(Y_2) = \mathrm{Var}\left(\frac{X_1}{\sqrt{3}} + X_2\right) = \frac{1}{3}\,\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \frac{2}{\sqrt{3}}\,\mathrm{Cov}(X_1, X_2).$$

Substituting the values,

$$\mathrm{Var}(Y_1) = 1 + 1 - 2\rho = 2(1-\rho), \qquad \mathrm{Var}(Y_2) = 1 + 1 + 2\rho = 2(1+\rho).$$

4. Independence of Y_1 and Y_2.
The vector Y is a linear transformation of the Gaussian vector X; therefore, Y is also Gaussian. Since

$$\mathrm{Cov}(Y_1, Y_2) = 0,$$

the random variables Y_1 and Y_2 are independent.

Computation of the variances of Y_1 and Y_2. We now compute the variances of Y_1 and Y_2. We obtain

$$\begin{aligned}\mathrm{Var}(Y_1) = \mathbb{E}\big[{Y_1}^2\big] &= \mathbb{E}\left[\left(\frac{X_1}{\sqrt{3}} - X_2\right)^2\right] \\ &= \mathbb{E}\left[\frac{X_1^2}{3} + X_2^2 - \frac{2}{\sqrt{3}} X_1 X_2\right] \\ &= \frac{1}{3}\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] - \frac{2}{\sqrt{3}}\mathbb{E}[X_1 X_2].\end{aligned}$$

Using $\mathbb{E}[X_1^2] = \mathrm{Var}(X_1) = 3$, $\mathbb{E}[X_2^2] = \mathrm{Var}(X_2) = 1$, and $\mathbb{E}[X_1 X_2] = \mathrm{Cov}(X_1, X_2) = \rho\sqrt{3}$, we obtain

$$\mathrm{Var}(Y_1) = 1 + 1 - 2\rho = 2(1-\rho).$$

Similarly,

$$\begin{aligned}\mathrm{Var}(Y_2) = \mathbb{E}\big[{Y_2}^2\big] &= \mathbb{E}\left[\left(\frac{X_1}{\sqrt{3}} + X_2\right)^2\right] \\ &= \mathbb{E}\left[\frac{X_1^2}{3} + X_2^2 + \frac{2}{\sqrt{3}} X_1 X_2\right] \\ &= \frac{1}{3}\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] + \frac{2}{\sqrt{3}}\mathbb{E}[X_1 X_2] \\ &= 1 + 1 + 2\rho = 2(1+\rho).\end{aligned}$$

Matrix approach. We can also proceed more efficiently using matrix notation. Let

$$Y = AX, \qquad A = \begin{pmatrix} \frac{1}{\sqrt{3}} & -1 \\ \frac{1}{\sqrt{3}} & 1 \end{pmatrix}.$$

Since X is a Gaussian random vector, any linear transformation of X is also Gaussian. Thus,

$$Y \sim \mathcal{N}\big(\mathbb{E}[Y],\, K_Y\big),$$

where K_Y is the covariance matrix of Y.

We have

$$\mathbb{E}[Y] = \mathbb{E}[AX] = A\mathbb{E}[X] = 0,$$

since $\mathbb{E}[X] = 0$. The covariance matrix of Y is given by

$$K_Y = \mathbb{E}[YY^{\mathsf{T}}] = A\,\mathbb{E}[XX^{\mathsf{T}}]\,A^{\mathsf{T}} = AK_XA^{\mathsf{T}}.$$

A direct computation yields

$$K_Y = \begin{pmatrix} 2(1-\rho) & 0 \\ 0 & 2(1+\rho) \end{pmatrix}.$$

This matrix approach confirms all the results obtained previously.

Independence of Y_1 and Y_2. The random vector Y is obtained as a linear transformation of the Gaussian vector X; hence, Y is also Gaussian. For Gaussian random vectors, independence of the components is equivalent to zero covariance. Since

$$\mathrm{Cov}(Y_1, Y_2) = 0,$$

the covariance matrix K_Y is diagonal. It follows that the components Y_1 and Y_2 are uncorrelated and therefore independent.

Exercise 3.11. *(Weak Law of Large Numbers). Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with*

$$\mathbb{E}[X_1] = \mu.$$

Then the sample mean

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges in probability to μ, that is,

$$\overline{X}_n \xrightarrow{\mathbb{P}} \mu.$$

Solution

Assume that $\sigma^2 = \mathrm{Var}(X_1) < \infty$. (This assumption is not necessary for the result, but it simplifies the proof.)

By Chebyshev's inequality, for any $\varepsilon > 0$,

$$\mathbb{P}\Big(\big|\overline{X}_n - \mu\big| > \varepsilon\Big) \le \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2}.$$

Since the random variables are independent and identically distributed,

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Therefore,

$$\mathbb{P}\Big(\big|\overline{X}_n - \mu\big| > \varepsilon\Big) \leq \frac{\sigma^2}{n\varepsilon^2},$$

which tends to 0 as $n \to \infty$. This proves that

$$\overline{X}_n \xrightarrow{\mathbb{P}} \mu.$$

□

Exercise 3.12. *True or False?*

1. *Let $(X_n)_{n\geq 1}$ be a sequence of real-valued random variables and let X be a real-valued random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that $X_n \xrightarrow{\mathcal{L}} X$. Show that $f(X_n) \xrightarrow{\mathcal{L}} f(X)$ for every continuous function $f : \mathbb{R} \to \mathbb{R}$.*

2. *Let $(\mu_n)_{n\geq 0}$ be a sequence of probability measures and let μ be a positive measure. Then μ_n converges narrowly (weakly) to μ if and only if, for every continuous function f with compact support, we have*

$$\int f\, d\mu_n \longrightarrow \int f\, d\mu.$$

3. *If the sequence of random variables $(X_n)_{n\geq 0}$ converges in distribution to X, then*

$$\mathbb{E}[X_n] \longrightarrow \mathbb{E}[X].$$

Solution.

1. **True.** Let $g : \mathbb{R} \to \mathbb{R}$ be a bounded continuous function. Then $g \circ f$ is also bounded and continuous. By the definition of convergence in distribution,

$$\mathbb{E}\Big[g(f(X_n))\Big] \longrightarrow \mathbb{E}\Big[g(f(X))\Big].$$

   Hence, $f(X_n) \xrightarrow{\mathcal{L}} f(X)$.

2. **False.** The implication is true, but the converse is false. Indeed, by a standard result, $(\mu_n)_{n\geq 0}$ converges narrowly to μ if and only if μ is a probability measure and, for every continuous function f with compact support,

$$\int f\, d\mu_n \longrightarrow \int f\, d\mu.$$

   The converse fails in general; for example, take $\mu_n = \delta_n$, the Dirac measure at n.

3. **False.** Consider $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}([0,1]), dx)$ and define $X_n(t)$ as the tent function such that

$$X_n(0) = 0, \qquad X_n\Big(\tfrac{1}{n}\Big) = n, \qquad X_n\Big(\tfrac{2}{n}\Big) = 0.$$

   Then $X_n \to 0$ almost surely, but

$$\mathbb{E}[X_n] = 1 \neq \mathbb{E}[0].$$

A more probabilistic example is the following. Let $(X_n)_{n\geq 1}$ be a sequence of independent and identically distributed random variables such that

$$\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = \frac{1}{2}.$$

Define

$$Z_n = X_1 + X_2 + \cdots + X_n, \qquad T = \inf\{n \geq 1 : Z_n = 1\},$$

and

$$W_n = Z_{\min(n,T)}.$$

Then (W_n) is a simple random walk starting from 0 that stays at 1 once it reaches it. It can be shown that $T < \infty$ almost surely, so that $W_n \to 1$ almost surely. However, for every $n \geq 1$,

$$\mathbb{E}[W_n] = 0,$$

and thus $\mathbb{E}[W_n]$ does not converge to $\mathbb{E}[1]$.

In the language of stochastic processes, this provides an example of a martingale that converges almost surely but not in L^1.

Exercise 3.13. *(Bernstein–Weierstrass Theorem).*

Let f be a continuous function from $[0,1]$ to $\mathbb{C}$. The n-th Bernstein polynomial of f is defined by

$$B_n(x) := \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} f\Big(\frac{k}{n}\Big), \quad x \in [0,1].$$

1. *Let $S_n(x) \sim \mathrm{Bin}(n,x)$ be a binomial random variable with parameters n and x. Show that*

$$B_n(x) = \mathbb{E}\Big[f(S_n(x)/n)\Big].$$

2. *Deduce the Bernstein–Weierstrass theorem:*

$$\|B_n - f\|_\infty \longrightarrow 0 \quad \text{as } n \to \infty.$$

Solution.

1. This is immediate from the definition of expectation:

$$\mathbb{E}[f(S_n(x)/n)] = \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big)\mathbb{P}(S_n(x) = k) = \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big)\binom{n}{k} x^k (1-x)^{n-k} = B_n(x).$$

2. Let $\varepsilon > 0$ and let $\eta > 0$ be the modulus of uniform continuity of f associated with ε. Then

$$\begin{aligned}
|f(x) - B_n(x)| &= \Big|f(x) - \mathbb{E}[f(S_n(x)/n)]\Big| \\
&\leq \mathbb{E}\Big[|f(x) - f(S_n(x)/n)|\Big] \\
&= \mathbb{E}\Big[|f(x) - f(S_n(x)/n)|\mathbf{1}_{\{|S_n(x)/n - x| \leq \eta\}}\Big] + \mathbb{E}\Big[|f(x) - f(S_n(x)/n)|\mathbf{1}_{\{|S_n(x)/n - x| > \eta\}}\Big] \\
&\leq \varepsilon + 2\|f\|_\infty \mathbb{P}\Big(|S_n(x)/n - x| \geq \eta\Big).
\end{aligned}$$

To estimate $\mathbb{P}(|S_n(x)/n - x| \geq \eta)$, we use Markov's inequality:

$$\mathbb{P}(|S_n(x)/n - x| \geq \eta) \leq \frac{\mathrm{Var}(S_n(x)/n)}{\eta^2} = \frac{x(1-x)}{n\eta^2} \leq \frac{1}{4n\eta^2}.$$

This bound is uniform in $x \in [0,1]$, so we conclude that

$$\|B_n - f\|_\infty \longrightarrow 0 \quad \text{as } n \to \infty.$$

Hence, the Bernstein polynomials converge uniformly to f, proving the Bernstein–Weierstrass theorem.

Exercise 3.14. *By applying Markov's inequality to $|X|$, where X is a standard normal random variable, show that for every $x > 0$, we have*

$$\int_0^x e^{-t^2/2}\,dt \geq \sqrt{\frac{\pi}{2}} - \frac{1}{x}.$$

Solution. Let $X \sim \mathcal{N}(0,1)$ and denote its density by

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \quad t \in \mathbb{R}.$$

The random variable $|X|$ is non-negative and has a finite expectation. By Markov's inequality, for any $x > 0$,

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

Now,

$$\begin{aligned}
\mathbb{P}(|X| \geq x) &= \mathbb{P}(X \geq x) + \mathbb{P}(X \leq -x) \\
&= \int_x^{+\infty} f(t)\,dt + \int_{-\infty}^{-x} f(t)\,dt \\
&= \int_{-\infty}^{+\infty} f(t)\,dt - \int_{-x}^{x} f(t)\,dt \\
&= 1 - 2\int_0^x f(t)\,dt,
\end{aligned}$$

since f is even and integrates to 1.

On the other hand, by the definition of expectation (law of the unconscious statistician),

$$\mathbb{E}[|X|] = \int_{-\infty}^{+\infty} |t| f(t)\,dt = 2\int_0^{+\infty} t f(t)\,dt,$$

because $t \mapsto |t| f(t)$ is even.

Now, for any $A > 0$,

$$\int_0^A t f(t)\,dt = \frac{1}{\sqrt{2\pi}} \int_0^A t e^{-t^2/2}\,dt = \frac{1}{\sqrt{2\pi}} \Big[-e^{-t^2/2}\Big]_0^A \longrightarrow \frac{1}{\sqrt{2\pi}} \quad \text{as } A \to +\infty.$$

Finally, for any $x > 0$, Markov's inequality gives

$$1 - 2\int_0^x f(t)\,dt = \mathbb{P}(|X| \geq x) \leq \frac{\mathbb{E}[|X|]}{x} = \frac{2}{x}\int_0^{+\infty} t f(t)\,dt \approx \frac{2}{x} \cdot \frac{1}{\sqrt{2\pi}}.$$

Multiplying both sides by $\frac{\sqrt{2\pi}}{2} = \sqrt{\pi/2}$, we obtain

$$\sqrt{\frac{\pi}{2}} - \frac{1}{x} \leq \int_0^x e^{-t^2/2}\,dt,$$

as required.

Exercise 3.15. *Convergence of Random Variables*

1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $(A_n)_{n\geq 1}$ be a sequence of events, and $p \geq 1$ a real number. Determine, for each type of convergence below, the condition on the sequence $(A_n)_{n\geq 1}$ under which it holds.*

   (a) *The sequence $(\mathbf{1}_{A_n})_{n\geq 1}$ converges in probability to 0.*
   Solution: *Suppose that $(\mathbf{1}_{A_n})$ converges in probability to 0. Then in particular,*

   $$P(\mathbf{1}_{A_n} > 1/2) = P(A_n) \longrightarrow 0.$$

   Conversely, if $P(A_n) \to 0$, then for any $\varepsilon > 0$,

   $$P(\mathbf{1}_{A_n} > \varepsilon) \leq P(A_n) \to 0.$$

   Hence, the condition is

   $$\lim_{n\to+\infty} P(A_n) = 0.$$

   (b) *The sequence $(\mathbf{1}_{A_n})_{n\geq 1}$ converges in L^p to 0.*
   Solution: *Since $\mathbb{E}[\mathbf{1}_{A_n}^p] = P(A_n)$, the condition is again*

   $$\lim_{n\to+\infty} P(A_n) = 0.$$

   (c) *The sequence $(\mathbf{1}_{A_n})_{n\geq 1}$ converges almost surely to 0.* ***Solution:*** *For $\omega \in \Omega$, the sequence $(\mathbf{1}_{A_n}(\omega))$ converges to 0 if and only if it is eventually 0. This occurs if and only if $\omega \in \liminf A_n^c$, which is the complement of $\limsup A_n$. Therefore, almost sure convergence holds if and only if*

   $$P(\limsup A_n) = 0.$$

2. *Let $(X_n)_{n\geq 1}$ be a sequence of independent random variables. Suppose that $\sum_{n\geq 1} X_n$ converges almost surely. Show that for every $c > 0$,*

   $$\sum_{n\geq 1} P(|X_n| > c) < +\infty.$$

   Solution: *If $\sum_{n\geq 1} P(|X_n| > c) = +\infty$, then by independence and the second part of the Borel-Cantelli lemma, $|X_n| > c$ infinitely often with probability 1. On this event, the series $\sum X_n$ does not converge, which contradicts the assumption.*

3. *Construct a sequence of integrable random variables $(X_n)_{n\geq 1}$ and an integrable random variable X such that $X_n \xrightarrow{\mathcal{L}} X$ but $\lim_{n\to\infty} \mathbb{E}[X_n] \neq \mathbb{E}[X]$.*

   Solution: *For $n \geq 1$, define X_n by*

   $$P(X_n = 0) = 1 - \frac{1}{n}, \quad P(X_n = n) = \frac{1}{n}.$$

   For any bounded continuous function f, we have

   $$\mathbb{E}[f(X_n)] = \left(1 - \frac{1}{n}\right) f(0) + \frac{1}{n} f(n) \quad \Rightarrow \quad |\mathbb{E}[f(X_n)] - f(0)| \leq \frac{\|f\|_\infty}{n} \to 0,$$

   so (X_n) converges in law to 0. However, $\mathbb{E}[X_n] = 1$ for all n, which does not converge to 0.

4. *Show that if a sequence of random variables converges in law and each term is exponentially distributed, then the limiting law is either exponential or a Dirac mass at* 0.

   Solution: *Convergence in law is equivalent to pointwise convergence of Laplace transforms. If* $X_n \sim \mathrm{Exp}(\lambda_n)$*, then*

$$\phi_{X_n}(t) = \mathbb{E}[e^{-tX_n}] = \frac{\lambda_n}{\lambda_n + t}.$$

   If X_n *converges in law, the sequence* (λ_n) *converges in* $(0, +\infty]$*. If the limit* $\lambda_\infty := \lim \lambda_n$ *is finite, the limiting law is* $\mathrm{Exp}(\lambda_\infty)$*; if* $\lambda_\infty = +\infty$*, the limit is almost surely* 0*.*

Exercise 3.16. *Let the sequence of random variables* (X_n) *be defined for* $n \in \mathbb{N}$ *by*

$$P(X_n = 0) = 1 - \frac{1}{n}, \quad P(X_n = n) = \frac{1}{n}.$$

Show that (X_n) *converges in probability to* $X = 0$ *but does not converge in mean square. What about almost sure convergence?*

Solution.

1. **Convergence in probability:**

   Let $\varepsilon > 0$. We have

$$P(X_n > \varepsilon) \leq P(X_n > 0) = \frac{1}{n} \longrightarrow 0 \quad \text{as } n \to \infty,$$

   so (X_n) converges in probability to 0.

