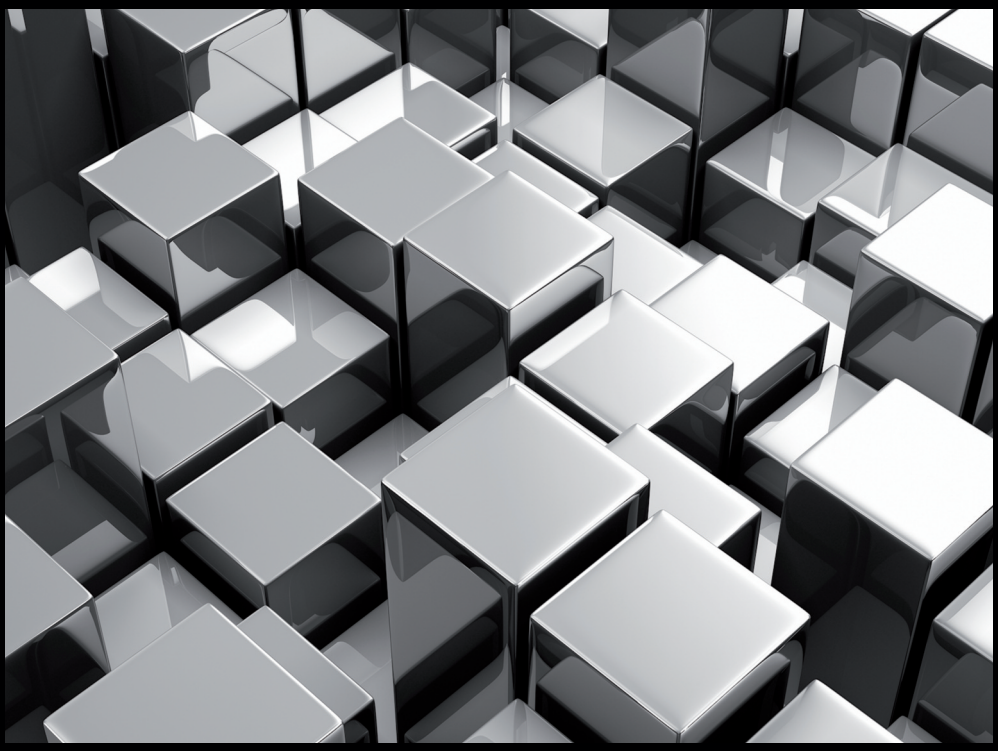
Enseignement des mathématiques

Théorie des probabilités

Cours d'introduction avec application à la statistique mathématique

Charles-Edouard Pfister



Presses polytechniques et universitaires romandes

Enseignement des mathématiques

Théorie des probabilités

Cours d'introduction avec application à la statistique mathématique

Cet ouvrage constitue une première introduction à la théorie des probabilités. A la fois rigoureux et didactique, il présente l'ensemble des notions et outils de base, et de manière approfondie, les deux théorèmes fondamentaux que sont la loi des grands nombres et le théorème de la limite centrale. Certains sujets, comme celui de l'espérance d'une variable aléatoire, sont traités plus en détail qu'usuellement dans un texte d'introduction. La théorie ainsi développée est appliquée d'une part à l'étude des chaînes de Markov, marches aléatoires et au modèle d'Ising, et d'autre part à des sujets classiques de statistique mathématique, estimations, tests, populations normalement distribuées. Les résultats sont démontrés dans leur intégralité, et de nombreux exemples jalonnent le texte.

Cette référence s'adresse principalement aux étudiants de physique ou de mathématiques des universités et grandes écoles, maîtrisant au préalable les bases du calcul différentiel et intégral.

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Charles-Edouard Pfister est Docteur en physique mathématique de l'Ecole polytechnique fédérale de Zurich (ETHZ). Il a poursuivi sa carrière de chercheur en Allemagne, au Centre pour la recherche interdisciplinaire (ZiF) de l'Université de Bielefeld, en France à Marseille, au Centre national de la recherche scientifique (CNRS) et à l'Université Rutgers aux USA. Il a ensuite rejoint la section de mathématiques de l'Ecole polytechnique fédérale de Lausanne (EPFL), où il est professeur titulaire depuis 1996. Ses domaines de recherches sont principalement la mécanique statistique ainsi que les systèmes dynamiques.

ISBN 978-2-88074-981-1

Presses polytechniques et universitaires romandes

Théorie des probabilités

Enseignement des mathématiques

Théorie des probabilités

Cours d'introduction avec application à la statistique mathématique

Charles-Edouard Pfister

Presses polytechniques et universitaires romandes

Les auteurs et l'éditeur remercient l'Ecole polytechnique fédérale de Lausanne (EPFL)
pour le soutien apporté à la publication de cet ouvrage.

Illustration de couverture: © FreshPaint, www.fotolia.com

Les Presses polytechniques et universitaires romandes sont une fondation
scientifique dont le but est principalement la diffusion des travaux
de l'Ecole polytechnique fédérale de Lausanne ainsi que d'autres
universités et écoles d'ingénieurs francophones.

PPUR, EPFL – Rolex Learning Center, Station 20, CH-1015 Lausanne
info@epflpress.org, tél.: +41 21 693 21 30

www.epflpress.org

Première édition

ISBN 978-2-88074-981-1, version imprimée
ISBN 978-2-88914-286-6, version ebook (pdf), doi.org/55430/3129TDPCEP

Avant-propos

Ce texte est issu d'un enseignement donné depuis plusieurs années à l'Ecole polytechnique fédérale de Lausanne aux étudiants de la section de physique pendant le troisième semestre. Le but est de présenter les bases et les outils principaux du calcul des probabilités de manière à faciliter par la suite la lecture et la consultation de textes plus avancés. Ce texte convient aussi à un enseignement pour des étudiants de mathématiques.

Le texte est divisé en trois parties. Dans la première partie, constituée des chapitres 1 à 8, les notions et les outils de base sont exposés. La notion d'espérance d'une variable aléatoire (v.a.) est traitée de façon approfondie. Dans la deuxième partie, constituée par les chapitres 9 à 12, les notions et les outils introduits sont utilisés pour construire des modèles importants et pour discuter en détail deux théorèmes fondamentaux. Cette partie commence par un chapitre qui donne une brève introduction aux chaînes de Markov. Le chapitre 10 expose le théorème de la loi des grands nombres et ses conséquences. Le chapitre 11 est consacré aux marches aléatoires et le chapitre 12 au théorème de la limite centrale. Dans ces deux parties, le modèle d'Ising (dans la version de Curie-Weiss) est introduit pour illustrer des concepts de la théorie. Ceci permet aussi de faire le lien avec la mécanique statistique. Enfin, dans les chapitres 13 à 16, la théorie est appliquée à quelques situations classiques de statistique mathématique, estimations ponctuelles, populations normalement distribuées, méthode des moindres carrés, estimations par intervalle et notion de test.

Tout au long du texte des exercices sont proposés afin de mieux maîtriser le contenu des chapitres. Lorsque la matière d'un chapitre est assimilée, la plupart des exercices ne présentent pas de difficultés majeures. Il est bon cependant de rappeler que pour n'importe quel sujet mathématique, les exercices les plus utiles sont les questions qu'on se pose et résout soi-même. A la fin du livre les solutions de certains exercices sont données.

Les prérequis sont peu nombreux. Les notions du calcul différentiel et intégral et de l'algèbre linéaire qui sont utilisées font l'objet de rappels qui suivent cet avant-propos. Il est utile de prendre connaissance de ces rappels car certaines conventions et notations y sont définies.

La bibliographie correspond aux sources principales utilisées, en particulier l'ouvrage classique [Fe1] pour le chapitre 11. Les références [Ch], [MiUp] et [Si] développent la théorie de points de vue assez différents. Les références [Bi], [Br], [Fe2], [Ar] et [Pe] sont de niveau plus avancé, mais devraient être abordables, au moins en partie, une fois que l'étudiant a assimilé le contenu de ce texte. Le petit traité de combinatoire [Be] est utile. Il existe de nombreux autres ouvrages d'introduction à la théorie des probabilités et à la statistique, qui sont excellents et qui sont écrits avec des points de vue différents.

Je remercie très chaleureusement Mireille pour sa relecture attentive du manuscrit. Noé Cuneo a fait toutes les simulations ainsi que tous les graphiques. Je le remercie aussi très chaleureusement pour sa disponibilité et son travail de grande qualité. Mes remerciements vont également à Ken Duffy qui m'a montré comment simplifier la preuve de l'inégalité de Hoeffding (théorème 8.1).

Table des matières

Conventions et rappels de mathématiques

Le signe □ est utilisé pour indiquer la fin d'une preuve et si nécessaire la fin d'une remarque ou d'un exemple.

I.1 Rappel sur les ensembles

La notion d'ensemble est fondamentale pour la théorie des probabilités. On utilise le terme « ensemble » seulement dans les situations suivantes :

a) Il s'agit d'une collection d'objets qu'on peut énumérer sans ambiguïté. Par exemple

$$E = \{1, 2, 3\} \quad A = \{a, b, c, d\} \quad V = \{a, e, i, o, u\}\,.$$

Les énumérations se réfèrent à un contexte qui est précisé ou qui est évident pour le lecteur.

b) Il s'agit d'une collection d'objets définie par un critère permettant de dire sans ambiguïté si un objet appartient ou non au dit ensemble. L'assertion « cet objet appartient à l'ensemble » est alors, pour un objet donné, une proposition décidable qui est vraie ou fausse.