2. **Convergence in mean square:**

   We compute

$$\mathbb{E}[X_n^2] = n^2 \cdot \frac{1}{n} = n \longrightarrow +\infty,$$

   so (X_n) does not converge in mean square.

3. **Almost sure convergence:**

   For almost sure convergence, the sufficient condition

$$\sum_{n=0}^{+\infty} P(X_n > \varepsilon) < +\infty$$

   is not satisfied. We apply the Borel-Cantelli lemma to the events $E_n = \{X_n = n\}$. These events are independent and

$$\sum_{n \in \mathbb{N}} P(E_n) = \sum_{n \in \mathbb{N}} \frac{1}{n} = +\infty.$$

   Hence, with probability 1, infinitely many of the events E_n occur. Therefore, for any $n \in \mathbb{N}$,

$$P\Big(\sup_{N \geq n} X_N > \varepsilon\Big) = P(\exists N \geq n \text{ such that } X_N = N) = 1.$$

   Thus, (X_n) does not converge almost surely.

Exercise 3.17. *Let the sequence of random variables (X_n) be defined for $n \in \mathbb{N}$ by*

$$P(X_n = 0) = 1 - \frac{1}{n}, \quad P(X_n = n) = \frac{1}{n}.$$

Show that (X_n) converges in probability to $X = 0$ but does not converge in mean square. What about almost sure convergence?

Solution.

1. **Convergence in probability:**

   Let $\varepsilon > 0$. We have

   $$P(X_n > \varepsilon) \leq P(X_n > 0) = \frac{1}{n} \longrightarrow 0 \quad \text{as } n \to \infty,$$

   so (X_n) converges in probability to 0.

2. **Convergence in mean square:**

   We compute

   $$\mathbb{E}[X_n^2] = n^2 \cdot \frac{1}{n} = n \longrightarrow +\infty,$$

   so (X_n) does not converge in mean square.

3. **Almost sure convergence:**

   For almost sure convergence, the sufficient condition

   $$\sum_{n=0}^{+\infty} P(X_n > \varepsilon) < +\infty$$

   is not satisfied. We apply the Borel-Cantelli lemma to the events $E_n = \{X_n = n\}$. These events are independent and

   $$\sum_{n \in \mathbb{N}} P(E_n) = \sum_{n \in \mathbb{N}} \frac{1}{n} = +\infty.$$

   Hence, with probability 1, infinitely many of the events E_n occur. Therefore, for any $n \in \mathbb{N}$,

   $$P\Big(\sup_{N \geq n} X_N > \varepsilon\Big) = P(\exists N \geq n \text{ such that } X_N = N) = 1.$$

   Thus, (X_n) does not converge almost surely.

Exercise 3.18. *Let X and Y be two independent random variables following a Bernoulli distribution with the same parameter p. Define*

$$U = X + Y \quad \textit{and} \quad V = X - Y.$$

Determine the joint distribution of (U, V). Are U and V independent?

Solution 3.1. *The random variables X and Y are independent and follow a Bernoulli distribution with parameter p. Thus,*

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

Joint distribution of (U,V)**.**

Since $U=X+Y$ and $V=X-Y$, the possible values of (X,Y) and the corresponding values of (U,V) are given in the table below:

(X,Y)	(U,V)	*Probability*
$(0,0)$	$(0,0)$	$(1-p)^2$
$(1,0)$	$(1,1)$	$p(1-p)$
$(0,1)$	$(1,-1)$	$p(1-p)$
$(1,1)$	$(2,0)$	p^2

Hence, the joint distribution of (U,V) is

$$\begin{aligned}\mathbb{P}(U=0,V=0)&=(1-p)^2,\\ \mathbb{P}(U=1,V=1)&=p(1-p),\\ \mathbb{P}(U=1,V=-1)&=p(1-p),\\ \mathbb{P}(U=2,V=0)&=p^2,\end{aligned}$$

and all other probabilities are zero.

Independence of U ***and*** V**.**

The marginal distributions are given by

$$\mathbb{P}(U=0)=(1-p)^2, \qquad \mathbb{P}(U=1)=2p(1-p), \qquad \mathbb{P}(U=2)=p^2,$$

and

$$\mathbb{P}(V=0)=(1-p)^2+p^2, \qquad \mathbb{P}(V=1)=p(1-p), \qquad \mathbb{P}(V=-1)=p(1-p).$$

For example,

$$\mathbb{P}(U=0,V=0)=(1-p)^2\neq\mathbb{P}(U=0)\mathbb{P}(V=0)=(1-p)^2\Big((1-p)^2+p^2\Big),$$

for $p\in(0,1)$.

Therefore, U and V are not independent*.*

Exercise 3.19. *We define a pair of random variables (X,Y) with joint probability density*

$$f(x,y)=\begin{cases}x+y, & (x,y)\in[0,1]^2,\\ 0, & \textit{otherwise.}\end{cases}$$

(a) Compute the marginal densities f_X and f_Y of X and Y, respectively. Are these variables independent?

Solution. *The marginal density of X is*

$$f_X(x) = \int_0^1 f(x,y)\,dy = \int_0^1 (x+y)\,dy = x + \frac{1}{2}.$$

Similarly, the marginal density of Y is

$$f_Y(y) = \int_0^1 f(x,y)\,dx = \int_0^1 (x+y)\,dx = y + \frac{1}{2}.$$

We observe that

$$f_X(0)f_Y(0) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4} \neq f(0,0) = 0,$$

so X and Y are ***not independent****.*

(b) Determine the density of $Z = X + Y$.

Hint: *Draw the region corresponding to $Z \leq z$ depending on z.*

Solution. *Let $z \in \mathbb{R}$. We compute $P(Z \leq z)$.*
- If $z < 0$, $P(Z \leq z) = 0$. - If $z > 2$, $P(Z \leq z) = 1$ since $X + Y \in [0,2]$.
So we consider $z \in [0,2]$:

$$P(X+Y \leq z) = \int_{[0,1]^2, x+y\leq z} (x+y)\,dx\,dy.$$

Case 1: $z \in [0,1]$

$$P(X+Y \leq z) = \int_0^z \int_0^{z-x} (x+y)\,dy\,dx = \int_0^z \frac{(z-x)^2}{2}\,dx = \frac{z^3}{3}.$$

Case 2: $z \in [1,2]$

$$\begin{aligned} P(X+Y \leq z) &= \int_0^1 \int_0^{\min(1,z-x)} (x+y)\,dy\,dx \\ &= \int_0^{z-1} \left(x + \frac{1}{2}\right) dx + \int_{z-1}^1 \frac{(z-x)^2}{2}\,dx \\ &= z^2 - \frac{1}{3}(z^3+1). \end{aligned}$$

Finally, the density $f_Z(z)$ is obtained by differentiating $P(Z \leq z)$ with respect to z.

(b) Density of $Z = X+Y$ (continued) *Finally, we can give the density of the random variable Z:*

$$f_Z(z) = \begin{cases} 0, & z < 0, \\ \frac{z^2}{2}, & z \in [0,1], \\ 2z - \frac{z^2}{2}, & z \in [1,2], \\ 0, & z > 2. \end{cases}$$

(c) Covariance of (X, Y) *[2]*

Solution. *The covariance is defined as*

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

First, we compute

$$\mathbb{E}[X] = \int_0^1 x\, f_X(x)\, dx = \int_0^1 x \left(x + \frac{1}{2}\right) dx = \frac{7}{12},$$

and by symmetry $\mathbb{E}[Y] = \frac{7}{12}$.

Next,

$$\mathbb{E}[XY] = \int_0^1 \int_0^1 xy\, f(x, y)\, dy\, dx = \int_0^1 \int_0^1 xy(x + y)\, dy\, dx = \frac{1}{3}.$$

Finally,

$$\mathrm{Cov}(X, Y) = \frac{1}{3} - \frac{7}{12} \cdot \frac{7}{12} = -\frac{1}{144}.$$

Chapter 4

Statistical Estimation and Hypothesis Testing

Statistical estimation and hypothesis testing are two fundamental approaches for drawing conclusions about a population based on sample data. Although they address different questions, they are complementary and are often used together. Statistical estimation focuses on determining the value of unknown population parameters. These parameters, such as the mean income of a city or the proportion of defective products in a factory, are fixed but unknown. Since observing the entire population is generally impractical or too costly, a sample is used to estimate these quantities.

A *point estimate* provides a single best guess of a population parameter, whereas an *interval estimate*, such as a confidence interval, gives a range of plausible values that reflects the uncertainty of the estimation process. Not all estimators perform equally well: some may be biased, while others may exhibit large variability across different samples. A good estimator is one that is, on average, close to the true parameter value and has low variability. In general, larger and higher-quality samples lead to more precise estimates, resulting in narrower confidence intervals. Hypothesis testing addresses the question of whether there is sufficient evidence to support a specific claim about a population. The procedure begins with a *null hypothesis*, which typically represents no effect, no difference, or a reference value. The *alternative hypothesis* reflects the research question of interest, such as the presence of an effect or a change in a parameter.

Sample data are summarized by a test statistic and a *p-value*, which measures how unlikely the observed data would be if the null hypothesis were true. If the p-value is sufficiently small, the null hypothesis is rejected in favor of the alternative hypothesis.

Since hypothesis testing is based on probabilistic reasoning, incorrect decisions are possible. A *Type I error* occurs when a true null hypothesis is rejected, while a *Type II error* occurs when a false null hypothesis is not rejected. The *significance level* controls the probability of committing a Type I error, whereas the *power* of a test measures its ability to detect a true effect [**?**, 2, 3, 6–8, 12, 15, 17, 19, 21].

Relationship Between Estimation and Testing Estimation and hypothesis testing represent two perspectives on the same inferential goal: learning about populations from samples. Estimation emphasizes the magnitude of effects and the uncertainty associated with them, while hypothesis testing focuses on evaluating specific claims and determining whether the available evidence is sufficient. In practice, these two approaches are most effective when used together: estimates describe the size of an effect, and hypothesis tests assess the strength of the evidence supporting it.

4.1 Normal distribution

As discussed in the previous chapter, we have already introduced the normal or Gaussian) distribution.

Properties and Examples

Let Φ be the standard normal distribution. Then, we have the following properties.

- $\Phi(-x) = 1 - \Phi(x)$,
- $\Phi(0) = 0.5$,
- $\Phi(1.645) \approx 0.95$, $\Phi(1.960) \approx 0.975$.

For $|x| < 2$, a Taylor series approximation of Φ at order 5 near 0 can be used:

$$\Phi(x) \approx 0.5 + \frac{1}{\sqrt{2\pi}}\left(x - \frac{x^3}{6} + \frac{x^5}{40}\right).$$

Conversely, given a probability, one can determine the corresponding bound for which this probability is achieved.

4.2 Chi-Square Distribution χ^2

Definition 4.1. *Let $Z_1, Z_2, \ldots, Z_\nu$ be a sequence of independent random variables, each following the standard normal distribution $N(0,1)$ (with $Z_i \sim N(0,1)$ i.i.d). Then the random variable*

$$\sum_{i=1}^{\nu} Z_i^2$$

follows a distribution called the Chi-square distribution with ν degrees of freedom, denoted by $\chi^2(\nu)$.

Proposition 4.1. *1. Its characteristic function is*

$$\phi(t) = (1 - 2it)^{-\nu/2}.$$

2. The probability density function of $\chi^2(\nu)$ is

$$f_\nu(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} & \text{if } x > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where Γ is Euler's Gamma function defined by

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx.$$

3. The expectation of the $\chi^2(\nu)$ distribution is ν, and its variance is 2ν.

4. The sum of two independent random variables following $\chi^2(\nu_1)$ and $\chi^2(\nu_2)$, respectively, also follows a χ^2 distribution with $\nu_1 + \nu_2$ degrees of freedom.

Proof. We first calculate the characteristic function of Z^2 when $Z \sim N(0,1)$:

$$\phi(t) = \mathbb{E}[e^{itZ^2}] = \int_{-\infty}^{\infty} e^{itz^2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(1-2it)z^2} dz.$$

Setting $u = \sqrt{1-2it}\, z$, we get

$$\phi(t) = (1-2it)^{-1/2}.$$

For the sum of ν independent variables Z_i^2, the characteristic function becomes

$$\phi(t) = (1-2it)^{-\nu/2}.$$

We now show that the density function is correct. For this, we calculate the characteristic function from the density:

$$\begin{aligned}
\phi(t) = \mathbb{E}[e^{itX}] &= \int_0^{+\infty} e^{itx} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} dx \\
&= \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^{+\infty} x^{\nu/2-1} e^{-(\frac{1}{2}-it)x} dx \\
&= \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \frac{1}{(\frac{1}{2}-it)^{\nu/2}} \int_0^{+\infty} u^{\nu/2-1} e^{-u} du \quad \text{with } u = (\frac{1}{2}-it)x \\
&= \frac{1}{(1-2it)^{\nu/2}}.
\end{aligned}$$

Now we calculate the expectation and variance. By definition of the χ^2 distribution, each variable Z_i follows the standard normal distribution. Thus,

$$\mathbb{E}[Z_i^2] = \mathrm{Var}(Z_i) = 1 \quad \text{and} \quad \mathbb{E}\Big[\sum_{i=1}^{\nu} Z_i^2\Big] = \nu.$$

Similarly,

$$\mathrm{Var}(Z_i^2) = \mathbb{E}[Z_i^4] - (\mathbb{E}[Z_i^2])^2 = \mu_4 - 1.$$

For a standard normal distribution, $\mu_4 = 3$, so $\mathrm{Var}(Z_i^2) = 2$ and

$$\mathrm{Var}\Big(\sum_{i=1}^{\nu} Z_i^2\Big) = 2\nu.$$

The last proposition regarding the sum of independent χ^2 variables is straightforward from the definition of the χ^2 distribution. □

4.3 Student's t-Distribution

Definition 4.2. *Let Z and Q be two independent random variables such that $Z \sim N(0,1)$ and $Q \sim \chi^2(\nu)$. Then the random variable*

$$T = \frac{Z}{\sqrt{Q/\nu}}$$

follows a distribution called the Student's t-distribution with ν degrees of freedom, denoted by $St(\nu)$.

Proposition 4.2. *1. The density of the Student's t-distribution with ν degrees of freedom is*

$$f(x) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\left(1+\frac{x^2}{\nu}\right)^{(\nu+1)/2}}.$$

2. The expectation is not defined for $\nu = 1$ and equals 0 if $\nu \geq 2$. Its variance does not exist for $\nu \leq 2$ and equals $\nu/(\nu-2)$ for $\nu \geq 3$.

3. The Student's t-distribution converges in distribution to the standard normal distribution as $\nu \to \infty$.

Remark 4.1. *For $\nu = 1$, the Student's t-distribution is called the Cauchy distribution or Lorentz distribution.*

4.4 Fisher-Snedecor F-Distribution

Definition 4.3. *Let Q_1 and Q_2 be two independent random variables such that $Q_1 \sim \chi^2(\nu_1)$ and $Q_2 \sim \chi^2(\nu_2)$. Then the random variable*

$$F = \frac{Q_1/\nu_1}{Q_2/\nu_2}$$

follows a Fisher-Snedecor distribution with (ν_1, ν_2) degrees of freedom, denoted $F(\nu_1, \nu_2)$.

Proposition 4.3. *The density of $F(\nu_1, \nu_2)$ is*

$$f(x) = \begin{cases} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{\nu_1/2-1} \left(1+\frac{\nu_1}{\nu_2}x\right)^{-(\nu_1+\nu_2)/2}, & x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Its expectation exists only if $\nu_2 \geq 3$ and equals $\nu_2/(\nu_2-2)$. Its variance exists only if $\nu_2 \geq 5$ and equals

$$Var(F) = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}.$$

Proposition 4.4. *1. If $F \sim F(\nu_1, \nu_2)$, then $1/F \sim F(\nu_2, \nu_1)$.*

2. If T follows a Student's t-distribution with ν degrees of freedom, then $T^2 \sim F(1, \nu)$.

4.5 Sampling and Estimation

4.5.1 Sampling

We will study how a sample (randomly drawn elements) behaves in a population for which we know the statistical characteristics (distributions, etc.) of a considered variable X. In this case, taking a random sample of size n consists of considering n realizations of X or, equivalently, considering n independent random variables $X_1, \ldots, X_n$ with the same distribution as X.

Definition 4.4. *Let X be a random variable on a probability space Ω. A sample of X of size n is an n-tuple $(X_1, \ldots, X_n)$ of independent random variables with the same distribution as X. The distribution of X is called the* parent distribution. *A realization of this sample is an n-tuple of real numbers $(x_1, \ldots, x_n)$ where $X_i(\omega) = x_i$.*

4.5.2 Empirical Mean and Variance

Definition 4.5. *A* statistic *on an n-sample is any function of* $(X_1, \ldots, X_n)$.

Empirical mean

Definition 4.6. *The* sample mean *or* empirical mean *is the statistic denoted by* $\overline{X}$ *defined as*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Proposition 4.5. *Let* X *be a random variable with mean* μ *and standard deviation* σ. *Then*

$$E(\overline{X}) = \mu, \quad V(\overline{X}) = \frac{\sigma^2}{n}.$$

Moreover, by the Central Limit Theorem, $\overline{X}$ *converges in distribution to* $N(\mu, \sigma/\sqrt{n})$ *as* $n \to \infty$.

Proof.

$$E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu.$$

And, due to the independence of the X_i,

$$V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Moreover, by the Central Limit Theorem, the random variable

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges in distribution to a normal random variable

$$\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{as } n \to \infty.$$

□

Theorem 4.1. *Any sum of independent normal random variables is itself a normal random variable. In particular, if* $X_i \sim \mathcal{N}(\mu, \sigma^2)$, *then for any* n,

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Proof. It suffices to establish the result for the sum of two independent random variables, the general case following by induction. Let X_1 and X_2 be independent random variables with respective distributions $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$. The result for the sum is obtained using characteristic functions. □

Empirical Variance

The empirical variance is the statistic denoted by $\widetilde{S}^2(X)$ and defined as

$$\widetilde{S}^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, \quad \text{where } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Proposition 4.6. *Let X be a random variable with standard deviation σ and fourth central moment μ_4. Then we have*

$$\mathbb{E}(\widetilde{S}^2) = \frac{n-1}{n}\sigma^2, \quad \mathrm{Var}(\widetilde{S}^2) = \frac{n-1}{n^3}\Big((n-1)\mu_4 - (n-3)\sigma^4\Big).$$

Moreover, as $n \to \infty$,

$$\mathrm{Var}(\widetilde{S}^2) \sim \frac{\mu_4 - \sigma^4}{n}.$$

Proof. We start from the definition of the empirical variance:

$$\widetilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}\Big[(X_i - \mu) - (\overline{X} - \mu)\Big]^2.$$

Expanding the square, we get

$$\widetilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + (\overline{X}-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\overline{X}-\mu)^2.$$

Taking expectations, we obtain

$$\mathbb{E}(\widetilde{S}^2) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(X_i) - \mathrm{Var}(\overline{X}) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Let us prove the other equality. Recall the notations: the X_i are independent and identically distributed and the k-th central moments are defined by

$$\mu_k = \mathbb{E}\Big((X-\mu)^k\Big).$$

In particular, $\mu_1 = 0$ and $\mu_2 = \sigma^2$.

The empirical variance can be written as

$$\widetilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2.$$

On the other hand, consider the sum of squared differences:

$$\sum_{i,j}(X_i - X_j)^2 = \sum_{i,j}\Big(X_i^2 - 2X_iX_j + X_j^2\Big) = \sum_{i,j}X_i^2 - 2\sum_{i,j}X_iX_j + \sum_{i,j}X_j^2.$$

Since $\sum_{i,j}X_i^2 = n\sum_{i=1}^{n}X_i^2$ and similarly for $\sum_{i,j}X_j^2$, and $\sum_{i,j}X_iX_j = \Big(\sum_{i=1}^{n}X_i\Big)\Big(\sum_{j=1}^{n}X_j\Big) = (n\overline{X})^2$, we obtain

$$\begin{aligned}\sum_{i,j}(X_i - X_j)^2 &= 2n\sum_{i=1}^{n}X_i^2 - 2n^2\overline{X}^2 \\ &= 2n^2\widetilde{S}^2,\end{aligned}$$

using the previous expression for $\widetilde{S}^2$. □

Variance of $\widetilde{S}^2$

We can calculate the variance of $\widetilde{S}^2$ using the following relation:

$$\mathrm{Var}(\widetilde{S}^2) = \mathrm{Cov}(\widetilde{S}^2, \widetilde{S}^2) = \frac{1}{(2n^2)^2} \sum_{i,j,k,l} \mathrm{Cov}\Big((X_i - X_j)^2, (X_k - X_l)^2\Big).$$

We then compute the different covariances according to the type of terms:

- Covariances of the form $\mathrm{Cov}((X_i - X_j)^2, (X_k - X_l)^2)$ with i, j, k, l all distinct,
- Covariances of the form $\mathrm{Cov}((X_i - X_j)^2, (X_k - X_j)^2)$ with i, j, k distinct,
- Covariances of the form $\mathrm{Cov}((X_i - X_j)^2, (X_i - X_j)^2)$ with $i \neq j$.

Note that if $i = j$ or $k = l$, then the covariance involves zero, either of the form $\mathrm{Cov}(0, (X_k - X_l)^2)$ or $\mathrm{Cov}((X_i - X_j)^2, 0)$, which equals zero.

Calculation of $\mathrm{Cov}((X_i - X_j)^2, (X_i - X_j)^2)$ for $i \neq j$

$$\mathrm{Cov}((X_i - X_j)^2, (X_i - X_j)^2) = \mathbb{E}\Big((X_i - X_j)^4\Big) - \Big[\mathbb{E}\Big((X_i - X_j)^2\Big)\Big]^2.$$

Computation of Covariances for Distinct Indices

Thus, for $i \neq j$, we have

$$\mathrm{Cov}((X_i - X_j)^2, (X_i - X_j)^2) = 2\mu_4 + 2\sigma^4.$$

Next, we compute $\mathrm{Cov}((X_i - X_j)^2, (X_k - X_j)^2)$ with i, j, k all distinct:

$$\mathrm{Cov}((X_i - X_j)^2, (X_k - X_j)^2) = \mathbb{E}\Big((X_i - X_j)^2 (X_k - X_j)^2\Big) - \mathbb{E}\Big((X_i - X_j)^2\Big)\mathbb{E}\Big((X_k - X_j)^2\Big).$$

Since $\mathbb{E}((X_i - X_j)^2) = 2\sigma^2$, this becomes

$$\mathrm{Cov}((X_i - X_j)^2, (X_k - X_j)^2) = \mathbb{E}\Big((X_i - X_j)^2 (X_k - X_j)^2\Big) - (2\sigma^2)^2.$$

Computation of $\mathrm{Cov}((X_i - X_j)^2, (X_k - X_j)^2)$

Expanding the product, we have

$$(X_i - X_j)^2 (X_k - X_j)^2 = \Big((X_i - \mu)^2 - 2(X_i - \mu)(X_j - \mu) + (X_j - \mu)^2\Big)\Big((X_k - \mu)^2 - 2(X_k - \mu)(X_j - \mu) + (X_j - \mu)^2\Big)$$

Multiplying out all terms gives

$$\begin{aligned}(X_i - X_j)^2 (X_k - X_j)^2 = {} & (X_i - \mu)^2 (X_k - \mu)^2 - 2(X_i - \mu)(X_j - \mu)(X_k - \mu)^2 + (X_j - \mu)^2 (X_k - \mu)^2 \\ & - 2(X_i - \mu)^2 (X_k - \mu)(X_j - \mu) + 4(X_i - \mu)(X_k - \mu)(X_j - \mu)^2 \\ & - 2(X_k - \mu)(X_j - \mu)^3 + (X_i - \mu)^2 (X_j - \mu)^2 \\ & - 2(X_i - \mu)(X_j - \mu)^3 + (X_j - \mu)^4.\end{aligned}$$

Taking expectations, we get

$$\mathbb{E}\Big((X_i - X_j)^2 (X_k - X_j)^2\Big) = 3(\mu_2)^2 + \mu_4.$$

Thus, for i, j, k all distinct,

$$\mathrm{Cov}((X_i - X_j)^2, (X_k - X_j)^2) = \mu_4 - \sigma^4.$$

Covariance for all distinct indices

Finally, if i, j, k, l are all distinct, then by independence of the X_i,

$$\mathrm{Cov}((X_i - X_j)^2, (X_k - X_l)^2) = 0.$$

Counting the number of terms

- Covariances of the form $\mathrm{Cov}((X_i - X_j)^2, (X_i - X_j)^2)$ occur when $(k = i, l = j)$ or $(k = j, l = i)$ with $i \neq j$, giving $2n(n-1)$ terms.
- Covariances of the form $\mathrm{Cov}((X_i - X_j)^2, (X_k - X_j)^2)$ occur when $(l = j$ or $l = i)$ and i, j, k distinct, or $(k = i$ or $k = j)$ and i, j, l distinct, giving $4n(n-1)(n-2)$ terms.

Hence, summing over all indices,

$$\sum_{i,j,k,l} \mathrm{Cov}((X_i - X_j)^2, (X_k - X_l)^2) = 2n(n-1)(2\mu_4 + 2\sigma^4) + 4n(n-1)(n-2)(\mu_4 - \sigma^4)$$

$$= 4n(n-1)^2(\mu_4 - \frac{n-3}{n-1}\sigma^4).$$

Corollary 4.1 (Asymptotic normality of the empirical variance)**.** *Let $(X_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with mean μ, variance $\sigma^2 > 0$, and finite fourth central moment*

$$\mu_4 = \mathbb{E}\Big[(X - \mu)^4\Big] < \infty.$$

Let

$$\widetilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then,

$$\sqrt{n}\Big(\widetilde{S}^2 - \sigma^2\Big) \xrightarrow{d} \mathcal{N}\Big(0,\ \mu_4 - \sigma^4\Big).$$

Equivalently,

$$\frac{\sqrt{n}\Big(\widetilde{S}^2 - \sigma^2\Big)}{\sqrt{\mu_4 - \sigma^4}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Proof. We start from the identity

$$\widetilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2.$$

Hence,

$$\sqrt{n}\Big(\widetilde{S}^2 - \sigma^2\Big) = \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[(X_i - \mu)^2 - \sigma^2\Big]}_{A_n} - \underbrace{\sqrt{n}(\overline{X} - \mu)^2}_{B_n}.$$

We first study the term A_n. The random variables

$$Y_i = (X_i - \mu)^2 - \sigma^2$$

are i.i.d., centered, and satisfy

$$\mathrm{Var}(Y_i) = \mathbb{E}\Big[(X - \mu)^4\Big] - \sigma^4 = \mu_4 - \sigma^4 < \infty.$$

By the Central Limit Theorem,

$$A_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \xrightarrow{d} \mathcal{N}\Big(0,\ \mu_4 - \sigma^4\Big).$$

Next, consider the term B_n. By the Central Limit Theorem,

$$\sqrt{n}(\overline{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

which implies

$$\sqrt{n}(\overline{X} - \mu)^2 = \frac{(\sqrt{n}(\overline{X} - \mu))^2}{\sqrt{n}} \xrightarrow{\mathbb{P}} 0.$$

Finally, since $A_n \xrightarrow{d} \mathcal{N}(0, \mu_4 - \sigma^4)$ and $B_n \xrightarrow{\mathbb{P}} 0$, Slutsky's theorem yields

$$\sqrt{n}\Big(\widetilde{S}^2 - \sigma^2\Big) \xrightarrow{d} \mathcal{N}\Big(0,\ \mu_4 - \sigma^4\Big).$$

This completes the proof. □

4.5.3 Frequency in a Bernoulli Sample

Let $(X_i)_{i=1}^n$ be a random sample of size n from a Bernoulli distribution with parameter p. Define

$$F = \frac{X_1 + \cdots + X_n}{n}$$

as the sample frequency of the value 1. Then nF follows a binomial distribution with parameters n and p. Hence,

$$\mathbb{E}(F) = p \quad \text{and} \quad \mathrm{Var}(F) = \frac{pq}{n},$$

where $q = 1 - p$.

Therefore, as $n \to \infty$, by the Central Limit Theorem, Therefore, by the Central Limit Theorem,

$$\sqrt{n}(F - p) \xrightarrow{d} \mathcal{N}(0, pq).$$

Therefore, for large n, the distribution of F can be approximated by

$$\mathcal{N}\left(p, \frac{pq}{n}\right).$$

Indeed,

$$\mathbb{E}(F) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = p.$$

Variance of F

$$\mathrm{Var}(F) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \quad (\text{independent } X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{npq}{n^2} = \frac{pq}{n},$$

where $q = 1 - p$.

Exercise 4.1. *Show that*

$$\mathbb{E}\Big(F(1-F)\Big) = pq\left(1 - \frac{1}{n}\right).$$

4.6 Parametric Estimation

The aim of this section is to estimate certain statistical characteristics of a distribution (mean, variance, cumulative distribution function) from a series of observations $x_1, x_2, \ldots, x_n$. This is the inverse problem of sampling.

From the characteristics of a sample, what can we deduce about the characteristics of the population from which it is drawn?

Estimation consists in providing approximate values for the parameters of a population using a sample of n observations drawn from that population. The exact value may not be known, but we aim to give the "best possible value" that can be reasonably assumed.

4.6.1 Point Estimator

We aim to estimate a parameter θ of a population (this could be its mean μ, standard deviation σ, or a proportion p). An estimator of θ is a statistic T (i.e., a function of $(X_1, \ldots, X_n)$) whose realization is considered as a "good value" of the parameter θ.

The estimation of θ associated with this estimator is the observed value during the experiment, that is, the value taken by the function at the observed point $(x_1, \ldots, x_n)$.

Example 4.1. *To estimate the expectation $\mathbb{E}(X)$ of the distribution of X, a natural estimator is the sample mean*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

which produces the estimate $\bar{x}$, the descriptive mean of the observed values.

4.6.2 Quality of an Estimator

Definition 4.7. *The* bias *of an estimator T for a parameter θ is defined as*

$$b_\theta(T) = \mathbb{E}(T) - \theta.$$

An estimator T is said to be unbiased *if*

$$\mathbb{E}(T) = \theta.$$

4.6.3 Convergence and Consistency

An estimator T is said to be *convergent* if $\mathbb{E}(T) \to \theta$ as $n \to \infty$. It is said to be *consistent* if T converges in probability to θ as $n \to \infty$.

Theorem 4.2. *If T is convergent and its variance tends to 0 as $n \to \infty$, then T is consistent.*

Proof. For any real θ and any $\alpha > 0$,

$$|T - \theta| > \alpha \implies |T - \mathbb{E}(T)| > \alpha - |\theta - \mathbb{E}(T)|.$$

If $\lim \mathbb{E}(T) = \theta$, then there exists some N such that for all $n \geq N$, $|\theta - \mathbb{E}(T)| < \frac{\alpha}{2}$. Thus,

$$\mathbb{P}(|T - \theta| > \alpha) \leq \mathbb{P}(|T - \mathbb{E}(T)| > \alpha - |\theta - \mathbb{E}(T)|) \leq \mathbb{P}\left(|T - \mathbb{E}(T)| > \frac{\alpha}{2}\right) \leq \frac{4}{\alpha^2}\mathrm{Var}(T),$$

by the Bienaymé–Chebyshev inequality. The upper bound tends to 0 as $n \to \infty$. □

Remark 4.2. *In general, we can write*

$$T - \theta = (T - \mathbb{E}(T)) + (\mathbb{E}(T) - \theta),$$

where

- $T - \mathbb{E}(T)$ *represents the fluctuations of* T *around its mean,*
- $\mathbb{E}(T) - \theta$ *represents the systematic error (bias).*

4.6.4 Mean Squared Error

The quality of an estimator can also be measured by the *mean squared error* (or quadratic risk), defined as

$$\mathrm{MSE}(T) = \mathbb{E}\Big((T - \theta)^2\Big).$$

Theorem 4.3. *Let* T *be an estimator of the parameter* θ*. Then*

$$\mathbb{E}\Big((T - \theta)^2\Big) = \mathrm{Var}(T) + \Big(\mathbb{E}(T) - \theta\Big)^2.$$

Proof.

$$\begin{aligned}\mathbb{E}\Big((T - \theta)^2\Big) &= \mathbb{E}\Big((T - \mathbb{E}(T) + \mathbb{E}(T) - \theta)^2\Big) \\ &= \mathbb{E}\Big((T - \mathbb{E}(T))^2\Big) + \mathbb{E}\Big((\mathbb{E}(T) - \theta)^2\Big) + 2\,\mathbb{E}\Big((T - \mathbb{E}(T))(\mathbb{E}(T) - \theta)\Big) \\ &= \mathrm{Var}(T) + (\mathbb{E}(T) - \theta)^2,\end{aligned}$$

since $\mathbb{E}(T - \mathbb{E}(T)) = 0$. □

Remark 4.3.

- *Among two unbiased estimators, the "better" one is the one with smaller variance; this is called* efficiency*.*
- *The mean squared error (MSE) criterion is not perfect, but it is preferred over other seemingly more natural criteria, such as the mean absolute error* $\mathbb{E}(|T - \theta|)$*, because it can be expressed in terms of simple concepts like bias and variance and is relatively easy to handle analytically.*

4.6.5 Some Classical Estimators

1. $\overline{X}$ is an unbiased estimator of the mean μ. Its estimate x is the sample mean observed in a realization of the sample.
2. $\widetilde{S}^2$ is a consistent estimator of σ^2 (but biased).
3. $S^2 = \frac{n}{n-1}\widetilde{S}^2$ is an unbiased and consistent estimator of σ^2. Its estimate is
$$s^2 = \frac{n}{n-1}\sigma_e^2,$$
where σ_e is the sample standard deviation observed in a realization of the sample.
4. If p is the frequency of a characteristic, F is an unbiased and consistent estimator of p. Its estimate is denoted f.