L'expression $x \in A$ signifie que x est un élément de A. Dans ce livre les mots *famille*, *collection*, *espace* sont des synonymes du mot *ensemble*. Soit A et B deux ensembles ; la notation $A \subset B$ signifie que A est un sous-ensemble de B ; on peut avoir $A = B$. Si $A \subset \Omega$, l'*ensemble complémentaire de A par rapport à* Ω est $A^c := \Omega \backslash A$. Plus généralement, si $A \subset \Omega$ et $B \subset \Omega$,

$$A \backslash B := \{x \in A \colon x \notin B\} = A \cap B^c\,.$$

Les opérations d'union et d'intersection sont distributives,

$$A \cap (C \cup D) = (A \cap C) \cup (A \cap D)$$

et

$$A \cup (C \cap D) = (A \cup C) \cap (A \cup D)\,.$$

Soit Ω un ensemble et A_t, $t \in I$, une famille de sous-ensembles de Ω, indexée par les éléments d'un ensemble quelconque I ; l'identité suivante est appelée *formule de De Morgan* (1806-1871)

$$\Big(\bigcup_{t \in I} A_t\Big)^c = \bigcap_{t \in I} A_t^c\,.$$

En effet

$$x \in \Big(\bigcup_{t\in I} A_t\Big)^c \iff x \notin A_t \;\forall\, t \iff x \in \bigcap_{t\in I} A_t^c\,.$$

Le *produit cartésien de deux ensembles* A *et* B est l'ensemble

$$A \times B := \{(x,y)\colon x \in A \text{ et } y \in B\}\,.$$

Le produit cartésien de n copies d'un ensemble A est l'ensemble

$$A^n = A \times \cdots \times A := \{(x_1,\ldots,x_n)\colon x_i \in A\,,\ \forall\, i = 1,\ldots,n\}\,.$$

Les éléments de A^n correspondent aux applications définies sur $\{1,\ldots,n\}$ et à valeur dans A. Un élément de A^n est aussi appelé *n-uple.*

$\mathbb{N} := \{1,2,\ldots\}$ est l'ensemble des nombres naturels ;

$\mathbb{Z} := \{\ldots,-1,01,\ldots\}$ est l'ensemble des entiers relatifs ;

$\mathbb{R}$ est l'ensemble des nombres réels ;

$\mathbb{R}^k$ est le produit cartésien de k copies de $\mathbb{R}$. Les éléments de $\mathbb{R}^k$ sont notés par $(x_1,\ldots,x_k)$ ou par $\mathbf{x}$.

La *cardinalité d'un ensemble fini* E est le nombre d'éléments de cet ensemble ; elle est notée $|E|$, ou card$E$. Un ensemble E est *dénombrable* si E possède une infinité d'éléments et s'il existe une bijection

$$\varphi\colon \mathbb{N} \to E\,, \quad n \mapsto \varphi(n)$$

qui donne une *énumération* des éléments de E : $E = \{e_1, e_2, e_3, \ldots\}$ avec $e_n \equiv \varphi(n)$. L'ensemble $\mathbb{Z}$ est dénombrable, mais non l'ensemble $\mathbb{R}$.

Proposition I.1 *a) Soit A un ensemble dénombrable et $B \subset A$. Si B a une infinité d'éléments, alors B est dénombrable.*
b) Soit A_n, $n \geq 1$, une collection dénombrable d'ensembles. Si chaque A_n est fini ou dénombrable, et si l'union $\bigcup_{n\geq 1} A_n$ a une infinité d'éléments, alors $\bigcup_{n\geq 1} A_n$ est dénombrable.

I.2 Rappel d'analyse

Un intervalle est un sous-ensemble de $\mathbb{R}$ qui est borné et connexe ; il peut être ouvert, fermé, ouvert à gauche et fermé à droite ou ouvert à droite et fermé à gauche. Ces différents cas sont notés respectivement,

$$(a,b) := \{x \in \mathbb{R}\colon a < x < b\} \qquad [a,b] := \{x \in \mathbb{R}\colon a \leq x \leq b\}$$

et

$$(a,b] := \{x \in \mathbb{R}\colon a < x \leq b] \qquad [a,b) := \{x \in \mathbb{R}\colon a \leq x < b\}\,.$$

Si la suite n'est pas bornée supérieurement, on pose $\limsup_n x_n = \infty$; si pour tout $M \in \mathbb{R}$ il existe n tel que $\sup_{m \geq n} x_m < M$, alors on pose $\limsup_n x_n = -\infty$. Des conventions similaires sont faites pour $\liminf_n x_n$ qui vérifie l'identité

$$\limsup_{n\to\infty}(-x_n) = -\liminf_{n\to\infty} x_n\,.$$

Une suite est convergente, et sa limite est $a \in \mathbb{R}$, si et seulement si

$$\limsup_{n\to\infty} x_n = \liminf_{n\to\infty} x_n = a\,.$$

L'expression $f\colon (a,b) \to \mathbb{R}$ désigne une fonction f définie sur l'intervalle ouvert (a,b) à valeur réelle. Lorsque le domaine de définition de f est évident par le contexte, on désigne une fonction par $x \mapsto f(x)$ ou simplement par f. La fonction f est *continue en* $c \in (a,b)$ si et seulement si

$$\big(\forall\, x_n \in (a,b)\,,\ \lim_{n\to\infty} x_n = c\big) \implies \lim_{n\to\infty} f(x_n) = f(c)\,.$$

Lemme I.1 *Soit a_n, $n \geq 1$, une suite convergente de limite $a \in \mathbb{R}$. Alors*

$$\lim_{n\to\infty}\Big(1+\frac{a_n}{n}\Big)^n = \mathrm{e}^a\,.$$

Preuve Soit $a = \lim_n a_n$; pour tout $b \in \mathbb{R}$,

$$b = \lim_{x\to 0}\frac{\ln(1+bx)}{x} = \lim_{n\to\infty} n\ln\Big(1+\frac{b}{n}\Big) = \lim_{n\to\infty}\ln\Big(1+\frac{b}{n}\Big)^n\,.$$

Pour tout $\varepsilon > 0$, si n est suffisamment grand, $a-\varepsilon \leq a_n \leq a+\varepsilon$. Par conséquent pour tout $\varepsilon > 0$

$$a-\varepsilon \leq \liminf_{n\to\infty}\ln\Big(1+\frac{a_n}{n}\Big)^n \leq \limsup_{n\to\infty}\ln\Big(1+\frac{a_n}{n}\Big)^n \leq a+\varepsilon\,.$$

Comme la fonction exponentielle est continue, on en déduit

$$\lim_{n\to\infty}\ln\Big(1+\frac{a_n}{n}\Big)^n = a \quad\text{et}\quad \lim_{n\to\infty}\Big(1+\frac{a_n}{n}\Big)^n = \mathrm{e}^a\,.$$

□

Une fonction f est de *classe* C^1 si elle est dérivable et si sa dérivée est continue sur (a,b). Une fonction continue $f\colon (a,b) \to \mathbb{R}$ est *convexe* sur (a,b) si pour tout $0 \leq \lambda \leq 1$ et pour tout $x_1, x_2 \in (a,b)$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)\,.$$

Géométriquement, une fonction convexe est caractérisée par le fait que pour tout $c \in (a,b)$ il existe $\alpha \in \mathbb{R}$ (pas nécessairement unique) tel que

$$f(c+t) \geq f(c) + t\alpha \quad \forall\, t\,,\ c+t \in (a,b)\,.$$

Par exemple, $x \mapsto \mathrm{e}^{-x}$ est une fonction convexe sur $\mathbb{R}$ et

$$\mathrm{e}^{-x} \geq 1 - x \quad \forall\, x\,.$$

Si la dérivée seconde $f''(x) \geq 0$ pour tout $x \in (a, b)$, alors f est convexe sur (a, b). Une *fonction f est concave* si et seulement si $-f$ est convexe.

On rappelle quelques résultats importants concernant l'intégrale de Riemann (1826-1866) et la convergence des séries.

Proposition I.2 *Si f est une fonction non négative, monotone décroissante, définie sur un intervalle $[a, b]$, alors f est Riemann intégrable.*

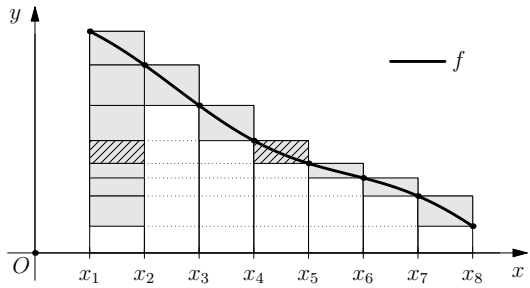


Figure 1 La différence entre la somme de Riemann supérieure et la somme de Riemann inférieure est égale à l'aire de la première colonne à gauche.

Preuve L'idée de la démonstration est due à Newton (1642-1727). On considère une subdivision de $[a, b]$ en n parties égales, $a = x_1 < \cdots < x_{n+1} = b$. Les sommes de Riemann inférieures, respectivement supérieures, de f pour cette partition vérifient

$$\underline{S}_n := \sum_{i=1}^{n} f(x_{i+1}) \frac{b-a}{n} \leq \overline{S}_n := \sum_{i=1}^{n} f(x_i) \frac{b-a}{n}\,.$$

Par définiton de $\overline{S}_n$ et $\underline{S}_n$ (voir figure 1)

$$0 \leq \overline{S}_n - \underline{S}_n \leq f(a) \frac{b-a}{n}\,.$$

A partir de ce résultat on déduit aisément l'existence de l'intégrale de Riemann de f sur $[a, b]$ lorsque $n \to \infty$. □

Proposition I.3 *Soit $a_n \geq 0$, $n \geq 1$, une suite décroissante et une fonction f décroissante définie sur $[0, \infty)$ telle que $a_n = f(n)$. Alors*

$$\int_1^{n+1} f(t)\, dt \leq \sum_{k=1}^{n} a_k \leq \int_0^n f(t)\, dt\,.$$

Des résultats analogues sont vrais dans le cas monotone croissant avec des inégalités inversées.

Preuve La somme en question est interprétée soit comme une somme de Riemann inférieure, soit comme une somme de Riemann supérieure (le pas des subdivisions est de longueur 1 ; voir figure 1). □

Proposition I.4 *Soit g_n une suite de fonctions non négatives, monotones décroissantes, définies sur l'intervalle borné $[0, N]$ et telles que*

$$g_n(t) \leq g_{n+1}(t) \ \ \forall\, t \quad et \quad \lim_n g_n(0) = g(0) < \infty\,.$$

Alors la fonction g, $g(t) := \lim_{n\to\infty} g_n(t)$, est Riemann intégrable et

$$\lim_{n\to\infty} \int_0^N g_n(t)\,dt = \int_0^N g(t)\,dt\,.$$

Preuve Sans restreindre la généralité on suppose que $N \in \mathbb{N}$. La condition $\lim_n g_n(0) = g(0) < \infty$ implique que $g(t)$ est finie pour tout $t \in [0, N]$ puisque g est non croissante. La suite g_m, $m \geq 1$, étant monotone,

$$\int_0^N g(t)\,dt \geq \lim_{m\to\infty} \int_0^N g_m(t)\,dt\,.$$

On minore l'intégrale de droite en minorant les sommes de Riemann inférieures pour $\int g_m$ par une somme de Riemann inférieure de $\int g$ de la façon suivante. L'intervalle $[0, N]$ est subdivisé en $N2^n$ sous-intervalles de longueur 2^{-n}. Pour tout $n \in \mathbb{N}$ et pour tout $\delta > 0$ il existe $m_{\delta,n}$ tel que si $m \geq m_{\delta,n}$

$$g_m(k\,2^{-n}) \geq g(k\,2^{-n}) - \delta \quad \forall\, k\,,\ 1 \leq k < N2^n\,,$$

car le nombre de points de la subdivision est fini. Pour $m \geq m_{\delta,n}$,

$$\int_0^N g_m(t)\,dt \geq \sum_{k=1}^{N\,2^n} 2^{-n} g(k\,2^{-n}) - 2^{-n}\,\delta N 2^n\,.$$

On prend successivement les limites $m \to \infty$, $\delta \to 0$ puis $n \to \infty$; on obtient

$$\lim_{m\to\infty} \int_0^N g_m(t)\,dt \geq \int_0^N g(t)\,dt\,.$$

□

Proposition I.5 *Soit $a_1, a_2, \ldots$ une suite de nombres réels et $\varphi : \mathbb{N} \to \mathbb{N}$ une bijection. On pose $b_n := a_{\varphi(n)}$. Si $a_i \geq 0$ pour tout i, alors*

$$\sum_{n\geq 1} a_n = \sum_{n\geq 1} b_n < \infty \quad ou \quad \sum_{n\geq 1} a_n \ et \ \sum_{n\geq 1} b_n \ divergent.$$

Si la série $\sum_i |a_i| < \infty$, alors il existe $a = \sum_{n\geq 1} a_n = \sum_{n\geq 1} b_n$.

Preuve Pour tout n il existe N_n tel que $\varphi(m) \leq N_n$ si $m \leq n$. Si $a_i \geq 0$,

$$\sum_{m=1}^{n} b_m \leq \sum_{k=1}^{N_n} a_k \leq \sum_{k=1}^{\infty} a_k \quad \text{et} \quad \sum_{m=1}^{\infty} b_m \geq \sum_{m=1}^{N_n} b_m \geq \sum_{k=1}^{n} a_k \,.$$

Dans le deuxième cas, pour tout $\varepsilon > 0$ il existe n_ε tel que $\sum_{j \geq n_\varepsilon} |a_j| \leq \varepsilon$. La suite des sommes partielles $\sum_{k=1}^{n} a_k$, $n \geq 1$, est une suite de Cauchy (1789-1857) qui converge vers $a = \sum_{k \geq 1} a_k$. Si $n \geq n_\varepsilon$ et $m \geq N_n$

$$\Big|\sum_{k=1}^{n} a_k - a\Big| \leq \varepsilon \quad \text{et} \quad \Big|\sum_{j=1}^{m} b_j - \sum_{k=1}^{n} a_k\Big| \leq \sum_{j \geq n_\varepsilon} |a_j| \leq \varepsilon \,.$$

□

La *série harmonique* est la série dont le terme général est $1/n$. Soit

$$H_n := \sum_{k=1}^{n} \frac{1}{k} \,.$$

Cette série diverge et (utiliser la proposition I.3)

$$\lim_{n \to \infty} \frac{H_n}{\ln n} = 1 \,.$$

On a le résultat plus fort $\lim_{n \to \infty} (H_n - \ln n) = \gamma$, où $\gamma = 0{,}5772\cdots$ est la *constante d'Euler* (1707-1783).

I.3 Rappel d'algèbre linéaire

Le *produit scalaire* sur $\mathbb{R}^k$ est noté $\langle \mathbf{x} | \mathbf{y} \rangle = \sum_{i=1}^{k} x_i y_i$. La norme d'un vecteur est $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x} | \mathbf{x} \rangle}$. Deux vecteurs $\mathbf{x}$ et $\mathbf{y}$ sont *orthogonaux* si et seulement si $\langle \mathbf{x} | \mathbf{y} \rangle = 0$.

Une matrice carrée $\mathbf{U}$ est *orthogonale* si et seulement si $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$ où $\mathbf{I}$ est la matrice identité et $\mathbf{U}^\top$ la matrice transposée de $\mathbf{U}$. Une matrice $\mathbf{U}$ est orthogonale si et seulement si ses vecteurs colonnes sont orthogonaux et de norme 1. Une matrice $\mathbf{U}$ est orthogonale si et seulement si ses vecteurs lignes sont orthogonaux et de norme 1. Si $\mathbf{U}$ est orthogonale, $|\det \mathbf{U}| = 1$.

Une matrice $\mathbf{A}$ de type $k \times k$ est *définie positive* si et seulement si

$$\langle \mathbf{x} | \mathbf{A}\mathbf{x} \rangle > 0 \quad \forall\, \mathbf{x} \neq \mathbf{0} \,.$$

Soit $\mathbf{A}$ une matrice réelle de type $k \times k$, symétrique et définie positive ; cette matrice est diagonalisable, plus précisément, il existe une matrice orthogonale $\mathbf{U}$ telle que

$$\mathbf{A} = \mathbf{U}^\top \mathbf{D}\, \mathbf{U}$$

avec $\mathbf{D}$ diagonale. Tous les éléments de $\mathbf{D}$ sont strictement positifs et coincident avec les valeurs propres de $\mathbf{A}$. Leur produit est égal au déterminant $\det\mathbf{A} > 0$; $\det\mathbf{A}^{-1} = (\det\mathbf{A})^{-1}$.

Soit $\mathbf{A}$ une matrice de type $k \times k$; les *mineurs principaux* de $\mathbf{A}$ sont les déterminants des matrices de type $m \times m$, $m = 1, \ldots, k$, dont les éléments sont $\mathbf{A}_{ij}$ $1 \leq i, j \leq m$. La proposition suivante est due à Sylvester (1814-1897) (voir par exemple *Matrix Analysis*, R. A. Horn, C. R. Johnson, Cambridge University Press (2006), théorème 7.2.5.)

Proposition I.6 *Une matrice réelle symétrique est définie positive si et seulement si ses mineurs principaux sont positifs.*

Chapitre 1

Introduction

La théorie des probabilités a des applications dans pratiquement tous les domaines, sciences sociales, assurances, finance, linguistique, génétique, télécommunications, physique, etc. Cette théorie traite de phénomènes aléatoires, terme qui apparaît en droit à la fin du XVI^{e} siècle pour qualifier un contrat qui prévoit des conditions liées à la chance. Un tel phénomène peut se manifester de plusieurs manières possibles, sans qu'on puisse prévoir de quelle manière il se manifestera. Deux mots-clés caractérisent les phénomènes aléatoires : imprédictibilité et incertitude. Les raisons à l'origine de l'imprédictibilité ou de l'incertitude sont très diverses. Dans certaines situations le mécanisme qui en est la cause est connu, comme pour le tirage des numéros gagnants d'une loterie ou dans un algorithme aléatoire où le caractère aléatoire est introduit artificiellement. Dans d'autres situations, comme dans le cas de transactions boursières, c'est l'ignorance de certains paramètres qui en est la cause au moins partiellement. Dans le cas du chaos déterministe, la sensibilité aux conditions initiales et l'impossibilité de contrôler celles-ci font que des techniques probabilistes sont adéquates pour étudier ces systèmes dynamiques. En raison des multiples applications de la théorie des probabilités à des domaines si variés, il n'est pas étonnant que le contenu sémantique du terme *probabilité* ait reçu des interprétations différentes. L'aspect sémantique de ce terme joue un rôle en statistique dont le but est de récolter et d'analyser des données empiriques. Le sujet de ce livre est l'exposition des bases mathématiques du calcul des probabilités. Le terme *probabilité* a une définition mathématique précise et ce calcul est le même quelle que soit son interprétation dans un domaine d'application particulier.

La théorie des jeux de hasard et les problèmes de tirage d'urne ont eu une place prépondérante au début de la théorie des probabilités que l'on situe généralement au XVII^{e} siècle avec la correspondance entre Pierre de Fermat (1608-1665) et Blaise Pascal (1623-1662). Parmi les oeuvres majeures on peut citer celle de Jacob Bernoulli (1654-1705), *Ars Conjectandi* (1713), et celle de Abraham De Moivre (1667-1754), *The Doctrine of Chances : or, A Method of Calculating the Probabilities of Events in Play* (1718). Un ouvrage important qui clôt cette première phase de développement est le traité de Pierre-Simon de Laplace (1749-1827), *Théorie analytique des probabilités*. Au début du livre II, « Théorie générale des probabilités », il énonce un principe de déterminisme absolu et en conséquence en déduit que le mot *hasard* n'est au fond que l'expression de

notre ignorance. La probabilité est relative, en partie, à cette ignorance, et en partie, à nos connaissances. Laplace donne ensuite la définition suivante [1] :

« La théorie des probabilités consiste à réduire tous les événements qui peuvent avoir lieu dans une circonstance donnée, à un certain nombre de cas également possibles, c'est-à-dire tels que nous soyons également indécis sur leur existence, et à déterminer parmi ces cas, le nombre de ceux favorables à l'événement dont on cherche la probabilité. Le rapport de ce nombre à celui de tous les cas possibles, est la mesure de cette probabilité qui n'est donc qu'une fraction dont le dénominateur est celui de tous les cas possibles. »

A la suite de cette définition Laplace considère l'exemple suivant. On a trois urnes A, B et C dont une ne renferme que des boules noires, tandis que les autres ne contiennent que des boules blanches. On tire une boule de l'urne C. Quelle est la probabilité que cette boule soit noire ? Si l'on ignore laquelle de ces urnes contient les boules noires, il n'y a pas de motifs de croire que c'est plutôt l'urne C que l'urne B ou A. La probabilité d'extraire une boule noire est donc 1/3. Si par contre on sait que l'urne A ne contient que des boules blanches, l'indécision ne porte que sur les urnes B et C. Sachant que l'urne A ne contient que des boules blanches, la probabilité de tirer une boule noire est 1/2.