Proposition 4.7. *If the mean* μ *of* X *is known, then*

$$T = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

is a better estimator of σ^2 *than* S^2 *(exercise).*

4.6.6 Estimation by the Maximum Likelihood Method

Let X be a real random variable with a parametric distribution (discrete or continuous), and we want to estimate the parameter θ. Define a function f such that

$$f(x;\theta) = \begin{cases} f_\theta(x) & \text{if } X \text{ is a continuous random variable with density } f_\theta, \\ f_{P_\theta}(x) & \text{if } X \text{ is a discrete random variable with probability mass function } f_{P_\theta}. \end{cases}$$

Definition 4.8. *(Likelihood Function)*
For a realization $(x_1, \ldots, x_n)$ of a sample, the likelihood function *of θ is defined by*

$$L(x_1, \ldots, x_n; \theta) = f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Maximum Likelihood Estimator

The method of estimating θ by the value that maximizes L (the likelihood) is called the *maximum likelihood method*. The maximum likelihood estimator (MLE) is defined as

$$\hat{\theta} = \arg\max_\theta L(\theta) = \{\theta : L(\theta) = \sup_\theta L(\theta)\}.$$

Optimization Problem in Maximum Likelihood Estimation

This is an optimization problem. Generally, we use the fact that if L is differentiable and has a global maximum at some value, then the first derivative vanishes there and the second derivative is negative.

Conversely, if the first derivative vanishes at $\theta = \hat{\theta}$ and the second derivative is negative at $\theta = \hat{\theta}$, then $\hat{\theta}$ is a local maximum (not necessarily global) of $L(x_1, \ldots, x_n; \theta)$. It is then necessary to verify that it is indeed a global maximum.

Since the likelihood is positive and the natural logarithm is a monotonically increasing function, it is equivalent and often simpler to maximize the logarithm of the likelihood (the product becomes a sum, which is easier to differentiate).

In Practice

1. **Necessary condition:**

$$\frac{\partial L(x_1, \ldots, x_n; \theta)}{\partial \theta} = 0 \quad \text{or} \quad \frac{\partial \ln L(x_1, \ldots, x_n; \theta)}{\partial \theta} = 0$$

   allows us to find the value $\hat{\theta}$.

2. **Sufficient condition:** $\theta = \hat{\theta}$ is a local maximum if the second derivative condition is satisfied at the critical point:

$$\frac{\partial^2 L(x_1, \ldots, x_n; \theta)}{\partial \theta^2}(\hat{\theta}) \leq 0 \quad \text{or} \quad \frac{\partial^2 \ln L(x_1, \ldots, x_n; \theta)}{\partial \theta^2}(\hat{\theta}) \leq 0.$$

Example 4.2. *Discrete Distribution (Poisson)*

We want to estimate the parameter λ of a Poisson distribution from a sample of size n. The probability mass function is

$$f(x;\lambda) = P_\lambda(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

The likelihood function is

$$L(x_1,\ldots,x_n;\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = e^{-n\lambda}\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}.$$

Since the likelihood is positive, it is simpler to work with the log-likelihood:

$$\begin{aligned}\ln L(x_1,\ldots,x_n;\lambda) &= \ln e^{-n\lambda} + \ln\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}\\ &= -n\lambda + \sum_{i=1}^{n}\ln\lambda^{x_i} - \sum_{i=1}^{n}\ln(x_i!)\\ &= -n\lambda + \ln\lambda\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\ln(x_i!).\end{aligned}$$

The first derivative with respect to λ is

$$\frac{\partial \ln L(x_1,\ldots,x_n;\lambda)}{\partial\lambda} = -n + \frac{\sum_{i=1}^{n}x_i}{\lambda},$$

which vanishes for

$$\hat\lambda = \frac{1}{n}\sum_{i=1}^{n}x_i = \overline{X}.$$

The second derivative is

$$\frac{\partial^2 \ln L(x_1,\ldots,x_n;\lambda)}{\partial\lambda^2} = -\frac{\sum_{i=1}^{n}x_i}{\lambda^2} \le 0,$$

confirming that $\hat\lambda$ is a maximum. Thus, the maximum likelihood estimator (MLE) of λ is

$$\hat\lambda = \overline{X},$$

the sample mean, which is also the best estimator of λ (and of the expectation of a Poisson distribution).

Example 4.3. *Continuous Distribution (Normal)*

We want to estimate the parameters μ and σ of a normal distribution from a sample of size n. The normal density function is

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

Maximum Likelihood Estimation for a Normal Sample

Consider a sample of n independent variables. The likelihood function is

$$f(x_1, \ldots, x_n; \mu, \sigma) = \prod_{i=1}^{n} f(x_i; \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left[-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right].$$

By the König theorem,

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2,$$

where $\bar{x}$ is the sample mean.

Thus, the likelihood function can be rewritten as

$$f(x_1, \ldots, x_n; \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left[-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2\sigma^2} - \frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right].$$

Taking the derivative of the log-likelihood with respect to μ:

$$\frac{\partial}{\partial \mu} \ln L = \frac{\partial}{\partial \mu}\left[-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2\sigma^2} - \frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right] = \frac{n(\bar{x} - \mu)}{\sigma^2} = 0.$$

Solving this gives the maximum likelihood estimator of the mean:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Maximum Likelihood Estimation for the Variance

For the second parameter σ, we calculate the derivative of the log-likelihood with respect to σ:

$$\frac{\partial}{\partial \sigma} \ln L = \frac{\partial}{\partial \sigma}\left[-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2}\right] = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{\sigma^3}.$$

Setting this derivative to zero gives the MLE of the variance:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Verification of Local Maxima

The second derivatives of the log-likelihood are

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\sigma^2} \le 0, \quad \frac{\partial^2 \ln L}{\partial \sigma^2} = \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right).$$

At the MLE point $\hat{\sigma}$,

$$\frac{\partial^2 \ln L}{\partial \sigma^2}(\hat{\sigma}) = \frac{n}{\hat{\sigma}^2} - \frac{3}{\hat{\sigma}^4}\left(n\hat{\sigma}^2 + n(\bar{x} - \mu)^2\right) \le 0,$$

confirming that it is indeed a local maximum.

Remark 4.4.

- *The MLE provides an* unbiased *estimator of the mean,* $\mathbb{E}(\hat{\mu}) = \mu$.
- *However, the MLE of the variance is* biased*:*

$$\mathbb{E}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2.$$

Nevertheless, the estimator is asymptotically unbiased *as* $n \to \infty$.

4.6.7 Confidence Intervals

Instead of providing a single value (point estimator) for a parameter, we seek an interval in which the parameter lies with a controlled (and usually high) probability.

Estimation of a Proportion by Confidence Interval

Consider a population in which the proportion p of a certain category is unknown. We want to estimate this proportion p from a sample of size n, in which the observed frequency of the category is f.

Let F be the random variable that associates, to each sample of size n, the frequency of elements belonging to the chosen category. For sufficiently large n ($n > 30$), F is approximately normally distributed:

$$F \sim N\left(p, \sigma^2\right), \quad \text{with } \sigma = \sqrt{\frac{pq}{n}}, \quad q = 1 - p.$$

Since p is unknown, we use the sample frequency f to estimate the standard deviation:

$$\sigma_0 = \sqrt{\frac{f(1-f)}{n}}.$$

Adjusting for the sample, we write

$$\sigma = \sigma_0\sqrt{\frac{n}{n-1}} = \sqrt{\frac{f(1-f)}{n-1}}.$$

Thus, the standardized random variable

$$Z = \frac{F - p}{\sigma}$$

can be used to construct a confidence interval for p. Thus, Z is approximately standard normal, $Z \sim N(0,1)$. We want to construct a confidence interval for the proportion p, that is, an interval such that the probability that p does *not* belong to this interval is equal to α, where $\alpha \in [0,1]$.

This interval is called a confidence interval with risk α, or equivalently, with confidence level $c = 1 - \alpha$. The risk taken by stating that p belongs to this interval is α, i.e., the probability that p does not belong to the interval is the risk α.

To determine this confidence interval, recall that $z_{\alpha/2}$ is defined as the value such that

$$P(Z > z_{\alpha/2}) = \frac{\alpha}{2},$$

where $Z \sim N(0,1)$. Using the properties of the standard normal distribution, we also have

$$P(Z < -z_{\alpha/2}) = \frac{\alpha}{2} \quad \text{and} \quad P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

Confidence Interval for a Proportion

From the previous discussion, we have

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

Replacing Z by its expression in terms of F and p:

$$P\left(-z_{\alpha/2} < \frac{F-p}{\sigma} < z_{\alpha/2}\right) = 1 - \alpha,$$

$$P\left(-z_{\alpha/2} \cdot \sigma < F - p < z_{\alpha/2} \cdot \sigma\right) = 1 - \alpha,$$

$$P\left(F - z_{\alpha/2} \cdot \sigma < p < F + z_{\alpha/2} \cdot \sigma\right) = 1 - \alpha.$$

Using the estimated standard deviation $\sigma = \sqrt{\frac{f(1-f)}{n-1}}$, we obtain

$$P\left(F - z_{\alpha/2}\sqrt{\frac{f(1-f)}{n-1}} < p < F + z_{\alpha/2}\sqrt{\frac{f(1-f)}{n-1}}\right) = 1 - \alpha.$$

Thus, the confidence interval for the proportion p with confidence level $1 - \alpha$ is

$$\left[f - z_{\alpha/2}\sqrt{\frac{f(1-f)}{n-1}},\ f + z_{\alpha/2}\sqrt{\frac{f(1-f)}{n-1}}\right].$$

Remark: For large n, the difference between n and $n-1$ becomes negligible, so the formula simplifies to

$$\left[f - z_{\alpha/2}\sqrt{\frac{f(1-f)}{n}},\ f + z_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}\right],$$

which is the most commonly used version.

It can be further approximated for a risk $\alpha = 5\%$ and $f \approx 0.5$ by

$$\left[f - \frac{1}{\sqrt{n}},\ f + \frac{1}{\sqrt{n}}\right].$$

Confidence Interval for the Mean

Consider a random variable $X \sim N(\mu, \sigma^2)$ and $X_1, \ldots, X_n$, n independent and identically distributed (i.i.d.) variables with the same distribution as X. Recall that the sample mean and the corrected (or unbiased) sample variance are defined by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Let $z_{\alpha/2}$ be the positive real number such that

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha, \quad Z \sim N(0,1).$$

We have the sample mean $\bar{X}$ follows

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Therefore,

$$1-\alpha = P(-z_{\alpha/2} < Z < z_{\alpha/2}) = P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

Hence, the confidence interval for the mean of a population with known variance σ^2 is

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Confidence Interval for the Mean (Large Sample)

We have

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

so the confidence interval can be written as

$$I = \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

This interval remains valid when the variance σ^2 is unknown, provided the sample is large.

Proposition 4.8. *The variable*

$$\frac{(n-1)S^2}{\sigma^2}$$

follows a χ^2 distribution with $\nu = n-1$ degrees of freedom.

Proof. We have

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2.$$

Dividing by σ^2 gives

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 + \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2.$$

The first term is a sum of squares of independent standard normal variables $N(0,1)$, the second term equals

$$\frac{(n-1)S^2}{\sigma^2},$$

and the last term is the square of a variable following $N(0,1)$ according to previous proposition on the sample mean.

Assuming independence (admitted) between $\bar{X}$ and S^2, we can express this equality in terms of characteristic functions. Let $\varphi(t)$ denote the characteristic function of the random variable

$$\frac{(n-1)S^2}{\sigma^2},$$

and recall the characteristic function of the χ^2 distribution

$$(1-2it)^{-\nu/2}.$$

Using the decomposition

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\frac{(n-1)S^2}{\sigma^2}+\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2,$$

and the independence between $\bar{X}$ and S^2, we obtain the following relation between characteristic functions:

$$(1-2it)^{-n/2}=\varphi(t)\ (1-2it)^{-1/2},$$

or equivalently,

$$\varphi(t)=(1-2it)^{-(n-1)/2}.$$

Therefore, by Proposition 1.2.1, the random variable

$$\frac{(n-1)S^2}{\sigma^2}$$

follows a chi-square distribution with $\nu=n-1$ degrees of freedom.

When only n observations from a normal population with unknown standard deviation are available, the confidence interval for the mean must be modified. In this case, we rely on the sample mean and the estimated standard deviation to construct a confidence interval for the population mean μ.

Consider the statistic

$$T=\frac{\bar{X}-\mu}{S/\sqrt{n}}.$$

Let

$$Q=\frac{(n-1)S^2}{\sigma^2}.$$

Then we can write

$$T=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\cdot\sqrt{\frac{n-1}{Q}}.$$

Since

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1),\quad Q\sim\chi^2_{n-1},$$

and these variables are independent, so

$$T\sim t_{n-1},$$

that is, T follows a Student t distribution with $n-1$ degrees of freedom.

Consequently, the confidence interval for the population mean μ with confidence level $1-\alpha$ is given by

$$\boxed{\bar{X}-t_{\alpha/2,\,n-1}\frac{S}{\sqrt{n}}\ <\ \mu\ <\ \bar{X}+t_{\alpha/2,\,n-1}\frac{S}{\sqrt{n}}}$$

where $t_{\alpha/2,\,n-1}$ denotes the $(1-\alpha/2)$ quantile of the Student t distribution with $n-1$ degrees of freedom.

Confidence Interval with Unknown Variance and Characteristic Functions

We can write this equality in terms of characteristic functions, where ϕ is the characteristic function of $(n-1)S^2/\sigma^2$ and the characteristic function of the χ^2 distribution is used

$$(1-2it)^{n/2} = \phi(t) \cdot (1-2it)^{1/2},$$

or equivalently,

$$\phi(t) = (1-2it)^{(n-1)/2}.$$

Thus, according to previous proposition, we get

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

When we only have n observations from a normally distributed population with unknown standard deviation, the interval is modified. Indeed, we base it on the sample mean and the estimated standard deviation S of the population to construct a confidence interval for the population mean μ.

We have

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1} \quad \text{(Student's } t \text{ distribution with } n-1 \text{ degrees of freedom)},$$

since this variable can be written as a product. Let

$$Q = \frac{(n-1)S^2}{\sigma^2}, \quad \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}/\sqrt{Q/n-1}.$$

This variable follows a Student's t distribution.

Thus, the confidence interval is given by

$$\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}},$$

where $t_{\alpha/2} = t_{\alpha/2;n-1}$, that is, the critical value read from the Student t distribution at a significance level $\alpha/2$ with $\nu = n-1$ degrees of freedom.

Now, we consider the modified empirical variance S^2. we know that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

that is, it follows a chi-square distribution with $\nu = n-1$ degrees of freedom.

Moreover,

$$\mathbb{P}\left(\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{\alpha/2}\right) = 1-\alpha.$$

Hence,

$$1-\alpha = \mathbb{P}\left(\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right).$$

This implies

$$\mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha.$$

Here, $\chi^2_{\alpha/2} = \chi^2_{\alpha/2;(n-1)}$ denotes the quantile read from the chi-square table with $\nu = n-1$ degrees of freedom. We therefore determine the values such that

$$\mathbb{P}\left(K^2 > \chi^2_{\alpha/2;(n-1)}\right) = \frac{\alpha}{2}, \qquad \mathbb{P}\left(K^2 < \chi^2_{1-\alpha/2;(n-1)}\right) = \frac{\alpha}{2}.$$

□

4.7 Notion of Hypothesis Testing

In statistics, reality is described through variables, which consist of numerical observations obtained from data. These variables are frequently compared in order to determine whether they are equal or different, whether they may be regarded as arising from the same population, whether they follow a given probability distribution, or whether they are consistent with a specified statistical model. Since observed data represent only a sample of the population, such comparisons inevitably involve uncertainty.

To address this uncertainty, statisticians have developed a rigorous inferential framework that enables decision-making while quantifying the risk associated with those decisions.

The primary objectives of hypothesis testing are:

- to define clearly and rigorously the framework of statistical analysis;
- to provide a unified and precise methodology applicable to a wide range of problems;
- to determine whether observed differences are statistically *significant* at a prescribed level.

4.7.1 Null Hypothesis and Type I and Type II Errors

The mathematical framework of hypothesis testing is based on probabilistic events. The initial assumption or comparison is formulated as an event within a probabilistic model, which can then be tested and potentially rejected. Typically, the analysis considers only two hypotheses.

The first, called the *null hypothesis* and denoted H_0, represents the scenario where the observed difference is assumed to be zero (or, more precisely, not statistically significant with respect to a threshold known as the *Type I error risk*). The second hypothesis, called the *alternative hypothesis* and usually denoted H_1, is complementary to H_0 and encompasses all other possible outcomes. A hypothesis must specify a value, say θ_0, for a population parameter θ. We therefore test

$$H_0 : \theta = \theta_0.$$

A classical choice for the alternative hypothesis is

$$H_1 : \theta \neq \theta_0,$$

which tests both sides of the equality and is referred to as a *two-sided test*.

Another possible formulation of the hypotheses is

$$H_0 : \theta \geq \theta_0,$$

sometimes also written as $H_0 : \theta = \theta_0$. The corresponding alternative hypothesis is then

$$H_1 : \theta < \theta_0,$$

which tests only one side of the equality; this is referred to as a *one-sided (or one-tailed) test*.

The last case is easily obtained:

$$H_0 : \theta \leq \theta_0 \qquad \text{and} \qquad H_1 : \theta > \theta_0,$$

which is also a one-sided test.

In hypothesis testing, one may either reject the null hypothesis or fail to reject it. In reality, however, the null hypothesis may be either true or false. This leads to a 2×2 contingency table summarizing all possible combinations of decisions and reality:

Decision / Reality	H_0 is true	H_0 is false
Do not reject H_0	True Positive (TP)	False Positive (FP)
Reject H_0	False Negative (FN)	True Negative (TN)

Example 4.4. *Consider a self-administered pregnancy test. The result can be:*

- *A* false negative (FN)*, when the test indicates that the person is not pregnant even though fertilization has occurred.*
- *A* false positive (FP)*, when the test indicates pregnancy even though the person is not pregnant.*

In statistical hypothesis testing, let H_0 denote the null hypothesis that the person is not pregnant*. Then:*

- *Rejecting H_0 when it is true corresponds to a **Type I error** (FP), with risk α.*
- *Failing to reject H_0 when it is false corresponds to a **Type II error** (FN), with risk β.*
- *Correct decisions occur when H_0 is true and we do not reject it (true negative, TN), or when H_0 is false and we reject it (true positive, TP).*

There is no reason to assume that α and β are equal. In practice, α is often taken as 5% *(or* 1% *for more stringent tests), while a common choice for β is* 0.20*.*

- *The probability of rejecting H_0 when it is true is α, called the* significance level *of the test.*
- *The probability of rejecting a false H_0 is $1 - \beta$, called the* power *of the test.*

It is important to note that hypothesis tests do not allow us to accept H_0, only to reject it. Failing to reject H_0 indicates insufficient evidence against it, but does not prove it true.

This leads to the following summary table:

Decision / Reality	*H_0 is true*	*H_0 is false*
Do not reject H_0	*Correct decision (TN)*	*Type II error (FN): risk β*
Reject H_0	*Type I error (FP): risk α*	*Correct decision (TP)*

In practice, statistical hypothesis testing helps determine, for example, whether 0.21 *and* 0.22 *can be considered close, or whether* 15% *and* 20% *are not significantly different, based on the known distribution of the difference.*

4.7.2 Mechanics of Hypothesis Testing

To perform a hypothesis test, there is a strict sequence of steps to follow. The process begins with the formulation of the hypothesis in the relevant field (medical, economic, social, etc.) and its translation into probabilistic events related to the null hypothesis H_0. One must then consider a test statistic (the theoretical distribution of the difference) and choose a decision threshold (significance level) α.

Next, the value of the test statistic is computed from the observed data and compared with the theoretical critical value corresponding to the chosen threshold. From this comparison, a decision is made as to whether or not to reject H_0. Finally, the computation (or lookup) of the associated *p-value*, corresponding to the exceedance of the test statistic, allows for a more refined conclusion regarding whether the observed difference is statistically significant.

The expression "fail to reject H_0" at level α is sometimes abusively replaced by "accept H_0". In this course, however, the risk β is not taken into account, which explains this simplification.

4.8 Test of Independence

4.8.1 Chi-square Test of Independence for Two Qualitative Variables

In most of the tests presented so far, the observations in the sample are assumed to be independent. This assumption is necessary and should often be verified by a statistical test. Such a test is based on a random variable that follows a chi-square distribution; it is therefore called the *chi-square test of independence*.

This test is used to assess the independence of two categorical characteristics within a given population.

Let X and Y be two random variables. The possible values of X are divided into l categories (or classes) $X_1, \ldots, X_l$, and those of Y into k categories $Y_1, \ldots, Y_k$. For each intersection of categories X_i and Y_j, an observed frequency $n_{i,j}$ is recorded. Thus, the total sample size is

$$n = \sum_{i=1}^{l} \sum_{j=1}^{k} n_{i,j}.$$

Tested hypothesis.

$$H_0 : \text{The variables } X \text{ and } Y \text{ are independent.}$$

Test procedure. We construct the contingency table, which is a two-way table. At the intersection of the i-th row and the j-th column, we place the observed frequency $n_{i,j}$. We then compute the marginal frequencies:

$$S_i = \sum_{j=1}^{k} n_{i,j}, \qquad T_j = \sum_{i=1}^{l} n_{i,j},$$

where S_i is the sum of the entries in the i-th row and T_j is the sum of the entries in the j-th column.

	Y_1	$\cdots$	Y_k	Total
X_1	$n_{1,1}$	$\cdots$	$n_{1,k}$	S_1
$\vdots$	$\vdots$		$\vdots$	$\vdots$
X_l	$n_{l,1}$	$\cdots$	$n_{l,k}$	S_l
Total	T_1	$\cdots$	T_k	n

Expected frequencies. The theoretical (expected) frequencies under the null hypothesis are given by

$$C_{i,j} = \frac{S_i T_j}{n}.$$

Remark. Under the null hypothesis H_0, the observed frequencies $n_{i,j}$ are close to the expected frequencies $C_{i,j}$.

Test statistic. The value of the chi-square test statistic is computed as

$$\chi_c^2 = \sum_{i=1}^{l} \sum_{j=1}^{k} \frac{(n_{i,j} - C_{i,j})^2}{C_{i,j}}.$$

We look up the critical value χ_α^2 in the chi-square distribution table with

$$\nu = (l-1)(k-1)$$

degrees of freedom.

Decision rule. If $\chi_c^2 < \chi_\alpha^2$, we accept the null hypothesis H_0; otherwise, we reject it.

Decision rule. Reject the null hypothesis H_0 if $\chi_c^2 > \chi_{1-\alpha,\nu}^2$; otherwise, do not reject H_0.

A posteriori verification of applicability conditions. The expected frequencies must satisfy

$$C_{i,j} \geq 5 \quad \text{for all } i, j.$$

Example 4.5. *To compare the effectiveness of two drugs treating the same disease but with very different costs, the Social Security system conducted a survey on patient recoveries under each treatment. The results are summarized in the following table:*

	Drug	*Generic*
Recovered	*48*	*158*
Not recovered	*6*	*44*

The marginal frequencies are:

	Drug	*Generic*	*Total*
Recovered	*48*	*158*	*206*
Not recovered	*6*	*44*	*50*
Total	*54*	*202*	*256*

The expected frequencies are:

	Drug	*Generic*	*Total*
Recovered	$\frac{206 \times 54}{256}$	$\frac{206 \times 202}{256}$	*206*
Not recovered	$\frac{50 \times 54}{256}$	$\frac{50 \times 202}{256}$	*50*
Total	*54*	*202*	*256*

Finally, we compute the value of the chi-square test statistic:

$$\chi_c^2 = \frac{(48 - 43.45)^2}{43.45} + \frac{(158 - 162.55)^2}{162.55} + \frac{(6 - 10.55)^2}{10.55} + \frac{(44 - 39.45)^2}{39.45} \approx 3.1.$$

The value of the test statistic χ_c^2 *is approximately* 3.1*, whereas the critical value at a* 5% *significance level is* 3.84 *(from the chi-square distribution table with one degree of freedom). We may therefore reasonably conclude that the recovery rate does not depend on the price of the drug, and one may question the relevance of continuing to market the more expensive medication.*

4.8.2 Test of Independence for Two Quantitative Variables: Test of Zero Correlation

Let r be the sample correlation coefficient computed from n pairs of observations drawn from Gaussian populations. The null hypothesis to be tested is

$$H_0 : \rho = 0 \qquad \text{(no correlation between the populations)}$$

at significance level α.

It can be shown that, under H_0, the random variable

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

follows a Student's t distribution with

$$\nu = n - 2$$

degrees of freedom. We therefore compute

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Next, we look up the critical value t_α or $t_{\alpha/2}$ in the Student's t distribution table with $\nu = n - 2$ degrees of freedom, such that

$$\mathbb{P}(T_\nu > t_{\alpha/2}) = \frac{\alpha}{2}.$$

The decision rules are as follows:

- **Two-sided alternative:** $H_1 : \rho \neq 0$ Reject H_0 at significance level α if

$$t \notin [-t_{\alpha/2}, t_{\alpha/2}],$$

with $\nu = n - 2$ degrees of freedom.

- **One-sided alternative:** $H_1 : \rho > 0$ Reject H_0 at significance level α if

$$t > t_\alpha,$$

with $\nu = n - 2$ degrees of freedom.

- **One-sided alternative:** $H_1 : \rho < 0$ Reject H_0 at significance level α if

$$t < -t_\alpha,$$

with $\nu = n - 2$ degrees of freedom.

4.9 Goodness-of-Fit Tests

4.9.1 General Case

Chi-square Goodness-of-Fit Test

Let X be a random variable with distribution L (usually unknown). We want to test whether this distribution fits a known distribution L_0 (e.g., Poisson, Exponential, Normal, etc.) chosen as a suitable model. We therefore test the hypotheses

$$H_0 : L = L_0 \quad \text{versus} \quad H_1 : L \neq L_0.$$

The n observations of X are divided into k classes. Let O_i denote the observed frequency in class i, so that

$$\sum_{i=1}^{k} O_i = n.$$

For each class, the expected frequency is defined as

$$C_i = n \cdot \mathbb{P}(X \in \text{Class}_i \mid L_0).$$

Class	1	2	$\cdots$	k
Observed frequency	O_1	O_2	$\cdots$	O_k
Expected frequency	C_1	C_2	$\cdots$	C_k

The chi-square test statistic is computed as

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i}.$$

This value is compared with the theoretical chi-square critical value χ_α^2 read from the chi-square table with

$$\nu = k - 1 - r$$

degrees of freedom, where r is the number of parameters of L_0 that had to be estimated.

- $r = 0$ if the distribution is completely known or imposed,
- $r = 1$ for a Poisson distribution,
- $r = 2$ for a Normal distribution with unknown mean and variance.

We reject H_0 if

$$\chi_c^2 > \chi_\alpha^2.$$

Example 4.6. *A fish farmer has a pond containing three varieties of trout: common, salmon, and rainbow trout. He wants to know whether he can consider the pond to contain equal numbers of each variety. For this purpose, he randomly samples 399 fish with replacement and obtains the following results:*

Observed frequencies. *The observed frequencies of the three trout varieties are:*

Variety	*Common*	*Salmon*	*Rainbow*
Observed frequency O_i	*145*	*118*	*136*

We want to test whether the trout are equally distributed among the three species. Under the null hypothesis H_0*, we assume a uniform distribution, with probability* $1/3$ *for each class. The expected frequencies are therefore*

$$C_i = 399 \cdot \frac{1}{3} = 133 \quad \textit{for each variety.}$$

Variety	*Common*	*Salmon*	*Rainbow*
Observed O_i	*145*	*118*	*136*
Expected C_i	*133*	*133*	*133*

Chi-square test statistic.

$$\chi_c^2 = \frac{(145-133)^2}{133} + \frac{(118-133)^2}{133} + \frac{(136-133)^2}{133} \approx 2.84.$$

The theoretical chi-square value at a 5% *significance level with*

$$\nu = k - 1 - r = 3 - 1 - 0 = 2$$

degrees of freedom is $\chi_\alpha^2 = 5.99$.

Since $\chi_c^2 < \chi_\alpha^2$*, we cannot reject the null hypothesis. Therefore, there is no statistical evidence against the assumption that the pond contains an equal number of trout of each variety.*

4.10 Kolmogorov-Smirnov Test

As before, the objective is to assess the plausibility of the hypothesis that a sample was drawn from a population with a given distribution. The Kolmogorov-Smirnov test is *nonparametric*: it imposes no constraints on the reference distribution and does not require it to be known in analytical form (although this is the most common case).

Given:

1. A sample of size n of observations of a variable,
2. A reference cumulative distribution function (CDF) $F(x)$,

the Kolmogorov-Smirnov test evaluates the null hypothesis H_0 that the sample comes from a population with CDF $F(x)$.

To do this, it computes a quantity D, called the *Kolmogorov statistic*, whose distribution is known under H_0. The empirical Kolmogorov-Smirnov statistic D_n is defined as

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|,$$

where $F_n(x)$ is the empirical cumulative distribution function, i.e., the proportion of observations less than or equal to x.

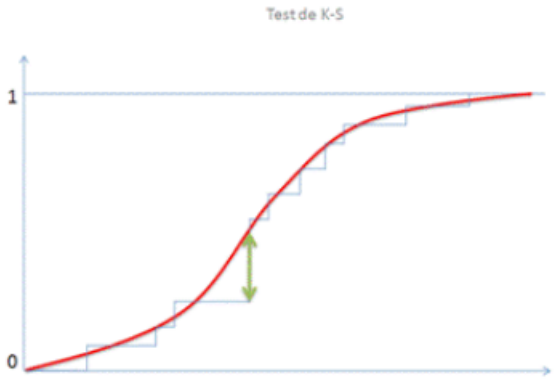

Figure 4.1: Kolmogorov-Smirnov Test

A high value of D indicates that the sample distribution deviates substantially from the reference distribution $F(x)$, and that H_0 is therefore unlikely to be true. More precisely,

Kolmogorov-Smirnov Critical Value. For any constant $c > 0$, the Kolmogorov distribution satisfies

$$\mathbb{P}\left(\sup_x |F_n(x) - F(x)| > \frac{c}{\sqrt{n}} \right) = \alpha(c) = 2 \sum_{r=1}^{\infty} (-1)^{r-1} \exp(-2r^2c^2),$$

where $\alpha(c) = 0.05$ for $c = 1.36$. For $n > 100$, the critical value of the test is approximately of the form $c/\sqrt{n}$.

Usual values of c as a function of the significance level α are:

α	0.20	0.10	0.05	0.02	0.01
c	1.073	1.224	1.358	1.517	1.628

If $D_n > c/\sqrt{n}$, we reject H_0.

Example 4.7. *A new group of foreign tourists is expected at a seaside resort. To understand their preferences, brewers conducted a market survey. At the beginning of the season, twenty tourists were asked to rank their preference among five types of beers, from the least bitter (beer 1) to the most bitter (beer 5). Using a Kolmogorov-Smirnov test, the researcher compares the results to a uniform distribution, i.e., a situation in which each beer would have been preferred by exactly four respondents.*

The survey results are:

$$1, 3, 2, 5, 1, 2, 2, 4, 1, 2, 2, 1, 3, 3, 2, 4, 5, 1, 1, 2.$$

We fix a significance level of 5%. *The null hypothesis* H_0 *to be tested is that the data follows a uniform distribution.*

Next, we summarize the deviations between the observed frequencies and the uniform distribution:

Kolmogorov-Smirnov Test Results.

Class	*Observed frequency*	*Uniform*	*Cumulative observed*	*Cumulative theoretical*	*D*
1	*6*	*4*	*0.30*	*0.20*	*0.10*
2	*7*	*4*	*0.65*	*0.40*	*0.25*
3	*3*	*4*	*0.80*	*0.60*	*0.20*
4	*2*	*4*	*0.90*	*0.80*	*0.10*
5	*2*	*4*	*1.00*	*1.00*	*0.00*

The maximum distance is

$$d = 0.25.$$

For $n = 20$ *and* $\alpha = 5\%$, *the critical value is*

$$c/\sqrt{n} = 0.303.$$

Although these tourists seem to prefer the less bitter beers, we cannot reject the null hypothesis that they have no particular preference.

4.11 Test of Normality

The previous tests are general and can be applied to any distribution. When the distribution to be tested is specifically the normal distribution, we refer to it as a *normality test.*

We test the hypotheses:

H_0 : The data follow a normal distribution, H_1 : The data do not follow a normal distribution.

4.12 Graphical Methods: Henry's Line

Henry's line is a method to visualize the likelihood that a distribution is Gaussian. It also allows one to quickly read off the mean and standard deviation of such a distribution.

Principle. The method consists of plotting the theoretical quantiles against the observed quantiles (a Q-Q diagram).

Let X be a Gaussian random variable with mean μ and variance σ^2, and let Z be a standard normal variable. Then, for each x_i,

$$\mathbb{P}(X < x_i) = \mathbb{P}\Big(\frac{X-\mu}{\sigma} < \frac{x_i-\mu}{\sigma}\Big) = \mathbb{P}(Z < y_i) = \Phi(y_i),$$

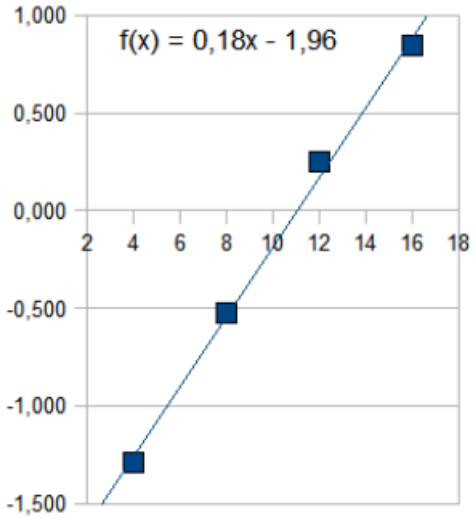

Figure 4.2: Droite de Henry

where $y_i = \frac{x_i - \mu}{\sigma}$ and Φ denotes the cumulative distribution function of the standard normal distribution.

For each value x_i of the variable X, one can compute $\mathbb{P}(X < x_i)$ and then deduce y_i using a table of the function Φ, such that $\Phi(y_i) = \mathbb{P}(X < x_i)$.