La monographie de A. N. Kolmogorov (1903-1987), *Grundbegriffe der Wahrscheinlichkeitsrechnung* parue en 1933, marque un tournant décisif dans la théorie des probabilités. Dans cette monographie Kolmogorov aborde la théorie des probabilités d'un point de vue nouveau. Il définit mathématiquement un *espace de probabilité*, sans faire référence au champ sémantique du mot probabilité, en donnant une définition axiomatique des notions d'*événement* et de *probabilité d'un événement*. La théorie des espaces de probabilité ainsi formulée est une branche des mathématiques au même titre que la théorie des espaces vectoriels ou la théorie des groupes. Elle est le cadre mathématique de la théorie moderne des probabilités. Néanmoins, la définition donnée par Laplace reste valable dans le cadre où elle a été formalisée. Ce cadre est principalement restreint à des problèmes de nature combinatoire où l'incertitude porte sur un nombre fini de cas *et* lorsque chaque cas est également possible. Cette définition a le mérite de mettre en évidence de façon simple et naturelle les propriétés mathématiques élémentaires du terme *probabilité*.

La théorie mathématique des probabilités s'est toujours développée en relation étroite avec ses applications, ce qui confère à cette branche des mathématiques un caractère spécial. Dans toute démarche scientifique pour étudier un phénomène on est conduit à faire des idéalisations, des simplifications et des hypothèses, afin de construire un modèle adéquat. Dans le cas du lancer d'une pièce de monnaie on n'hésite pas à attribuer la probabilité 1/2 à chacun des deux résultats possibles, Pile et Face. Même si des études statistiques poussées dans des cas réels ont toujours montré que les deux résultats ne sont pas également probables, ce modèle de la pièce idéale reste un modèle important, non

1. *Théorie Analytique des Probabilités* ; par M. le Comte Laplace, M^{me} V^{e} Courcier (Paris 1812), p. 178.

seulement à cause de sa simplicité logique, mais parce qu'il permet d'analyser un grand nombre de situations concrètes avec précision et exactitude. De plus, les résultats mathématiques sur ce modèle permettent de tester dans quelle mesure l'hypothèse « chacun des deux résultats est également probable » est vérifiée. Un des rôles de la théorie mathématique est précisément de proposer des modèles comme le « dé équilibré à six faces », le « générateur de nombre aléatoire », « la marche aléatoire sur un graphe », le « mouvement Brownien », la « chaîne de Markov » etc. Ces modèles se sont révélés très efficaces dans de très nombreuses situations. Un autre but de la théorie est d'introduire des concepts comme ceux « d'indépendance », « d'espérance » et d'énoncer des théorèmes, comme celui de la « loi des grands nombres » ou le « théorème de la limite centrale ».

Il y a un point important qu'on ne souligne souvent pas assez : une fois un modèle posé pour l'étude d'un phénomène aléatoire, l'analyse mathématique porte sur ce modèle et *les conclusions concernent ce modèle.* Il faut toujours avoir cela à l'esprit lorsqu'on applique la théorie des probabilités à des situations concrètes. La vérification de l'adéquation des modèles utilisés dans un domaine spécifique se base sur des méthodes qui relèvent principalement de la statistique ou du domaine lui-même. La justification d'un modèle pour décrire un phénomène aléatoire concret dépend en grande partie de son succès à rendre compte de ce phénomène.

Chapitre 2

Axiomes de Kolmogorov

L'analyse d'un phénomène aléatoire est faite en utilisant deux notions. La notion d'événement, qui est une propriété décidable du phénomène, et la notion de probabilité d'un événement, qui est une pondération sur l'ensemble des événements. Ceci est formalisé mathématiquement par la notion d'espace de probabilité, définition 2.1.

Dans son ouvrage *Théorie analytique des Probabilités* Laplace définit la probabilité d'un événement par le rapport

$$\frac{\text{nombre de cas favorables à la réalisation de l'événement}}{\text{nombre total des résultats possibles}}. \tag{2.1}$$

Pour appliquer cette définition il faut d'une part donner la liste de tous les résultats possibles du phénomène aléatoire qu'on étudie, et d'autre part donner la définition d'un événement. Un évènement est une affirmation sur le phénomène, telle qu'on peut décider si elle est vraie sur la base de l'observation du phénomène si et seulement si on connaît le résultat de l'observation. Lorsque l'on a observé le phénomène, on peut donc dire pour tout événement s'il est vrai ou non. Les cas favorables sont les cas pour lesquels l'événement est vrai. La probabilité d'un événement est un nombre compris entre 0 et 1 qui exprime un degré de vraisemblance de l'événement ; si cette probabilité est très proche de 1, mais non égale à 1, il est très vraisemblable, mais pas certain, que la propriété décrite par l'événement soit vraie lors de l'observation du phénomène.

2.1 Expérience aléatoire

Pour formaliser la théorie il est commode d'introduire la notion d'expérience aléatoire qui peut être réelle ou virtuelle (expérience de pensée). Ces expériences ont en commun avec celles de physique le point 1) ci-dessous. Une *expérience aléatoire* a les caractéristiques suivantes.

1) Il y a un protocole qui fixe entre autres les hypothèses qu'on fait, les buts et la manière de faire l'expérience, de sorte que les résultats possibles de celle-ci sont identifiables et peuvent être formulés sans ambiguïté. L'ensemble des résultats possibles est appelé *espace fondamental* et noté souvent par Ω.

2) Les résultats d'une expérience aléatoires sont imprédictibles, mais une fois l'expérience faite le résultat est connu sans ambiguïté. On peut répéter une expérience aléatoire. Si l'on répète l'expérience, le résultat observé est en général différent.

Lors de la modélisation d'un phénomène aléatoire le choix de l'espace fondamental Ω est une étape essentielle. Il faut par exemple décider si l'on veut faire l'expérience en fixant certains paramètres on non. On a une certaine liberté dans le choix de Ω, mais les événements qu'on peut définir pour une expérience aléatoire dépendent de ce choix. Le point principal à retenir est que chaque résultat est codé sans ambiguïté par un et un seul $\omega \in \Omega$. En allemand l'ensemble Ω est appelé *Ergebnissraum* (espace des résultats).

Exemple 2.1 Marche aléatoire à une dimension. Toutes les τ secondes un marcheur fait un pas sur la droite $\mathbb{R}$ de la manière suivante :
– la position du marcheur au temps $t = 0$ est $\ell_0 := 0$;
– au temps $t = \tau$ on lance une pièce de monnaie. Si c'est Pile le marcheur fait un pas de longueur h à droite et sa nouvelle position est $\ell_1 = \ell_0 + h$. Si c'est Face il fait un pas de longueur h à gauche et sa nouvelle position est $\ell_1 = \ell_0 - h$;
– on recommence cette opération n fois.

L'espace fondamental est l'ensemble de toutes les marches possibles compatibles avec le protocole de l'expérience. Il y a 2^n marches différentes. La position du marcheur au temps $t = n\tau$ est donnée par ℓ_n. □

Exemple 2.2 Mesure d'une quantité scalaire m_*. On fait n fois la mesure de cette quantité sous les mêmes conditions ; les résultats sont donnés par n nombres réels $x_1, \ldots, x_n$ qui sont en général différents. Ces différences sont attribuées entre autres à des causes aléatoires qui résultent du fait que l'on ne peut pas reproduire exactement des conditions identiques lors de la répétition d'une mesure ; on a donc une expérience aléatoire et l'on écrit $x_i := m_* + z_i$, où z_i représente l'incertitude due à ces causes aléatoires lors de la $i^{\text{ième}}$ mesure. L'espace fondamental est $\Omega = \mathbb{R}^n$.

Exemple 2.3 On lance successivement dix fois une pièce de monnaie. Un résultat possible est *Pile Face Face Face Pile Pile Face Face Pile Face*. On pose

$$\omega_i := \begin{cases} 1 & \text{si le } i^{\text{ième}} \text{ lancer donne Pile,} \\ 0 & \text{si le } i^{\text{ième}} \text{ lancer donne Face.} \end{cases}$$

On code le résultat de l'expérience par $\omega = (\omega_1, \ldots, \omega_{10})$; pour le résultat ci-dessus le code est $(1, 0, 0, 0, 1, 1, 0, 0, 1, 0)$ et l'espace fondamental est

$$\Omega := \{\omega = (\omega_1, \ldots, \omega_{10}) \colon \forall i,\ \omega_i \in \{0, 1\}\} = \{0, 1\}^n .$$

2.2 Espace de probabilité

Un *événement* $\mathcal{E}$ est une *propriété décidable* d'une expérience aléatoire. Cela signifie qu'un événement possède la caractéristique suivante : lorsque $\omega \in \Omega$ est

connu, un et un seul des énoncés ci-dessous est correct.

a) $\mathcal{E}$ est vrai pour ω ($\mathcal{E}$ *est réalisé*, ω est une *réalisation de* $\mathcal{E}$)

b) $\mathcal{E}$ est faux pour ω ($\mathcal{E}$ *n'est pas réalisé*).

Lorsque le résultat de l'expérience est connu il n'y a aucune ambiguïté sur la réalisation ou non de n'importe quel événement $\mathcal{E}$. Une conséquence immédiate de ce fait est qu'on peut identifier un événement $\mathcal{E}$ avec le sous-ensemble $E \subset \Omega$ défini par

$$E := \{\omega \in \Omega \colon \mathcal{E} \text{ réalisé pour } \omega\}.$$

Désormais on fait cette identification ; un événement est un sous-ensemble de Ω. Dans l'exemple 2.1, l'affirmation « la position du marcheur au temps $t = n\tau$ est $\ell_n = 0$ » est un événement important qui correspond au retour à l'origine du marcheur en $t = n\tau$. Il suffit de connaître ℓ_n pour savoir s'il est réalisé ou non. Dans l'exemple 2.3, l'affirmation « le résultat de l'expérience contient au plus cinq Piles » est un événement qui est identifié au sous-ensemble

$$E := \left\{\omega \in \Omega \colon \sum_{j=1}^{10} \omega_j \leq 5\right\}.$$

L'événement $\{\omega\}$ est réalisé si et seulement si le résultat de l'expérience aléatoire est ω. Cet événement est appelé *événement élémentaire*. On ne peut pas le décomposer en d'autres événements.