If the variable is Gaussian, the points with coordinates (x_i, y_i) are aligned on a straight line with equation

$$y = \frac{x - \mu}{\sigma}.$$

Example 4.8. *During an exam scored out of 20, the results are as follows:*

- *10% of candidates scored less than 4,*
- *30% of candidates scored less than 8,*
- *60% of candidates scored less than 12,*
- *80% of candidates scored less than 16.*

We want to determine whether the distribution of scores is Gaussian and, if so, to estimate its mean and standard deviation. We thus have four values x_i, and for these values, we know $\mathbb{P}(X < x_i)$. Using the Table of the Standard Normal Cumulative Distribution Function*, we determine the corresponding y_i values:*

x_i	$\mathbb{P}(X < x_i) = \Phi(y_i)$	y_i
4	*0.10*	*-1.282*
8	*0.30*	*-0.524*
12	*0.60*	*0.253*
16	*0.80*	*0.842*

The points appear to be aligned. The line intersects the x-axis at $x = 11$ and has a slope of approximately 0.18*, which gives a standard deviation of*

$$\sigma = \frac{1}{0.18} \approx 5.6.$$

This suggests that the distribution is approximately Gaussian with parameters

$$\mu = 11 \quad \textit{and} \quad \sigma = 5.6.$$

Remark 4.5. *One can perform a similar analysis by comparing the theoretical cumulative probabilities with the empirical cumulative probabilities on a graph (comparison of cumulative distribution functions: P-P plot). This is essentially a graphical validation similar to the Kolmogorov-Smirnov test.*

4.13 Jarque-Bera Test (or Bowman-Shelton Test)

The Jarque-Bera test is a test for normality. Define:

$$S = \mathbb{E}\Big[\Big(\frac{X-\mu}{\sigma}\Big)^3\Big] \quad \text{(Skewness: 3rd moment of a standardized variable)}$$

$$K = \mathbb{E}\Big[\Big(\frac{X-\mu}{\sigma}\Big)^4\Big] \quad \text{(Kurtosis: 4th moment of a standardized variable)}$$

Recall that a normal distribution has skewness $S = 0$ and kurtosis $K = 3$. The hypotheses can be written as:

$$H_0 : S = 0 \text{ and } K = 3, \qquad H_1 : S \neq 0 \text{ or } K \neq 3.$$

Note that if H_0 is rejected, the test does not indicate whether the deviation from normality is due to skewness or kurtosis.

The Jarque-Bera statistic is computed as

$$JB = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right].$$

where n is the number of observations. The sample size must be sufficiently large ($n > 50$). The JB statistic asymptotically follows a chi-square distribution with 2 degrees of freedom. If the data are normally distributed, the test statistic is close to 0, and we do not reject H_0 at significance level α.

4.14 Tests on Percentages

4.14.1 Relation Between Tests and Confidence Intervals

A hypothesis test can be viewed as constructing a confidence interval around a value based on a sample and checking whether the value assumed under H_0 falls within this interval, built using a specified risk level. The key quantity in a test is the risk of rejecting H_0, which allows assessing the plausibility of H_0 versus H_1.

The distributions involved in the calculations are the same as those used for confidence intervals, but instead of constructing an interval for each risk level, one compares a fixed part (computed from the observations) with a part that depends solely on the chosen risk level.

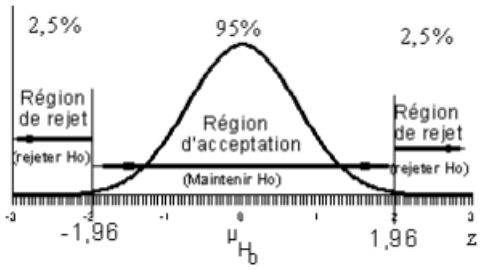

Figure 4.3: test bilatéral pour $\alpha = 5\%$

4.15 Conformity Test

Let p_r be the known proportion of a characteristic in a reference population. We want to test whether the proportion p in another population, from which a sample of size n has been drawn with observed frequency f for this characteristic, corresponds to that of the reference population. The hypotheses are

$$H_0 : p = p_r, \qquad H_1 : p \neq p_r.$$

Let F be the random variable representing the observed frequency in the sample. Under H_0, the distribution of F can be approximated by

$$F \sim N\Big(p_r, \frac{p_r(1-p_r)}{n}\Big).$$

We fix α, the risk of incorrectly rejecting H_0, which corresponds to finding an interval I centered on p_r such that

$$\mathbb{P}(p \notin I) = 1 - \alpha,$$

i.e.,

$$\mathbb{P}\Big(- z_{\alpha/2} < \frac{F - p_r}{\sqrt{p_r(1-p_r)/n}} < z_{\alpha/2}\Big) = 1 - \alpha.$$

We then test whether the calculated value

$$z = \frac{f - p_r}{\sqrt{p_r(1-p_r)/n}}$$

lies within the interval

$$(-z_{\alpha/2},\, z_{\alpha/2}).$$

Decision. We accept H_0 if

$$z \in (-z_{\alpha/2},\, z_{\alpha/2})$$

at significance level α, and we reject H_0 otherwise.

When part of the alternative hypothesis H_1 can be ruled out a priori (e.g., it is impossible or nonsensical), the risk is no longer split on both sides of the inequality but concentrated on one side. This is called a *one-sided test*.

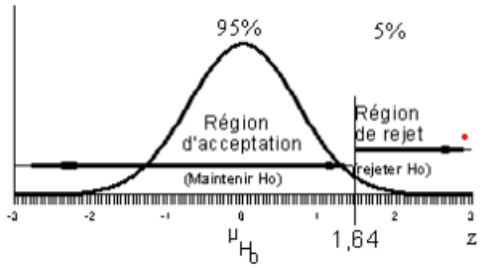

Figure 4.4: test unilatéral pour $\alpha = 5\%$

We then test either

$$H_0 : p = p_r \quad \text{against} \quad H_1 : p > p_r,$$

or

$$H_0 : p = p_r \quad \text{against} \quad H_1 : p < p_r.$$

We reject H_0 if p is much greater than p_r or, respectively, much smaller than p_r.

Unilateral case example. Consider the hypotheses:

$$H_0 : p = p_r, \qquad H_1 : p > p_r.$$

This corresponds to finding an interval I such that

$$\mathbb{P}\Big(F - p_r < z_\alpha \sqrt{\frac{p_r(1-p_r)}{n}}\Big) = 1 - \alpha.$$

We then compare the calculated value

$$z = \frac{f - p_r}{\sqrt{p_r(1-p_r)/n}}$$

with the critical value z_α from the standard normal table.

Decision rule. We accept H_0 if

$$z \in [0, z_\alpha]$$

at significance level α, and we reject H_0 otherwise (i.e., if $z > z_\alpha$).

Second one-sided case. Consider the hypotheses:

$$H_0 : p = p_r, \qquad H_1 : p < p_r.$$

This corresponds to finding an interval I such that

$$\mathbb{P}\Big(-z_\alpha < \frac{F - p_r}{\sqrt{p_r(1-p_r)/n}}\Big) = 1 - \alpha.$$

We then compare the calculated value

$$z = \frac{f - p_r}{\sqrt{p_r(1-p_r)/n}}$$

with the critical value z_α from the standard normal table (using a negative sign for the lower tail, read at risk 2α).

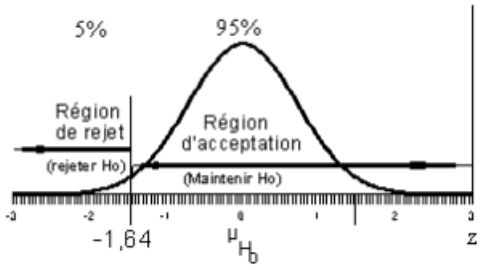

Figure 4.5: test unilatéral pour $\alpha = 5\%$

Decision rule. We accept H_0 if

$$z \in [-z_\alpha, 0]$$

at significance level α, and we reject H_0 otherwise (i.e., if $z < -z_\alpha$).

4.16 Examples of Proportion Tests

1. Two-sided test (A two-sided test rejects values that are too far from the expected proportion.) We want to test whether the claimed proportion of 20% of people listening to a certain radio program corresponds to reality. A survey of 1000 listeners gives a sample proportion of 0.1875.

$$H_0 : p = 0.2, \qquad H_1 : p \neq 0.2$$

A two-sided test is chosen because we have no prior knowledge of the actual proportion. Calculated test statistic:

$$z \approx -0.99$$

2. One-sided test to the right (A right-tailed test rejects values that are too large.) A magician claims that he can often guess the color of a randomly drawn card from a well-shuffled deck with two colors in equal numbers. In a sample of size 100, the magician achieved 64 successes. We want to know what level of risk we take to declare that the magician is not an impostor.

$$H_0 : p = 0.5, \qquad H_1 : p > 0.5$$

Calculated test statistic:

$$z \approx 2.8$$

One-sided test to the left (A left-tailed test rejects values that are too small.) It is known that influenza affects 30% of a population during an epidemic. To test the effectiveness of a flu vaccine, 300 people were vaccinated. At the end of the flu season, 50 vaccinated people contracted the flu. Can this result help assess the vaccine's effectiveness?

$$H_0 : p = 0.3, \qquad H_1 : p < 0.3$$

Calculated test statistic:

$$z \approx -5.04$$

4.17 Test of Homogeneity

Let X be a qualitative variable with two categories (success $X = 1$, failure $X = 0$) observed in two populations with two independent samples drawn from these populations. We observe a frequency f_1 in population 1 of size n_1 and f_2 in population 2 of size n_2.

We assume that the two samples come from populations in which the probabilities of success are identical:

$$H_0 : p_1 = p_2, \qquad H_1 : p_1 \neq p_2.$$

The sampling distribution of the success frequency in population 1, F_1, converges in law to

$$F_1 \sim N\Big(p_1, \sqrt{\frac{p_1 q_1}{n_1}}\Big),$$

and similarly for F_2:

$$F_2 \sim N\Big(p_2, \sqrt{\frac{p_2 q_2}{n_2}}\Big),$$

where nF follows a binomial distribution with parameters (n, p).

Since F_1 and F_2 are independent, we have

$$\mathbb{E}(F_1 - F_2) = \mathbb{E}(F_1) - \mathbb{E}(F_2) = p_1 - p_2,$$
$$\mathrm{Var}(F_1 - F_2) = \mathrm{Var}(F_1) + \mathrm{Var}(F_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}.$$

Under the normal approximation conditions

$$n_1 p_1,\ n_1 q_1,\ n_2 p_2,\ n_2 q_2 > 5 \quad \text{and} \quad n_1, n_2 > 30,$$

the random variable $F_1 - F_2$ follows approximately the normal distribution

$$F_1 - F_2 \sim N\Big(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\Big).$$

Thus, the standardized normal variable

$$Z = \frac{F_1 - F_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

follows, under H_0, the standard normal distribution $N(0, 1)$. The standardized variable is

$$Z = \frac{F_1 - F_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}}.$$

The common success probability p in the two populations is unknown and is estimated from the observed results of the two samples:

$$\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2},$$

where f_1 and f_2 are the observed frequencies in sample 1 and sample 2, respectively.

An observed value z of the random variable Z is calculated as

$$z = \frac{f_1 - f_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

with $\hat{q} = 1 - \hat{p}$.

This value is compared with the critical value z_α from the standard normal table $N(0, 1)$ for a fixed significance level α.

Decision rule.

- If $z \in (-z_{\alpha/2}, z_{\alpha/2})$, we accept H_0: the two samples come from populations with the same success probability p.
- If $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$ (i.e., $z \notin (-z_{\alpha/2}, z_{\alpha/2})$), we reject H_0 at risk level α: the two samples come from populations with different success probabilities p_1 and p_2.

Remark 4.6. *A one-sided test can also be performed (keeping $H_0 : p_1 = p_2$). The z value is calculated in the same way:*

$$z = \frac{f_1 - f_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

and the decision and conclusion are made according to the chosen alternative hypothesis.

One-sided test cases.

- If the alternative hypothesis is $H_1 : p_1 > p_2$ (right-tailed test), reject H_0 at significance level α if $z > z_\alpha$.
- If the alternative hypothesis is $H_1 : p_1 < p_2$ (left-tailed test), reject H_0 at significance level α if $z < -z_\alpha$.

Example 4.9. *We want to test the impact of attendance in tutorial sessions (TD) on success in a statistics exam:*

	Group 1	*Group 2*
TD hours	$18h$	$30h$
Number of students	180	150
Number of students who passed	126	129

We choose a one-sided test because we assume that success is higher with more TD hours. Thus, we test:

$$H_0 : p_1 = p_2 \quad \textit{against} \quad H_1 : p_1 < p_2.$$

Calculations.

$$z = \frac{f_1 - f_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = -3.45, \quad \hat{p} = 0.773$$

Decision. *At $\alpha = 0.05$, the theoretical critical value from the standard normal table is $-z_\alpha = -1.64$ (for a one-sided test). Since $z < -z_\alpha$, H_0 is rejected at the 5% significance level.*

We can also look at the p-value*, which is the minimum risk needed to reject H_0. Here, $z = -3.45$ corresponds to a p-value of approximately $\alpha \approx 0.001$.*

Since $\alpha < 0.001$, the type I error probability (rejecting H_0 when it is true) is very low. Therefore, we can reject H_0 with an almost negligible risk of error.

As expected, the success rate is significantly higher for students who attended more tutorial hours.

4.18 Tests on Means and Variances

4.18.1 Test on Means

Conformity Test

Consider a sample of n observations drawn from a Gaussian population with mean μ. We want to test this mean against a given value μ_0. The conformity test of a mean relative to the null hypothesis

$$H_0 : \mu = \mu_0$$

is performed using the sample mean $\bar{X}$ and the estimated standard deviation s.

We define the test statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}},$$

which follows a Student's t-distribution with $\nu = n - 1$ degrees of freedom.

Decision rules.

- If the alternative hypothesis is $H_1 : \mu \neq \mu_0$ (two-tailed test), reject H_0 at risk α if

  $$t \notin (-t_{\alpha/2}, t_{\alpha/2})$$

  with $\nu = n - 1$ degrees of freedom.

- If the alternative hypothesis is $H_1 : \mu > \mu_0$ (right-tailed test), reject H_0 at risk α if

  $$t > t_\alpha$$

  with $\nu = n - 1$ degrees of freedom.

- If the alternative hypothesis is $H_1 : \mu < \mu_0$ (left-tailed test), reject H_0 at risk α if

  $$t < -t_\alpha$$

  with $\nu = n - 1$ degrees of freedom.

Choice of significance level. There are two ways to proceed for decision-making:

- One can define a risk level *a priori*: a significance level of $\alpha = 5\%$ is often used in many fields (biology, medicine). This level can be lowered if necessary, for instance, if a type I error could have serious consequences.

Decision based on posterior risk. Alternatively, one can decide based on the *posterior risk*: most statistical software provides the minimum risk required to reject H_0. This is denoted as the *p-value*, which is the smallest significance level at which the null hypothesis is rejected. In other words, the p-value is the probability of committing a type I error, i.e., rejecting the null hypothesis when it is true (a false positive).

For example, in a two-tailed test:

$$\text{p-value} = 2P\Big(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > t_{\alpha/2} \;\Big|\; H_0 : \mu = \mu_0\Big).$$

The decision rule is then simplified: reject H_0 when

$$\text{p-value} < \alpha.$$

Known population variance. If the population variance σ^2 is known, the estimated standard deviation is replaced by the true value, and the theoretical value is read from the standard normal table instead of the Student's t table (corresponding to infinite degrees of freedom). In this case, the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

follows a standard normal distribution.

Compare

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

with the corresponding critical value from the standard normal table.

Decision rules.

- If $H_1 : \mu \neq \mu_0$ (two-tailed), reject H_0 at risk α if

$$z \notin (-z_{\alpha/2}, z_{\alpha/2}).$$

- If $H_1 : \mu > \mu_0$ (right-tailed), reject H_0 at risk α if $z > z_\alpha$.
- If $H_1 : \mu < \mu_0$ (left-tailed), reject H_0 at risk α if $z < -z_\alpha$. In this case, the p-value is

$$\text{p-value} = P\Big(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -t_\alpha \;\Big|\; H_0 : \mu = \mu_0\Big).$$

Example 4.10. *A company selling licenses for new e-commerce software advertises that businesses using this software can achieve, on average, a 10% return on their initial investments during the first year. The returns observed for a random sample of 10 of these franchises during their first year of operation are:*

$$6.1, 9.2, 11.5, 8.6, 12.1, 3.9, 8.4, 10.1, 9.4, 8.9$$

Assuming that the population returns are normally distributed, we test the company's claim. Here, $n = 10$, $\bar{x} = 8.82$, $s = 2.4$, $t = -1.55$, and the p-value is calculated as p-value $= T.DIST(1.55, 9, 2) = 0.1546$.

Since p-value $\geq \alpha = 0.05$, we accept H_0 at a 5% significance level. This is considered a two-tailed test (which may be debatable here). For a one-tailed test ($H_1 : r < 10\%$), we would reject H_0 at approximately 10% (p-value ≈ 0.08).

Test of homogeneity: independent populations *We are interested in the difference between the means μ_1 and μ_2 of two populations through two independent samples.*

Suppose the two samples, of sizes n_1 and n_2, are drawn from Gaussian populations with a common (unknown) variance σ^2, i.e.,

$$\sigma_1^2 = \sigma_2^2 = \sigma^2.$$

If the equality of variances cannot be assumed, it should be tested separately.

Consider the variable corresponding to the difference between $\bar{X}_1$ *and* $\bar{X}_2$*. It follows a normal distribution with mean* $(\mu_1 - \mu_2)$ *and variance*

$$V(X_1 - X_2) = V(X_1) + V(X_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right).$$

Thus, the standardized variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

follows the standard normal distribution. When the common variance is unknown, it is estimated by

$$\hat{s}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

The hypothesis test then uses a Student's t*-distribution. The test statistic is defined by*

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

which follows a Student's t*-distribution with*

$$\nu = n_1 + n_2 - 2$$

degrees of freedom.