On exige que la collection $\mathcal{F}$ des événements relatifs à une expérience aléatoire possède la structure d'une *algèbre de Boole* (1815-1864), ce qui signifie

1) $\Omega \in \mathcal{F}$, $\emptyset \in \mathcal{F}$;
2) si $A \in \mathcal{F} \Rightarrow A^c := \Omega \backslash A \in \mathcal{F}$;
3) si $A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$ et $A \cap B \in \mathcal{F}$.

Un exemple d'algèbre de Boole est la collection de tous les sous-ensembles de Ω. Cette algèbre est notée $\mathcal{P}(\Omega)$. Si Ω est fini, on choisit pour $\mathcal{F}$ en général cette algèbre de Boole.

Le fait que $\mathcal{F}$ est une algèbre de Boole signifie que les opérations élémentaires de la logique, la négation, la conjonction et la disjonction, sont définies sur la famille des événements :

1) si $A \in \mathcal{F}$, la *négation* de A est l'événement $A^c = \Omega \backslash A \in \mathcal{F}$, qui est vrai si et seulement si A est faux ;
2) si $A, B \in \mathcal{F}$, la *conjonction* de A et B est l'événement $A \cap B \in \mathcal{F}$, qui est vrai si et seulement si A et B sont vrais ;
3) si A,$B \in \mathcal{F}$, la *disjonction* de A et B est l'événement $A \cup B \in \mathcal{F}$, qui est vrai si et seulement si l'un des deux événements A, B est vrai. C'est le « ou » non exclusif, qu'il faut distinguer du *ou exclusif* (XOR eXclusive OR) de A et B, qui est l'événement $A \oplus B := (A \cup B) \backslash (A \cap B) \in \mathcal{F}$. $A \oplus B$ est vrai si et seulement si un et un seul des événements A, B est vrai. L'ensemble $A \oplus B$ est appelé *somme booléenne de A et B*. On utilise également la notation $A \Delta B$ à la place de $A \oplus B$, ainsi que la terminologie *différence symétrique de A et B* pour l'ensemble $A \Delta B$.

Ω est interprété comme l'événement *certain* puisque $\omega \in \Omega$ pour tout ω ; il est toujours réalisé. L'événement $\emptyset$ est interprété comme l'événement *impossible* puisque $\omega \notin \emptyset$ pour tout ω. Deux événements A et B sont *disjoints* ou *incompatibles* ou *mutuellement exclusifs* si et seulement si $A \cap B = \emptyset$. On utilise aussi la notation $A \cdot B$, ou même AB, à la place de $A \cap B$. Par récurrence on montre facilement

$$A_1, A_2, \ldots, A_p \in \mathcal{F} \implies \bigcup_{i=1}^{p} A_i \in \mathcal{F} \quad \text{et} \quad \bigcap_{i=1}^{p} A_i \in \mathcal{F}\,.$$

Si la collection $\mathcal{F}$ est infinie, on exige encore la propriété supplémentaire suivante pour les *familles dénombrables* d'événements :

$$A_1, A_2, \ldots \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F} \quad \text{et} \quad \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}\,. \tag{2.2}$$

Une algèbre de Boole qui possède la propriété (2.2) est une *σ-algèbre de Boole*. Cette propriété est primordiale pour la théorie des probabilités, comme on peut le voir ci-dessous dans la formulation de la proposition 2.1 et la remarque 2.1.

Pour compléter la description mathématique d'une expérience aléatoire on définit une pondération sur la collection $\mathcal{F}$ des événements. Cette pondération est une application $P\colon \mathcal{F} \to [0,1]$; la *probabilité de l'événement E est $P(E)$*. Une propriété fondamentale de la définition de Laplace est celle d'additivité de cette pondération :

$$\text{si } A \cap B = \emptyset, \text{ alors } P(A \cup B) = P(A) + P(B).$$

La propriété d'additivité s'étend immédiatement aux familles finies d'événements disjoints. Lorsque la collection $\mathcal{F}$ contient une infinité d'événements on ajoute une condition supplémentaire, la *σ-additivité*. P est *σ-additive* si et seulement si P est additive, et si $A_1, A_2, \ldots \in \mathcal{F}$ est une famille dénombrable d'événements disjoints deux à deux ($A_i \cap A_j = \emptyset$ pour tout $i \neq j$), alors

$$P\Big(\bigcup_{i \geq 1} A_i\Big) = \sum_{i \geq 1} P(A_i)\,.$$

Toute application $P\colon \mathcal{F} \to [0,1]$, qui est σ-additive et telle que $P(\Omega) = 1$, est *une mesure de probabilité sur $\mathcal{F}$*. La propriété de σ-additivité a des conséquences essentielles (voir par exemple remarques 2.1 et 7.1).

Définition 2.1 (Axiomes de Kolmogorov) *Un* espace de probabilité *est un triplet $(\Omega, \mathcal{F}, P)$ vérifiant les propriétés suivantes.*

P1. Ω est un ensemble qui est l'espace fondamental.

P2. $\mathcal{F}$ est une collection de sous-ensembles de Ω qui est la collection des événements ; $\mathcal{F}$ possède la structure d'une σ-algèbre de Boole.

P3. $P : \mathcal{F} \to [0,1]$ est une mesure de probabilité sur $\mathcal{F}$.

En résumé, une expérience aléatoire est décrite mathématiquement par un espace de probabilité.

On déduit quelques conséquences élémentaires des axiomes de Kolmogorov. On a toujours $P(\emptyset) = 0$, mais il est possible d'avoir $P(E) = 0$ et $E \neq \emptyset$.

Lemme 2.1

1) $P(A^c) = 1 - P(A)$ *;* $P(\emptyset) = 0$.
2) Si $A \subset B \Rightarrow P(A) \leq P(B)$.
3) $P(A) + P(B) = P(A \cup B) + P(A \cap B)$.

Preuve

$$P(\Omega) = P(A) + P(A^c) \implies P(A^c) = 1 - P(A)\,.$$

Si $A \subset B$,

$$P(B) = P(A) + P(B \backslash A) \geq P(A)\,,$$

car $P(B\backslash A) \geq 0$. Finalement, on écrit

$$A = (A \cap B) \cup (A \cap B^c) \quad \text{et} \quad B = (B \cap A) \cup (B \cap A^c)\,.$$

Par conséquent

$$\begin{aligned} P(A) + P(B) &= P(A \cap B) + \left[P(A \cap B^c) + P(B \cap A) + P(B \cap A^c)\right] \\ &= P(A \cap B) + P(A \cup B)\,. \end{aligned}$$

□

Une mesure de probabilité a des propriétés de continuité monotone séquentielle qui sont très utiles pour le calcul de $P(A)$. Une famille dénombrable A_n, $n \geq 1$, est une *suite monotone décroissante vers* A, ce qui est noté $A_n \downarrow A$, si et seulement si

$$A_n \supset A_{n+1} \text{ et } A = \bigcap_n A_n\,.$$

Une famille dénombrable A_n, $n \geq 1$, est une *suite monotone croissante vers* A, ce qui est noté $A_n \uparrow A$, si et seulement si

$$A_n \subset A_{n+1} \text{ et } A = \bigcup_n A_n\,.$$

Noter que dans ces définitions le sous-ensemble A est un événement parce que la famille des événements est une σ-algèbre.

Proposition 2.1

1) Si $A_n \downarrow A$, *alors* $\lim_n P(A_n) = P(A)$.
2) Si $A_n \uparrow A$, *alors* $\lim_n P(A_n) = P(A)$.

3) Pour toute famille finie ou dénombrable d'ensembles B_n,

$$P\Big(\bigcup_n B_n\Big) \leq \sum_n P(B_n)\,.$$

Preuve 1) Soit $A_n \downarrow A$; A_n est écrit comme une réunion d'ensembles disjoints,

$$A_n = A \cup \bigcup_{m\geq n} (A_m \backslash A_{m+1})\,.$$

En effet, les ensembles $A_i\backslash A_{i+1}$ et $A_j\backslash A_{j+1}$ sont disjoints si $i \neq j$, et chacun est disjoint de A. Si $\omega \in A_n$, ou bien $\omega \in A$ ou bien il existe un plus petit entier $m > n$ tel que $\omega \notin A_m$, et dans ce cas $\omega \in A_{m-1}\backslash A_m$. Ceci prouve l'identité. La σ-additivité permet d'écrire

$$P(A_n) = P(A) + \sum_{m\geq n} P(A_m \backslash A_{m+1})\,.$$

Comme

$$P(A_1) = P(A) + \sum_{m\geq 1} P(A_m \backslash A_{m+1}) \leq 1\,,$$

la série $\displaystyle\sum_{m\geq 1} P(A_m\backslash A_{m+1})$ est convergente et donc

$$\lim_{n\to\infty} \sum_{m\geq n} P(A_m \backslash A_{m+1}) = 0\,.$$

2) On utilise la formule de De Morgan

$$\Big(\bigcup_{t\in I} A_t\Big)^c = \bigcap_{t\in I} A_t^c\,.$$

Par conséquent $A_n \uparrow A$ si et seulement si $A_n^c \downarrow A^c$ et

$$\lim_n P(A_n^c) = 1 - \lim_n P(A_n) = P(A^c) = 1 - P(A)\,.$$

3) Comme l'union des B_n est égale à l'union disjointe des $B_k \backslash \bigcup_{j=1}^{k-1} B_j$, $k \geq 1$,

$$P(\bigcup_{n\geq 1} B_n) = \sum_{k\geq 1} P\Big(B_k \backslash \bigcup_{j=1}^{k-1} B_j\Big) \leq \sum_{n\geq 1} P(B_n)\,.$$

□

Remarque 2.1 Lorsque $A_n \downarrow \emptyset$ la propriété $\lim_n P(A_n) = 0$ est intuitivement naturelle. Si $\mathcal{F}$ est une σ-algèbre et si $P\colon \mathcal{F} \to [0,1]$ est une application additive, alors cette propriété ($A_n \downarrow \emptyset \implies \lim_n P(A_n) = 0$) implique la σ-additivité de P. Par conséquent, pour une σ-algèbre la propriété de σ-additivité de P et les propriétés de continuité monotone de P de la proposition 2.1 sont équivalentes.

En effet, de la propriété ($A_n \downarrow \emptyset \implies \lim_n P(A_n) = 0$) on obtient

$$A_n \uparrow A \implies \lim_{n\to\infty} P(A_n) = P(A)$$

car

$$A_n \uparrow A \iff A = (A\backslash A_n) \cup A_n \quad \text{et} \quad A\backslash A_n \downarrow \emptyset\,.$$

Si $A_1, A_2, \ldots$ est une famille dénombrable d'événements deux à deux disjoints,

$$B_m = \bigcup_{i=1}^{m} A_i \uparrow B = \bigcup_{i\geq 1} A_i\,,$$

et donc

$$P(B) = \lim_{m\to\infty} P(B\backslash B_m) + \lim_{m\to\infty} \sum_{i=1}^{m} P(A_i) = \sum_{m\geq 1} P(A_m)\,.$$

Ceci établit la σ-additivité de la mesure de probabilité P. □

De l'identité de base

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

on déduit aisément l'identité

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) = {} & P(A_1) + P(A_2) + P(A_3) \\ & - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\ & + P(A_1 \cap A_2 \cap A_3) \end{aligned}$$

en écrivant $P(A_1 \cup A_2 \cup A_3) = P(A_1 \cup (A_2 \cup A_3))$. Par récurrence, pour k événements, on obtient des identités analogues qui sont appelées *formules d'inclusion-exclusion.*

Proposition 2.2 *Soit $A_1, \ldots, A_n$ n événements. Alors* (formules d'inclusion-exclusion)

$$P\Big(\bigcup_i A_i\Big) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{J\subset\{1,\ldots,n\}:\ |J|=k} P\Big(\bigcap_{j\in J} A_j\Big)\,.$$

$$P\Big(\bigcap_j A_j\Big) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{J\subset\{1,\ldots,n\}:\ |J|=k} P\Big(\bigcup_{j\in J} A_j\Big)\,.$$

Une démonstration qui utilise la notion d'espérance est donnée à la fin du chapitre 7.

2.3 Espace de probabilité discret

Les espaces de probabilité les plus simples sont ceux où Ω est un ensemble fini ou dénombrable. On peut écrire $\Omega = \{\omega_1, \ldots, \omega_n\}$ si $|\Omega| = n$ et $\Omega = \{\omega_1, \omega_2, \ldots\}$ s'il est dénombrable. $(\Omega, \mathcal{F}, P)$ est un *espace de probabilité discret* si :

$$\Omega \text{ est fini ou dénombrable et } \mathcal{F} := \mathcal{P}(\Omega).$$

Une description complète de tous les espaces de probabilité discrets est donnée ci-dessous. Il suffit de spécifier $P(\{\omega\})$ pour tout $\omega \in \Omega$.

Dans un espace de probabilité les événements élémentaires distincts sont toujours disjoints. Pour simplifier les notations on écrit $P(\omega)$ à la place de $P(\{\omega\})$. Si l'espace de probabilité est discret, la σ-additivité de la mesure de probabilité P permet d'écrire

$$P(E) = \sum_{i\colon \omega_i \in E} P(\omega_i) \equiv \sum_{\omega \in E} P(\omega) \quad \forall\, E \subset \Omega\,.$$

La somme ne dépend pas du choix de l'énumération des éléments de l'espace fondamental (voir proposition I.5). Inversement, si $q\colon \Omega \to \mathbb{R}$ est n'importe quelle application telle que $q(\omega) \geq 0$ et $0 < \sum_{\omega\in\Omega} q(\omega) < \infty$, alors

$$E \mapsto P(E) := \frac{\sum_{\omega\in E} q(\omega)}{Z} \quad \text{avec} \quad Z := \sum_{\omega\in\Omega} q(\omega)$$

définit une mesure de probabilité sur $\mathcal{P}(\Omega)$. En posant $q'(\omega) = q(\omega)/Z$, il suffit de considérer le cas $Z = 1$. L'additivité est facile à vérifier et P est une mesure de probabilité si Ω est fini. Lorsque Ω est dénombrable il faut encore montrer la σ-additivité.

Lemme 2.2 *Sous les hypothèses ci-dessus, si Ω est dénombrable, l'application P est σ-additive.*

Preuve Soit $\Omega = \{\omega_1, \omega_2, \ldots\}$. Par hypothèse, pour tout $\varepsilon > 0$ il existe N_ε tel que

$$\sum_{n > N_\varepsilon} q'(\omega_n) \leq \varepsilon\,.$$

Soit $A_1, A_2, \ldots$ une famille dénombrable d'événements deux à deux disjoints. Il y a au plus un nombre fini de A_i tels que

$$A_i \cap \{\omega_1, \ldots, \omega_{N_\varepsilon}\} \neq \emptyset\,;$$

par conséquent il existe m_ε tel que $m \geq m_\varepsilon$ implique

$$A_m \cap \{\omega_1, \ldots, \omega_{N_\varepsilon}\} = \emptyset\,.$$

Pour tout $m \geq m_\varepsilon$ on pose

$$B_m := \bigcup_{i=1}^{m} A_i \quad \text{et} \quad C_m := \bigcup_{i>m} A_i\,.$$

Comme $C_m \cap \{\omega_1, \ldots, \omega_{N_\varepsilon}\} = \emptyset$,

$$P(C_m) = \sum_{\omega \in C_m} q'(\omega) \leq \sum_{i > N_\varepsilon} q'(\omega_i) \leq \varepsilon\,.$$

Si $A = \bigcup_{n \geq 1} A_n$, alors

$$P(A) = P(B_m \cup C_m) = P(B_m) + P(C_m) = \sum_{i=1}^{m} P(A_i) + P(C_m)\,;$$

on en déduit

$$0 \leq P(A) - \sum_{i=1}^{m} P(A_i) \leq \varepsilon\,.$$

On obtient la σ-additivité puisque ε est arbitraire. □

Exemple 2.4 Soit Ω l'ensemble des états possibles d'un système physique. On suppose que Ω est dénombrable et que $H(\omega)$ est l'énergie du système dans l'état ω. Soit $\beta := (k_B T)^{-1}$ où k_B est la constante de Boltzmann (1844-1906) et T la température absolue du système. La fonction

$$\beta \mapsto Z(\beta) := \sum_{\omega \in \Omega} \mathrm{e}^{-\beta H(\omega)}$$

est appelée *fonction de partition* du système. Si $Z(\beta) < \infty$, on définit en mécanique statistique la *mesure de probabilité de Gibbs* (1839-1903) par la formule

$$P(E) := \sum_{\omega \in E} \frac{\mathrm{e}^{-\beta H(\omega)}}{Z(\beta)}\,.$$

Cette mesure de probabilité décrit les propriétés du système à l'équilibre et à la température absolue T.

Un modèle important, en mécanique statistique, mais également en théorie moderne des probabilités, est celui du *modèle d'Ising* (1900-1998) qui a été proposé initialement comme un modèle de ferro-aimant.

On donne ici la version en champ moyen appelée aussi modèle de Curie-Weiss (Curie (1859-1906) et Weiss (1865-1940)). Le modèle consiste en n systèmes élémentaires, appelés « spins », qui peuvent se trouver dans deux états notés ± 1. L'espace des configurations est isomorphe à l'espace fondamental utilisé pour n lancers d'une pièce de monnaie,

$$\Omega_n = \{\omega = (\omega_1, \ldots, \omega_n) \colon \omega_i = \pm 1, \forall\, i\}\,.$$

L'énergie d'une configuration est donnée par

$$H_n(\omega) := -\frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_i \omega_j - h \sum_{i=1}^{n} \omega_i\,. \tag{2.3}$$

Le paramètre h est réel et en physique il représente un champ magnétique externe. Il est commode de définir des applications X_i, $i = 1, \ldots, n$, sur l'espace des configurations Ω par

$$X_i(\omega) := \omega_i \quad (X_i \text{ donne l'état du système indexé par } i). \tag{2.4}$$

On étudie entre autres l'aimantation par spin,

$$M_n(\omega) := \frac{1}{n} \sum_{i=1}^{n} X_i(\omega) = \frac{1}{n} \sum_{i=1}^{n} \omega_i \,, \tag{2.5}$$

pour les différentes valeurs de $\beta \geq 0$ et $h \in \mathbb{R}$, ainsi que de n. La limite, lorsque n tend vers l'infini, correspond en physique à la *limite thermodynamique*. On est intéressé au comportement asymptotique de M_n dans cette limite.

A partir de la fonction de partition on obtient, dans la limite thermodynamique, l'énergie libre du système par spin

$$f(h, \beta) := - \lim_{n \to \infty} \frac{1}{\beta n} \ln Z_n \,, \quad Z_n := \sum_{\omega \in \Omega_n} \mathrm{e}^{-\beta H_n(\omega)} \,. \tag{2.6}$$

Suite à l'exemple 6.4 section 6.1. □

Exemple 2.5 On considère des polynômes P et Q de même degré n de la variable réelle x et on aimerait savoir si $P = Q$. Dans ce but on utilise l'algorithme aléatoire suivant.

1) L'algorithme choisit de manière équiprobable un nombre entier positif k inférieur ou égal à $100n$.

2) L'algorithme calcule $P(k)$ et $Q(k)$; il indique que $P \neq Q$ si $P(k) \neq Q(k)$ et que $P = Q$ si $P(k) = Q(k)$.