The null hypothesis (the hypothesis to be tested) and the alternative hypothesis are given by

$$H_0 : \mu_1 = \mu_2 \quad \textit{or equivalently} \quad \mu_1 - \mu_2 = 0.$$

The observed t*-statistic is*

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Decision rule:

- *If the alternative hypothesis is*

  $$H_1 : \mu_1 \neq \mu_2 \quad \textit{(two-sided test)},$$

  then H_0 *is rejected at significance level* α *if*

  $$t \notin \left(-t_{\alpha/2,\nu},\, t_{\alpha/2,\nu} \right),$$

  where $t_{\alpha/2,\nu}$ *is the critical value of the Student's* t*-distribution with* $\nu = n_1 + n_2 - 2$ *degrees of freedom.*

- *If the alternative hypothesis is*

  $$H_1 : \mu_1 > \mu_2 \quad \textit{(right-tailed test)},$$

  then H_0 *is rejected at significance level* α *if*

  $$t > t_{\alpha,\nu}.$$

- *If the alternative hypothesis is*

$$H_1 : \mu_1 < \mu_2 \quad \textit{(left-tailed test)},$$

then H_0 is rejected at significance level α if

$$t < -t_{\alpha,\nu}.$$

In the case where the variances are unknown but assumed to be unequal, the test remains a Student's t*-test, with the number of degrees of freedom given by the Welch–Satterthwaite approximation. The number of degrees of freedom is given by the Welch–Satterthwaite formula:*

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

The test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

The statistic t is compared, at significance level α, with the critical value t_α or $t_{\alpha/2}$ depending on whether a one-sided or two-sided test is performed, using ν degrees of freedom.

4.18.2 Homogeneity test: paired populations

We observe a sample of n pairs of observations,

$$(x_1, y_1), \ldots, (x_n, y_n),$$

drawn from populations with means μ_X and μ_Y, respectively. Define the random variable

$$D = X - Y,$$

and denote by $\bar{D}$ and S_D the sample mean and the estimated standard deviation of the differences between the paired observations.

We assume that the distribution of the differences is Gaussian. The problem is then reduced to testing an observed mean against a theoretical mean. The null hypothesis is

$$H_0 : \mu_X - \mu_Y = D_0.$$

The random variable

$$\frac{\bar{D} - D_0}{S_D/\sqrt{n}}$$

follows a Student's t-distribution with

$$\nu = n - 1$$

degrees of freedom.

The observed t-statistic is computed as

$$t = \frac{\bar{d} - D_0}{s_D/\sqrt{n}}.$$

Decision rule:

- For a two-sided alternative hypothesis, reject H_0 at significance level α if

$$|t| > t_{\alpha/2,\nu}.$$

- For a right-tailed alternative hypothesis, reject H_0 at significance level α if

$$t > t_{\alpha,\nu}.$$

- For a left-tailed alternative hypothesis, reject H_0 at significance level α if

$$t < -t_{\alpha,\nu}.$$

If the alternative hypothesis is

$$H_1 : \mu_X - \mu_Y \neq D_0 \quad \text{(two-sided test)},$$

then H_0 is rejected at significance level α if

$$t \notin \left(-t_{\alpha/2,\nu},\, t_{\alpha/2,\nu}\right),$$

with $\nu = n - 1$ degrees of freedom.

If the alternative hypothesis is

$$H_1 : \mu_X - \mu_Y > D_0 \quad \text{(right-tailed test)},$$

then H_0 is rejected at significance level α if

$$t > t_{\alpha,\nu},$$

with $\nu = n - 1$ degrees of freedom.

If the alternative hypothesis is

$$H_1 : \mu_X - \mu_Y < D_0 \quad \text{(left-tailed test)},$$

then H_0 is rejected at significance level α if

$$t < -t_{\alpha,\nu},$$

with $\nu = n - 1$ degrees of freedom.

4.18.3 Tests on Variances

Goodness-of-fit test

This test consists in comparing an experimental variance with a theoretical variance, or equivalently, in studying the influence of a factor A on a population P using a sample.

In the population, the variance σ_0^2 is assumed to be known. Let E be a sample of size n. From this sample, we compute the sample mean $\bar{x}$ and the sample variance s^2.

The null hypothesis is

$$H_0 : \sigma^2 = \sigma_0^2,$$

that is, the sample variance is consistent with the population variance.

The alternative hypotheses are:

- $H_1 : \sigma^2 \neq \sigma_0^2$ (two-sided test),
- $H_1 : \sigma^2 > \sigma_0^2$ (right-tailed test),
- $H_1 : \sigma^2 < \sigma_0^2$ (left-tailed test).

Assuming that the data are normally distributed in the population, the random variable

$$Y^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

follows a chi-square distribution with

$$\nu = n - 1$$

degrees of freedom.

The observed value

$$y^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

is then compared with the critical values read from the χ^2 distribution table with $\nu = n-1$ degrees of freedom.

Decision:

- In the case of a two-sided test,

If $n \leq 30$ (the table does not contain degrees of freedom greater than 30), we determine a such that

$$\mathbb{P}(\chi^2 < a) = \frac{\alpha}{2}$$

(or equivalently $\mathbb{P}(\chi^2 \geq a) = 1 - \frac{\alpha}{2}$), and b such that

$$\mathbb{P}(\chi^2 \geq b) = \frac{\alpha}{2}.$$

Thus:

- If

$$y^2 \notin (a,\, b),$$

then H_0 is rejected (the experimental variance is not consistent with the theoretical variance; the sample variance differs from that of the population).

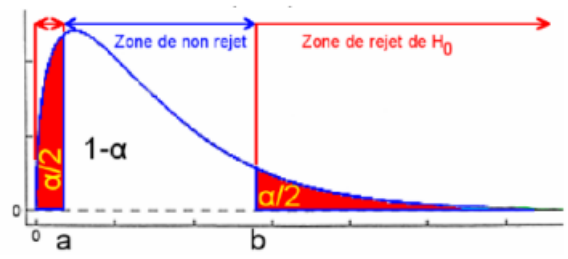

Figure 4.6: Loi χ^2 : Zones de rejet de l'hypothèse nulle

- Otherwise, H_0 is not rejected. There is no evidence to conclude that the experimental variance is not consistent with the population variance.

If $n > 30$, the random variable

$$Z = \sqrt{2\chi^2} - \sqrt{2\nu - 1}$$

is approximately distributed as a standard normal random variable.

The null hypothesis H_0 is rejected when

$$z = \sqrt{2y^2} - \sqrt{2n-3} \notin \left(-z_{\alpha/2},\, z_{\alpha/2}\right).$$

- If

$$H_1 : \sigma^2 > \sigma_0^2,$$

we determine b such that

$$\mathbb{P}(\chi^2 \geq b) = \alpha.$$

If $y^2 > b$, then H_0 is rejected; the experimental variance is greater than that of the population.

- If

$$H_1 : \sigma^2 < \sigma_0^2,$$

we determine a such that

$$\mathbb{P}(\chi^2 \leq a) = \alpha.$$

If $y^2 < a$, then H_0 is rejected; the experimental variance is smaller than that of the population.

Example 4.11. *A company produces electrical devices controlled by a thermostatic system. In practice, the standard deviation of the operating temperature of these controllers should not exceed* 2.0 *degrees.*

For a random sample of $n = 20$ *such devices, the sample standard deviation of the operating temperatures is* $s = 2.36$ *degrees. Perform a hypothesis test at the* 5% *significance level for the null hypothesis that the population standard deviation is* 2.0 *against the alternative that it is in fact larger.*

(Assume that the population distribution is normal.)

The test statistic is

$$\chi_c^2 = \frac{(n-1)s^2}{\sigma_0^2} = 26.45.$$

The critical value is

$$\chi_{\alpha,\nu}^2 = \chi_{0.05,\,19}^2 = 30.14.$$

Since $\chi_c^2 < \chi_{\alpha,\nu}^2$, *we do not reject* H_0.

Homogeneity Test

This test is required to validate the assumption of equality of variances used in Section 9.1.2.

The objective is to compare the variances of two populations P_1 and P_2. Two independent samples are available. Let s_1^2 be the variance of a random sample of size n_1 drawn from a Gaussian population P_1 with variance σ_1^2. Independently, a second random sample of size n_2 with variance s_2^2 is drawn from a Gaussian population P_2 with variance σ_2^2.

The random variable

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

follows an F-distribution, constructed as the ratio of two independent chi-square random variables divided by their respective degrees of freedom. The degrees of freedom are

$$\nu_1 = n_1 - 1 \quad \textit{and} \quad \nu_2 = n_2 - 1.$$

This distribution is denoted by F_{ν_1,ν_2}.

Let H_0 be the null hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2.$$

Under H_0 (that is, when the population variances are equal), the random variable reduces to

$$F = \frac{S_1^2}{S_2^2}.$$

Thus, we compute the ratio

$$f = \frac{s_1^2}{s_2^2}.$$

In practical applications, in order to compare correctly with the theoretical critical values from the F-distribution table, the samples are arranged so that this ratio is greater than 1, by interchanging the roles of the two samples if necessary.

Decision

- *If*

  $$H_1 : \sigma_1^2 > \sigma_2^2,$$

  we determine f_α such that

  $$\mathbb{P}\Big(F_{(n_1-1,\,n_2-1)} \geq f_\alpha\Big) = \alpha.$$

  If $f > f_\alpha$, then H_0 is rejected.

- *If*

  $$H_1 : \sigma_1^2 \neq \sigma_2^2,$$

  we determine $f_{\alpha/2}$ such that

  $$\mathbb{P}\Big(F_{(n_1-1,\,n_2-1)} \geq f_{\alpha/2}\Big) = \frac{\alpha}{2}.$$

  If $f > f_{\alpha/2}$, then H_0 is rejected. Although this decision rule appears to be one-sided, it in fact corresponds to a two-sided test at significance level α, the complementary case being tested through the condition $f > 1$.

Example 4.12. *It is assumed that the total sales of a company should vary more in an industry with active price competition than in a duopoly with tacit collusion.*

In a study of the goods-producing industry, it was observed that over a four-year period of active price competition, the variance of the company's total sales was 114.09*. During the following seven years, in which tacit collusion can be assumed, the variance was* 16.08 *Assume that the data can be considered as independent random samples drawn from two normal distributions. At the* 5% *significance level, test the null hypothesis that the two population variances are equal against the alternative hypothesis that the variance of total sales is higher during the years of active price competition.*

$$f = 7.095, \qquad f_\alpha = 4.76 \quad (\nu_1 = 3,\ \nu_2 = 6)$$

where

$$f_\alpha = F^{-1}(0.05; 3, 6) = \texttt{INVERSE.F}(0.05; 3; 6).$$

Since $f > f_\alpha$, the null hypothesis H_0 is rejected.

4.18.4 Exercise (Brambles)

The size of bramble leaves was measured in order to determine whether there is a difference between the size of leaves growing in full sunlight and those growing in the shade. The results are as follows (leaf width in cm):

Sunlight:

$$6.0,\ 4.8,\ 5.1,\ 5.5,\ 4.1,\ 5.3,\ 4.5,\ 5.1$$

Shade:

$$6.5,\ 5.5,\ 6.3,\ 7.2,\ 6.8,\ 5.5,\ 5.9,\ 5.5$$

The sample statistics are:

$$\bar{x}_1 = 5.05, \quad s_1 = 0.59, \quad n_1 = 8,$$

$$\bar{x}_2 = 6.15, \quad s_2 = 0.65, \quad n_2 = 8.$$

The pooled standard deviation is

$$s = 0.62,$$

and the test statistic is

$$t = 3.55.$$

The critical value for a two-sided test at significance level $\alpha = 0.05$ is

$$t_{\alpha/2} = 2.145.$$

4.19 Exercises

Exercise 4.2. *Let $X_1, \ldots, X_{20}$ be independent and identically distributed random variables following a normal distribution $\mathcal{N}(\mu, \sigma^2)$. We want to test the hypothesis*

$$H_0 : \sigma^2 = \sigma_0^2.$$

To this end, we use the estimator

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *denotes the sample mean.*

1. *What is the distribution of*

$$\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

under the null hypothesis H_0*?*

2. *Compute the critical value* c_α *for a one-sided test*

$$H_1 : \sigma^2 > \sigma_0^2$$

at the 5% *significance level.*

Solution

1. Under the null hypothesis H_0, the random variable

$$\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

follows a chi-square distribution with $(n-1)$ degrees of freedom, that is,

$$\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2_{n-1}.$$

2. We seek a critical value c_α such that

$$\mathbb{P}_{H_0}\Big(\widehat{\sigma}^2 < c_\alpha\Big) = 0.95.$$

Using the result of point 1, we obtain

$$\begin{aligned}
\mathbb{P}_{H_0}\Big(\widehat{\sigma}^2 < c_\alpha\Big) &= \mathbb{P}_{H_0}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 < c_\alpha\right) \\
&= \mathbb{P}_{H_0}\left(\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 < \frac{(n-1)\,c_\alpha}{\sigma_0^2}\right) \\
&= \mathbb{P}_{H_0}\left(D < \frac{(n-1)\,c_\alpha}{\sigma_0^2}\right),
\end{aligned}$$

where $D \sim \chi^2_{n-1}$.

Using the 95% quantile $\chi^2_{n-1,\,0.95}$ of the chi-square distribution with $(n-1)$ degrees of freedom, the critical value is therefore given by

$$c_\alpha = \frac{\sigma_0^2}{n}\,\chi^2_{n-1,\,0.95} \approx 1.59\,\sigma_0^2.$$

Exercise 4.3. *We consider a sample of* $n = 10$ *independent random variables following the same normal distribution* $\mathcal{N}(\mu, \sigma^2)$*.*

1. *The variance is known and equal to* $\sigma^2 = 2.5$. *The observed value of the sample mean* $\overline{X}$ *is* 1.15. *At the* 95% *significance level, can we accept the null hypothesis*

$$H_0 : \mu = 0.1\,?$$

2. *In fact, the variance* σ^2 *is unknown. The observed value of the corrected sample variance is*

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = 2.7.$$

Can we accept the null hypothesis $H_0 : \mu = 0.1$ *in this case?*

Solution

1. Under the null hypothesis H_0, each random variable X_k follows the normal distribution

$$\mathcal{N}(0.1,\, 2.5).$$

Therefore, the sample mean $\overline{X}$ follows the normal distribution

$$\overline{X} \sim \mathcal{N}\Big(0.1,\, \frac{2.5}{n}\Big) = \mathcal{N}(0.1,\, 0.25).$$

We use the sample mean as the test statistic. In the absence of additional information, the alternative hypothesis is

$$H_1 : \mu \neq 0.1.$$

The standardized test statistic is

$$U_n = \frac{\overline{X} - 0.1}{\sqrt{2.5/n}} = \frac{\overline{X} - 0.1}{\sqrt{0.25}},$$

which follows a standard normal distribution $\mathcal{N}(0,1)$ under H_0. Since

$$\mathbb{P}(|U_n| < 1.96) = 0.95,$$

the acceptance region for H_0 is determined by

$$\mathbb{P}\left(\left|\frac{\overline{X} - 0.1}{\sqrt{2.5/n}}\right| < 1.96\right) = 0.95.$$

This yields the acceptance interval

$$-0.88 < \overline{X} < 1.08.$$

Since the observed value $\overline{X} = 1.15$ does not belong to this interval, we reject H_0 at the 95% significance level.

2. Under the null hypothesis H_0, the mean μ is known but the variance σ^2 is unknown. We therefore replace σ^2 by the corrected sample variance S_n^2. The test statistic

$$U_n = \frac{\overline{X} - 0.1}{\sqrt{S_n^2/n}}$$

follows a Student's t distribution with $n-1 = 9$ degrees of freedom under H_0. From the Student's t table, we know that

$$\mathbb{P}(|T_n| < 2.26) = 0.95,$$

where T_n follows a Student's t distribution with $n - 1 = 9$ degrees of freedom.

Therefore, the acceptance region of the null hypothesis H_0 is defined by

$$\mathbb{P}\left(\left|\frac{\overline{X} - 0.1}{\sqrt{S_n^2/n}}\right| < 2.26\right) = 0.95.$$

This yields the acceptance interval

$$-1.03 < \overline{X} < 1.23,$$

which contains the observed value of the sample mean.

Hence, at the 95% significance level, we can accept the null hypothesis H_0.

Appendix:Integration Review

In this appendix, we recall several integration results used throughout the course. The framework is that of the Lebesgue integral. We adopt probabilistic notation: $(\Omega, \mathcal{T}, \mathbb{P})$ is a probability space, that is, a measure space such that $\mathbb{P}(\Omega) = 1$.

A.1 Convergence Theorems

Theorem 4.4 (Monotone Convergence Theorem)**.** *Let $f_n : \Omega \to \mathbb{R}_+$ be a sequence of nonnegative measurable functions. Suppose that, for almost every $\omega \in \Omega$, the sequence $\big(f_n(\omega)\big)_{n\in\mathbb{N}}$ is increasing, and denote by $f(\omega)$ its pointwise limit. Then*

$$\lim_{n\to\infty} \int_\Omega f_n(\omega)\, d\mathbb{P}(\omega) = \int_\Omega f(\omega)\, d\mathbb{P}(\omega).$$

Remark. The values of the integrals may be equal to $+\infty$.