Quelle est la probabilité que l'algorithme donne une fausse réponse ? Pour décrire la première opération de l'algorithme, on introduit l'espace de probabilité discret $\Omega := \{1, 2, \ldots, 100n\}$ muni de la mesure de probabilité uniforme, i.e. telle que $P(j) = (100n)^{-1}$. L'algorithme donne une réponse fausse si et seulement si $P \neq Q$ et $P(k) = Q(k)$. Cela signifie que k est une racine de $P(x) - Q(x) = 0$. Comme le degré de $P - Q$ est au plus n, il y a au plus n valeurs de $j \in \Omega$ qui peuvent conduire à ce résultat. Par conséquent la probabilité d'obtenir une réponse fausse est inférieure ou égale à $1/100$. □

2.4 Exercices

Exercice 2.1 Soit A_n, $n \geq 1$, une collection dénombrable d'événements tels que $P(A_n) = 1$ pour tout $n \geq 1$. Montrer que

$$P\Big(\bigcap_{n \geq 1} A_n \Big) = 1 \,.$$

Exercice 2.2 Soit A_n, $n \geq 1$, une collection d'événements. Vérifier que l'événement

$$E := \{\text{une infinité des événements } A_n \text{ sont réalisés}\}$$

s'écrit

$$E = \bigcap_{n\geq 1} \bigcup_{m\geq n} A_m \,.$$

Exercice 2.3 Vérifier qu'une algèbre de Boole $\mathcal{A}$ est une σ-algèbre si et seulement si pour chaque suite monotone croissante ou décroissante d'événements $A_n \in \mathcal{A}$, $n \geq 1$, $\bigcup_n A_n \in \mathcal{A}$ ou $\bigcap_n A_n \in \mathcal{A}$.

Exercice 2.4 Soit $\Omega := \{1,2,3,4,5,6\}$; on considère $(\Omega, \mathcal{F}, P)$ avec $\mathcal{F} = \mathcal{P}(\Omega)$ et P une mesure de probabilité telle que

$$P(\{1,2,3\}) = 0{,}60 \quad P(\{4\}) = 0{,}15 \quad P(\{5,6\}) = 0{,}25\,.$$

a) Quels sont les événements dont on peut calculer les probabilités avec l'information ci-dessus.
b) Définir explicitement toutes les mesures de probabilité P' qui coincident avec P sur les événements $\{1,2,3\}$, $\{4\}$ et $\{5,6\}$, i.e. telles que

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

Indication : la famille des mesures de probabilité P' dépend de trois paramètres.

Exercice 2.5 Soit $B \subset \Omega$ et A_n, $n \geq 1$, une collection de sous-ensembles de Ω. Vérifier les identités

$$B\backslash(\bigcup_{n\,:\,n\geq 1} A_n) = \bigcap_{n\,:\,n\geq 1} (B\backslash A_n) \quad , \quad B\backslash(\bigcap_{n\,:\,n\geq 1} A_n) = \bigcup_{n\,:\,n\geq 1} (B\backslash A_n)\,.$$

Exercice 2.6 Sur $\mathcal{P}(\Omega)$ on définit les opérations suivantes. La somme booléenne

$$(A,B) \mapsto A \oplus B := (A \cup B)\backslash(A \cap B)\,.$$

Le produit

$$(A,B) \mapsto A \cdot B := A \cap B\,.$$

Vérifier

$$(A \oplus B)\cdot C = A \cdot C \oplus B \cdot C\,.$$

Exprimer à l'aide de ces deux opérations les opérations $A \cup B$ et A^c.
Vérifier que le produit

1) est associatif : $(A \cdot B)\cdot C = A \cdot (B \cdot C)$;
2) est commutatif : $A \cdot B = B \cdot A$;
3) est idempotent : $A \cdot A = A$;
4) possède un élément neutre : $A \cdot \Omega = \Omega \cdot A = A$;
5) $A \cdot \emptyset = \emptyset \cdot A = \emptyset$.

Vérifier que la somme booléenne

1) est associative : $(A \oplus B) \oplus C = A \oplus (B \oplus C)$;
2) est commutatitive : $A \oplus B = B \oplus A$;
3) possède un élément neutre : $A \oplus \emptyset = \emptyset \oplus A = A$;
4) chaque élément a un inverse qui est lui-même, $A \oplus A = \emptyset$.

Indication : pour chaque sous-ensemble A on définit l'indicatrice de A : $I_A(\omega) := 1$ si $\omega \in A$ et $I_A(\omega) := 0$ si $\omega \notin A$; vérifier les identités

$$I_{AB}(\omega) = I_A(\omega) I_B(\omega) \quad \text{et} \quad I_{A \oplus B}(\omega) = I_A(\omega) + I_B(\omega) - 2 I_{AB}(\omega) \,.$$

Exercice 2.7 Vérifier par induction la première affirmation de la proposition 2.2.

Exercice 2.8 Soit $A \Delta B$ la différence symétrique de A et B. Montrer que l'application

$$(A, B) \mapsto d(A, B) := P(A \Delta B)$$

a les propriétés suivantes : $d(A, A) = 0$, $d(A, B) = d(B, A)$ et

$$d(A, C) \leq d(A, B) + d(B, C) \quad \text{(inégalité triangulaire)} \,.$$

Exercice 2.9 Si $A_i \supset B_i$ pour tout $i = 1, \ldots, n$, vérifier les formules

$$\Big(\bigcup_{n \geq 1} A_n \Big) \backslash \Big(\bigcup_{n \geq 1} B_n \Big) \subset \bigcup_{n \geq 1} (A_n \backslash B_n) \,,$$

$$\Big(\bigcap_{n \geq 1} A_n \Big) \backslash \Big(\bigcap_{n \geq 1} B_n \Big) \subset \bigcup_{n \geq 1} (A_n \backslash B_n) \,.$$

Exercice 2.10 On lance trois dés, un jaune, un rouge et un noir et on suppose que chaque résultat est également probable. On somme les résultats de ces trois dés. Les totaux 9 et 10 peuvent être obtenus de 6 manières différentes :

$$\begin{aligned} 9 &= 1+2+6 = 1+3+5 = 1+4+4 \\ &= 2+2+5 = 2+3+4 = 3+3+3 \end{aligned}$$

et

$$\begin{aligned} 10 &= 1+3+6 = 1+4+5 = 2+2+6 \\ &= 2+3+5 = 2+4+4 = 3+3+4 \,. \end{aligned}$$

Pourquoi le résultat 10 est-il plus probable que le résultat 9 ?

Chapitre 3

Des boules et des boîtes

Parmi les modèles classiques d'expériences aléatoires figurent ceux concernant des rangements de boules dans des boîtes ou de façon équivalente des tirages de jetons d'une urne. Les rangements peuvent être effectués de différentes manières, et pour chaque type de rangement on définit un espace fondamental approprié. On considère le cas des rangements et on énonce le problème équivalent en termes de tirages. Si chaque type de rangement (tirage) est fait *au hasard*, expression qui signifie simplement que chaque rangement (tirage) est également probable, alors la mesure de probabilité est définie sur l'algèbre de Boole de tous les sous-ensembles de Ω et son expression est celle donnée par la formule (2.1). Par conséquent il est essentiel de connaître la cardinalité de Ω.

Dans toute cette section on considère M boîtes désignées par $a_1, \ldots, a_M$; les boîtes sont toujours distinctes. On pose

$$\mathtt{A} := \{a_1, \ldots, a_M\}\,.$$

On place n boules dans les boîtes de différentes manières. Alternativement, on considère une urne contenant M jetons qui sont désignés par $a_1, \ldots, a_M$ et on tire n jetons. Dans ce chapitre on choisit d'étudier le modèle abstrait des boules dans des boîtes, car la situation de la section 3.4 n'a pas d'équivalent dans le cas des tirages de jetons. L'importance de ce modèle est qu'il possède de très nombreuses interprétations.

1) Placer n particules dans M niveaux d'énergie distincts.
2) Classer n accidents selon les jours de la semaine ($M = 7$).
3) Les résultats d'un lancer de n dés (à six faces) correspondent aux rangements de n boules dans six boîtes. Les six faces correspondent aux six boîtes. Si les dés sont distinguables les boules sont numérotées, sinon elles ne le sont pas.
4) Les distributions possibles de n erreurs typographiques dans un texte de M symboles ($M \geq n$).
5) Définir une fonction $f\colon \{1, \ldots, n\} \to \{1, \ldots, M\}$ est équivalent à placer n boules (numérotées) dans M boîtes.

3.1 Ranger n boules distinguables dans M boîtes

Les rangements de n boules distinguables (numérotées) dans M boîtes sont codés univoquement en donnant pour chaque boule le symbole de la boîte où elle est rangée. L'espace fondamental est

$$\Omega_a := \{\omega = (\omega_1, \ldots, \omega_n) \colon \forall i,\ \omega_i \in \mathtt{A}\} = \mathtt{A}^n\,; \quad |\Omega_a| = M^n\,.$$

En physique classique, lorsqu'on place des particules dans des niveaux d'énergie distincts, c'est cette situation qu'il faut considérer, car on admet qu'on peut en principe distinguer les particules. Les éléments de Ω_a codent aussi les *tirages ordonnés avec remise de n jetons*, i.e. on enregistre l'ordre dans lequel les jetons sont tirés et le jeton tiré est remis dans l'urne.

3.2 Ranger n boules distinguables, au plus une boule par boîte

On suppose que $M \geq n$. Chaque rangement est codé comme précédemment avec la condition supplémentaire $\omega_i \neq \omega_j$ si $i \neq j$ puisqu'il y a au plus une boule par boîte. L'espace fondamental est

$$\Omega_b := \{\omega = (\omega_1, \ldots, \omega_n) \colon \forall i,\ \omega_i \in \mathtt{A} \text{ et } \omega_i \neq \omega_j\ \forall i \neq j\}\,.$$

On utilise aussi Ω_b pour coder les *tirages ordonnés sans remise de n jetons*, i.e. on enregistre l'ordre dans lequel les jetons sont tirés mais on ne remet pas dans l'urne un jeton qui a été tiré. Pour calculer la cardinalité de Ω_b on utilise le principe de base suivant :

Si une opération globale peut se décomposer en k opérations successives, la première pouvant s'effectuer de n_1 manières différentes, la deuxième de n_2 manières différentes, quelle que soit la manière dont la première opération a été effectuée, la troisième de n_3 manières différentes, quelle que soit la manière dont la première et la deuxième opération ont été effectuées etc, alors l'opération globale peut se faire de $n_1 \cdot n_2 \cdot n_3 \cdot \cdots \cdot n_k$ manières différentes.