Special case. Applying this theorem to a sequence of indicator functions $\mathbf{1}_{A_n}$, where $(A_n)_{n\in\mathbb{N}}$ is an increasing sequence of sets (with respect to inclusion), we obtain

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

Theorem 4.5 (Fatou's Lemma)**.** *Let $f_n : \Omega \to \mathbb{R}_+$ be a sequence of nonnegative measurable functions. Then*

$$\int_\Omega \liminf_{n\to\infty} f_n(\omega)\, d\mathbb{P}(\omega) \le \liminf_{n\to\infty} \int_\Omega f_n(\omega)\, d\mathbb{P}(\omega).$$

Theorem 4.6 (Dominated Convergence Theorem)**.** *Let $f_n : \Omega \to \mathbb{R}$ be a sequence of measurable functions that converges almost everywhere to a function f. Assume that the sequence (f_n) is dominated by an integrable function $g : \Omega \to \mathbb{R}_+$, that is,*

$$|f_n(\omega)| \leq g(\omega) \quad \text{for almost every } \omega \in \Omega.$$

Then

$$\lim_{n\to\infty} \int_\Omega f_n(\omega)\, d\mathbb{P}(\omega) = \int_\Omega f(\omega)\, d\mathbb{P}(\omega).$$

Remark. We have assumed that $\mathbb{P}(\Omega) = 1$, so any bounded sequence (f_n) is dominated by a constant function, which is integrable. The theorem therefore applies to such sequences.

Theorem 4.7 (Interchange of Sum and Integral: Nonnegative Case)**.** *Let $(f_n)_{n\geq 1}$ be a sequence of nonnegative measurable functions $f_n : \Omega \to \mathbb{R}_+$. Then*

$$\int_\Omega \sum_{n=1}^{\infty} f_n(\omega)\, d\mathbb{P}(\omega) = \sum_{n=1}^{\infty} \int_\Omega f_n(\omega)\, d\mathbb{P}(\omega).$$

Remark. The value of the series may be equal to $+\infty$.

Theorem 4.8 (Interchange of Sum and Integral: Integrable Case)**.** *Let $(f_n)_{n\geq 1}$ be a sequence of measurable functions $f_n : \Omega \to \mathbb{R}$. Assume that*

$$\sum_{n=1}^{\infty} \int_\Omega |f_n(\omega)|\, d\mathbb{P}(\omega) < +\infty.$$

Then

$$\int_\Omega \sum_{n=1}^{\infty} f_n(\omega)\, d\mathbb{P}(\omega) = \sum_{n=1}^{\infty} \int_\Omega f_n(\omega)\, d\mathbb{P}(\omega).$$

Remark. The series appearing in the right-hand side is convergent.

A.2 Integrals Depending on a Parameter

Theorem 4.9 (Continuity Under the Integral Sign)**.** *Let I be an interval of $\mathbb{R}$ and let $f : I \times \Omega \to \mathbb{R}$ be a measurable function such that:*

- *for $\mathbb{P}$-almost every $\omega \in \Omega$, the mapping $t \mapsto f(t, \omega)$ is continuous on I;*
- *there exists an integrable function $g : \Omega \to \mathbb{R}$ such that, for all $t \in I$,*

$$|f(t, \omega)| \leq g(\omega) \quad \text{for } \mathbb{P}\text{-almost every } \omega \in \Omega.$$

Then the function

$$t \longmapsto \int_\Omega f(t, \omega)\, d\mathbb{P}(\omega)$$

is continuous on I. In particular, for every $t_0 \in I$,

$$\lim_{t\to t_0} \int_\Omega f(t, \omega)\, d\mathbb{P}(\omega) = \int_\Omega f(t_0, \omega)\, d\mathbb{P}(\omega).$$

Theorem 4.10 (Derivative Under the Integral Sign)**.** *Let I be an interval of $\mathbb{R}$ and let $f : I \times \Omega \to \mathbb{R}$ be a measurable function such that:*

- *for every $t \in I$, the mapping $\omega \mapsto f(t,\omega)$ is integrable;*
- *for $\mathbb{P}$-almost every $\omega \in \Omega$, the function $t \mapsto f(t,\omega)$ is differentiable for all $t \in I$;*
- *there exists an integrable function $g : \Omega \to \mathbb{R}$ such that, for all $t \in I$,*

$$\left|\frac{\partial}{\partial t} f(t,\omega)\right| \le g(\omega) \quad \text{for } \mathbb{P}\text{-almost every } \omega \in \Omega.$$

Then, for every $t \in I$,

$$\frac{d}{dt}\int_\Omega f(t,\omega)\, d\mathbb{P}(\omega) = \int_\Omega \frac{\partial}{\partial t} f(t,\omega)\, d\mathbb{P}(\omega).$$

A.3 Multiple Integrals

Let $(\Omega_1, \mathcal{T}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{T}_2, \mathbb{P}_2)$ be probability spaces.

Theorem 4.11 (Fubini's Theorem: Nonnegative Case)**.** *Let $f : \Omega_1 \times \Omega_2 \to \mathbb{R}_+$ be a $\mathcal{T}_1 \otimes \mathcal{T}_2$-measurable nonnegative function. Then*

$$\begin{aligned}\int_{\Omega_1\times\Omega_2} f(\omega_1,\omega_2)\, d(\mathbb{P}_1\otimes\mathbb{P}_2)(\omega_1,\omega_2) &= \int_{\Omega_2}\left(\int_{\Omega_1} f(\omega_1,\omega_2)\, d\mathbb{P}_1(\omega_1)\right) d\mathbb{P}_2(\omega_2)\\ &= \int_{\Omega_1}\left(\int_{\Omega_2} f(\omega_1,\omega_2)\, d\mathbb{P}_2(\omega_2)\right) d\mathbb{P}_1(\omega_1).\end{aligned}$$

Remark. The integrals may take the value $+\infty$.

Theorem 4.12 (Fubini's Theorem: Integrable Case)**.** *Let $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ be a $\mathcal{T}_1 \otimes \mathcal{T}_2$-measurable function. Assume that*

$$\int_{\Omega_1\times\Omega_2} |f(\omega_1,\omega_2)|\, d(\mathbb{P}_1\otimes\mathbb{P}_2)(\omega_1,\omega_2) < +\infty.$$

Then

$$\begin{aligned}\int_{\Omega_1\times\Omega_2} f(\omega_1,\omega_2)\, d(\mathbb{P}_1\otimes\mathbb{P}_2)(\omega_1,\omega_2) &= \int_{\Omega_2}\left(\int_{\Omega_1} f(\omega_1,\omega_2)\, d\mathbb{P}_1(\omega_1)\right) d\mathbb{P}_2(\omega_2)\\ &= \int_{\Omega_1}\left(\int_{\Omega_2} f(\omega_1,\omega_2)\, d\mathbb{P}_2(\omega_2)\right) d\mathbb{P}_1(\omega_1).\end{aligned}$$

Remark. The function f belongs to $L^1(\Omega_1 \times \Omega_2)$.

Theorem 4.13 (Change of Variables)**.** *Let U, V be two open subsets of $\mathbb{R}^d$, and let $\varphi : U \to V$ be a C^1-diffeomorphism. Let $f : V \to \mathbb{R}$ be a function measurable with respect to the Lebesgue measure on V, and assume that f is nonnegative or integrable. Then*

$$\int_U f(\varphi(u))\, J_\varphi(u)\, du = \int_V f(v)\, dv,$$

where $J_\varphi(u)$ denotes the Jacobian of φ,

$$J_\varphi(u) = \left|\det\Big(D\varphi(u)\Big)\right|.$$

Remark. In the case of the change of variables to polar coordinates, $u = (r,\theta)$ and $v = \varphi(u) = (r\cos\theta, r\sin\theta)$, we have

$$du = dr\, d\theta \quad \text{and} \quad J_\varphi(r,\theta) = r.$$

Lp Spaces

Reminder

For a measurable function $f : \Omega \to \mathbb{R}$ and $1 \leq p < \infty$, we define the L^p-norm by

$$\|f\|_p = \left(\int_\Omega |f|^p \, d\mathbb{P} \right)^{1/p},$$

and for $p = \infty$,

$$\|f\|_\infty = \inf\{M \geq 0 \mid |f(\omega)| \leq M \text{ for } \mathbb{P}\text{-almost every } \omega \in \Omega\}.$$

Theorem 4.14 (Normal Convergence in L^p)**.** *Let $p \in [1, \infty]$ and (f_n) be a sequence of functions in $L^p(\Omega)$ such that*

$$\sum_{n \in \mathbb{N}} \|f_n\|_p < \infty.$$

Then the series $\sum f_n$ converges almost everywhere and in L^p-norm to a function $f \in L^p(\Omega)$.

Theorem 4.15 (Inclusion of L^p Spaces)**.** *Let $p, q \in \mathbb{R}$ with $1 \leq p \leq q \leq \infty$. Then*

$$L^\infty(\Omega) \subset L^q(\Omega) \subset L^p(\Omega) \subset L^1(\Omega).$$

Moreover, for any measurable function $f : \Omega \to \mathbb{R}$,

$$\|f\|_1 \leq \|f\|_p \leq \|f\|_q \leq \|f\|_\infty.$$

Remark. *The case $p = 2$ is particularly important:*

$$L^\infty(\Omega) \subset L^2(\Omega) \subset L^1(\Omega).$$

Theorem 4.16 (Subsequence Extraction)**.** *Let $(f_n)_{n \in \mathbb{N}}$ be a sequence in $L^p(\Omega)$ converging in L^p-norm to a function $f \in L^p(\Omega)$. Then there exists a subsequence (f_{n_k}) converging almost everywhere to f.*

Remark. In general, L^p-convergence does not imply almost everywhere convergence.

Uniqueness of Measures

Theorem 4.17 (Uniqueness of Measures)**.** *Let Ω be a set, $\mathcal{T}$ a σ-algebra on Ω, and $\mathcal{A}$ a collection of elements of $\mathcal{T}$, stable under finite intersections and containing Ω. Consider two finite measures μ_1, μ_2 defined on $\mathcal{T}$. If μ_1 and μ_2 coincide on $\mathcal{A}$, then they coincide on the σ-algebra generated by $\mathcal{A}$.*

Remark. This theorem can be proved using the monotone class lemma. It is often applied when $\mathcal{A}$ is an algebra of subsets of $\mathcal{T}$, i.e., stable under finite unions, complements, and containing Ω. A classical example is $\Omega = \mathbb{R}^n$ and $\mathcal{A}$ the set of open rectangles of the form $Q_i = \prod_i (a_i, b_i)$.

Corollary on Uniqueness of Probability Measures

Corollary 4.2. *Let μ_1, μ_2 be two probability measures defined on the Borel σ-algebra of $\mathbb{R}^n$. If for every open rectangle $R \subset \mathbb{R}^n$,*

$$\mu_1(R) = \mu_2(R),$$

then $\mu_1 = \mu_2$. Consequently, for any Borel measurable function $f : \mathbb{R}^n \to \mathbb{R}$, positive or μ_1-integrable,

$$\int_{\mathbb{R}^n} f \, d\mu_1 = \int_{\mathbb{R}^n} f \, d\mu_2.$$

Remark. In other words, to verify the above equality, it suffices to check it for indicator functions of open rectangles. The result also holds if one considers closed rectangles or half-open rectangles of the form $Q_i = \prod_i (a_i, b_i]$ instead of open rectangles.

Inequalities in L^p Spaces

Theorem 4.18 (Minkowski Inequality)**.** *Let $p \in [1, \infty]$ and $f, g \in L^p(\Omega)$. Then*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

Remark. This is the triangle inequality for L^p norms.

Theorem 4.19 (Cauchy–Schwarz Inequality)**.** *Let $f, g \in L^2(\Omega)$. Then fg is integrable and*

$$\int_\Omega fg \, d\mathbb{P} \leq \|f\|_2 \, \|g\|_2.$$

Remark. Equality holds if and only if f and g are proportional.

Theorem 4.20 (Hölder's Inequality)**.** *Let $p, q \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$, and let $f \in L^p(\Omega)$, $g \in L^q(\Omega)$. Then $fg \in L^r(\Omega)$ and*

$$\|fg\|_r \leq \|f\|_p \, \|g\|_q.$$

Remark. The Cauchy–Schwarz inequality corresponds to $p = q = 2$, $r = 1$.

Theorem 4.21 (Jensen's Inequality)**.** *Recall that $\mathbb{P}(\Omega) = 1$. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function and $f : \Omega \to \mathbb{R}$ such that f and $\varphi \circ f$ are integrable. Then*

$$\varphi\Big(\int_\Omega f \, d\mathbb{P} \Big) \leq \int_\Omega \varphi \circ f \, d\mathbb{P}.$$

Fourier Inversion Formula

The theorem of interest is a pointwise version of the Fourier inversion formula, which is the analogue of Dirichlet's theorem for Fourier series. It holds for real- or complex-valued functions. We state here a slightly more general version than the one used in the course.

Fourier Transform Convention

The Fourier transform of a function f is defined by

$$\hat{f}(t) = \int_{\mathbb{R}} e^{-itx} f(x)\, dx.$$

When f is integrable, its Fourier transform $\hat{f}$ is continuous and tends to 0 at infinity, by the Riemann–Lebesgue lemma.

Lemma 4.1 (Riemann–Lebesgue)**.** *Let $f \in L^1(\mathbb{R})$. Then*

$$\lim_{|t|\to\infty} \int_{\mathbb{R}} e^{-itx} f(x)\, dx = 0.$$

Remark. This lemma can be proved explicitly when f is the indicator function of an interval. In the general case, it suffices to approximate f in L^1-norm by a finite linear combination of indicator functions.

Theorem 4.22 (Fourier Inversion Formula)**.** *Let $f \in L^1(\mathbb{R})$ and $t \in \mathbb{R}$. Suppose that f admits a left and a right limit at t, denoted $f(t^-)$ and $f(t^+)$, and that f is differentiable from the left and the right at t. Then*

$$\frac{1}{2}\big[f(t^-) + f(t^+)\big] = \lim_{A\to\infty} \frac{1}{2\pi} \int_{-A}^{A} e^{itx} \hat{f}(x)\, dx.$$

If f is integrable and of class C^1, and if $\hat{f}$ is integrable, then

$$f(t) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{itx} \hat{f}(x)\, dx \quad \textit{for all } t \in \mathbb{R}.$$

Remark. The function $\hat{f}$ is integrable if $f \in C^2$, integrable, with second derivative f'' integrable. Indeed, in that case, $\hat{f}$ is continuous and bounded by a constant times $1/|t|^2$, as shown by

$$\hat{f}(t) = -\frac{1}{t^2} \widehat{f''}(t), \quad t \in \mathbb{R}^*,$$

which follows by integration by parts.

This also holds if $f \in C^1$, integrable, with square-integrable derivative. With some work, the Fourier transform can be extended by density to square-integrable functions.

Bibliography

[1] J. Bass. *Éléments de calcul de probabilités*. Masson, 1974.

[2] P. Billingsley. *Probability and Measure*. John Wiley & Sons, Inc., 1995.

[3] D. Birkes and Y. Dodge. *Alternative Methods of Regression*. Wiley, 1993.

[4] M. Briane and G. Pagès. *Théorie de l'intégration : cours et exercices*. Vuibert, 2014. ISBN 978-2-311-40226-1.

[5] G. Calot. *Cours de calcul des probabilités*. Dunod, 1967.

[6] P. Dagnélie. *Statistique théorique et appliquée*. De Boeck Université, 1998.

[7] Y. Dodge. *Analyse de régression appliquée*. Dunod, Paris, 1999.

[8] Y. Dodge. *Premiers pas en statistique*. Springer, 1999.

[9] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2006.

[10] R. Durrett. *Elementary Probability for Applications*. Cambridge University Press, 2009.

[11] D. Foata and A. Fuchs. *Calcul des probabilités*. Dunod, Paris,1998.

[12] D. Foudrinier. *Statistique inférentielle : cours et exercices*. Dunod, Paris, 2002.

[13] J. Guégand and J. P. Gavini. *Probabilités*. 1998.

[14] J. Jacod and P. Protter. *L'essentiel en théorie des probabilités*. Cassini, Paris. 2003.

[15] R. A. Johnson and G. K. Bhattacharyya. *Statistics: Principles and Methods*. Wiley, 1996.

[16] A. Krief and S. Lévy. *Calcul des probabilités*. Hermann, 1972.

[17] J. P. Lecoutre, S. Legait, and P. Tassi. *Statistique : exercices corrigés et rappels de cours*. Masson, 1987.

[18] J. P. Lecoutre. *Statistique et probabilités : manuel et exercices corrigés*, 4^{e} édition. Masson, 2009.

[19] A. Mattei. *Inférence et décision statistiques : théorie et applications à la gestion des affaires*. Peter Lang, 2000.

[20] S. M. Ross. *Initiation aux probabilités.* Presses Polytechniques et Universitaires Romandes, 1994.

[21] G. Saporta. *Probabilités, analyse des données et statistique.* Technip, 1990.

[22] P. Tassi. *Méthodes statistiques.* Economica, 2004.