Dans notre cas la première boule peut être mise dans M boîtes, la seconde dans $M-1$ boîtes quel que soit le choix de la boîte pour la première boule etc, de sorte que

$$|\Omega_b| = M(M-1)\cdots(M-n+1) \equiv [M]_n\,.$$

Les éléments de Ω_b correspondent aussi aux applications injectives de $\{1, \ldots, n\}$ dans $\mathtt{A}$. Une telle application correspond à un choix ordonné de n objets parmi M objets distincts. Un tel choix est aussi appelé *arrangement sans répétition* (de n objets choisis parmi $\mathtt{A}$). Dans le cas particulier où $M = n$ on parle de *permutation* de n objets. Les permutations de n objets correspondent aux bijections de $\{1, \ldots, n\}$ sur $\{1, \ldots, n\}$. Il y a donc $n!$ (n *factoriel*) permutations, où

$$n! := \begin{cases} n(n-1)(n-2)\cdots 2 \cdot 1 & \text{si } n \geq 1 \\ 1 & \text{si n=0}\,. \end{cases}$$

Avec cette notation $|\Omega_b| = \dfrac{M!}{(M-n)!}$.

Exemple 3.1 On range au hasard n boules distinguables dans M boîtes, $M > n$. Quelle est la probabilité que chaque boîte contienne au plus une boule ? L'espace fondamental est Ω_a car on autorise que les boîtes contiennent plus d'une boule. Par le principe ci-dessus le nombre des cas favorables est $[M]_n$ et la probabilité vaut

$$\frac{[M]_n}{M^n} = \frac{M(M-1)\cdots(M-n+1)}{M^n} = \prod_{k=1}^{n-1}\Big(1-\frac{k}{M}\Big)\,.$$

Pour estimer cette probabilité on utilise l'inégalité $\mathrm{e}^{-x} \geq 1-x$ (voir chap. « Conventions et rappels de mathématiques »). On en déduit la borne supérieure

$$\begin{aligned}\prod_{k=1}^{n-1}\Big(1-\frac{k}{M}\Big) &\leq \exp\Big(-\sum_{k=1}^{n-1}\frac{k}{M}\Big)\\ &= \exp\Big(-\frac{n(n-1)}{2M}\Big) \approx \exp\Big(-\frac{n^2}{2M}\Big)\,.\end{aligned}$$

La probabilité cherchée est approximativement égale à $1/2$ lorsque

$$\ln 2 = \frac{n^2}{2M} \quad \text{i.e. } n = \sqrt{2M\ln 2}\,.$$

Cette approximation est bonne si M est grand par rapport à n, car dans ce cas k/M reste petit et $(1-k/M) \approx \mathrm{e}^{-k/M}$.

Par exemple on choisit au hasard un élément dans un ensemble de cardinalité 100. Si on répète 12 fois cette opération, la probabilité de choisir 12 éléments différents est environ $1/2$. □

3.3 Ranger n boules indistinguables, au plus une boule par boîte

On suppose que $M \geq n$. Pour coder ces rangements il suffit de donner le sous-ensemble ω de $\mathtt{A}$ formé par les boîtes occupées. L'espace fondamental est

$$\Omega_c := \{\omega\colon \omega \subset \mathtt{A}\,,\ \operatorname{card}\omega = n\}\,.$$

De la même façon on code les *tirages non ordonnés sans remise de n jetons*, i.e. on n'enregistre pas l'ordre dans lequel les jetons sont tirés. Pour calculer la cardinalité de Ω_c on utilise la méthode suivante. On suppose que les boules sont numérotées. Dans ce cas l'espace fondamental est Ω_b. Sur Ω_b on introduit une relation d'équivalence : $\omega \sim \omega'$ si et seulement si ω' s'obtient à partir de ω par une permutation des coordonnées. Par exemple si $M = 7$ et $n = 4$, $(a_1, a_4, a_6, a_2) \sim (a_4, a_1, a_2, a_6)$ mais $(a_1, a_4, a_6, a_2) \not\sim (a_7, a_5, a_3, a_2)$. Chaque

rangement des n boules indistinguables est codé univoquement par une classe d'équivalence puisque les classes d'équivalence sont entièrement spécifiées par les sous-ensembles des boîtes occupées. Comme chaque classe d'équivalence a le *même nombre d'éléments* $n!$,

$$|\Omega_c| = \frac{|\Omega_b|}{n!} = \frac{M!}{n!(M-n)!} \equiv \binom{M}{n}.$$

Le nombre $\binom{M}{n}$ est appelé *coefficient binomial* car le binôme de Newton s'écrit

$$(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k .$$

Le nombre de sous-ensembles à n éléments d'un ensemble à M éléments vaut $\binom{M}{n}$. Un élément de Ω_c est aussi appelé *combinaison sans répétition* (de n objets choisis parmi `A`).

Ce cas correspond en physique à distribuer n fermions dans M niveaux d'énergie distincts.

Exemple 3.2 Combien de mots de n lettres peuvent être écrits avec deux lettres a et b si chaque mot contient p fois la lettre a et q fois la lettre b, $p+q=n$? C'est équivalent à ranger p boules noires indistinguables et q boules blanches indistinguables dans $M=p+q$ boîtes, au plus une boule par boîte. Si l'on distingue les boules en numérotant les boules noires par $1,\ldots,p$ et les blanches par $p+1,\ldots,p+q$, on a la situation de la section 3.2 avec $n=M$, soit $(p+q)!$ cas différents. Les boules noires, respectivement blanches, étant indistinguables, le nombre de mots différents est

$$\frac{(p+q)!}{p!\,q!} = \binom{p+q}{p} = \binom{p+q}{q}.$$

Plus directement, le nombre de mots différents est obtenu en choisissant les p places où l'on écrit la lettre a parmi les $p+q$ places disponibles.

Proposition 3.1 *Soit $n_1 \geq 0,\ldots,n_p \geq 0$ tels que $n_1+\cdots+n_p = n$. Le nombre de rangements de n boules distinguables dans p boîtes $a_1,\ldots,a_p$, avec a_1 contenant n_1 boules, a_2 contenant n_2 boules, …, a_p contenant n_p boules est*

$$\frac{n!}{n_1!\,n_2!\,\cdots n_p!} \equiv \binom{n}{n_1\,,n_2\,,\cdots\,,n_p}.$$

Ce nombre est appelé coefficient multinomial.

Dans cette proposition les boîtes sont distinctes. Par exemple, si l'on a quatre boules numérotées et deux boîtes a_1, a_2, il y a 6 rangements distincts tels que $n_1=n_2=2$. Les rangements des boules 1 et 4 dans a_1 et des boules 2 et 3 dans a_2, respectivement des boules 2 et 3 dans a_1 et des boules 1 et 4 dans a_2, sont considérés comme différents.

Preuve En appliquant le principe de la section 3.2 on choisit n_1 boules qu'on met dans la boîte a_1, puis n_2 boules parmi les boules restantes qu'on met dans la boîte a_2 etc. Le nombre cherché est donc

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-\ldots-n_{p-2}}{n_{p-1}} = \frac{n!}{n_1!\,n_2!\cdots n_p!}\,.$$

□

Exemple 3.3 Soit $2n$ objets distinguables. Un *appariement de* $2n$ *objets* est une partition de ces objets en n paires (non ordonnées), chaque objet figurant dans une et une seule paire.

$$\#(\text{appariements de } 2n \text{ objets}) = \frac{(2n)!}{2^n\,n!} = 1\cdot 3\cdot 5\cdots(2n-1)\,. \tag{3.1}$$

Former un appariement de $2n$ objets distincts est équivalent à ranger $2n$ boules distinguables dans n boîtes *indistinguables*, avec deux boules par boîte. Le nombre de rangements de $2n$ boules numérotées dans n boîtes distinctes, chacune contenant 2 boules, est d'après la proposition 3.1

$$\frac{(2n)!}{2!\cdots 2!} = \frac{(2n)!}{2^n}\,.$$

Le nombre d'appariements de $2n$ objets est donné par (3.1) puisque les boîtes sont indistinguables.

Exemple 3.4 Un texte de M symboles contenant $n < M$ erreurs typographiques correspond à placer n boules dans M boîtes, au plus une boule par boîte. Si l'on ne distingue pas les erreurs typographiques et si celles-ci arrivent au hasard, on se trouve dans le cas de cette section.

3.4 Ranger n boules distinguables dans M boîtes ordonnées

Le type de rangement dans cette section diffère de celui de la section 3.1 ou de la proposition 3.1 par le genre des boîtes. Par définition une *boîte ordonnée* est une boîte où l'on tient compte de la manière dont les boules sont rangées à l'intérieur de la boîte : on distingue les rangements $(3,2,7)$ et $(7,2,3)$ des boules 2, 3, 7 dans la même boîte. Pour coder ces rangements, on écrit dans l'ordre, de gauche à droite, les numéros des boules dans la boîte a_1, puis les numéros des boules dans la boîte a_2 etc. On utilise le signe | pour indiquer qu'on passe à la boîte suivante. Par exemple, si $M = 5$ et $n = 7$ on a besoin de $M - 1 = 4$ signes | . Le codage

$$7,1\,|\quad|\,4,2,3\,|\,6,5\,|$$

correspond au rangement ordonné des boules 7 et 1 dans la boîte a_1, 4, 2, 3 dans a_3 et 6, 5 dans a_4 ; les boîtes a_2 et a_5 sont vides. Pour calculer le nombre

de rangements possibles on procède comme dans la section 3.3. Artificiellement on suppose que les signes | sont coloriés de façon différente de sorte qu'ils sont distinguables. Chaque codage colorié contient $(n+M-1)$ symboles différents et toutes les permutations de ces symboles correspondent à $(n+M-1)!$ codages coloriés distincts. On introduit une relation d'équivalence entre les codages coloriés : deux codages coloriés sont dans la même classe d'équivalence si et seulement si ces codages se distinguent uniquement par la couleur des signes | . *Chaque classe d'équivalence a* $(M-1)!$ *éléments* et correspond à un rangement des boules dans les boîtes ordonnées ; par conséquent

$$\#\text{rangements} = \frac{(n+M-1)!}{(M-1)!} = M(M+1)\cdots(M+n-1) \equiv [M]^n \,.$$

3.5 Ranger n boules indistinguables dans M boîtes

Ce cas est équivalent à effectuer des *tirages non ordonnés avec remise de n jetons* d'une urne contenant M jetons. Pour coder les rangements de n boules indistinguables dans M boîtes, il suffit de donner le nombre de boules que contient chaque boîte puisqu'on ne distingue pas les boules.

$$\Omega_e := \Big\{\omega = (\omega_1, \ldots, \omega_M) \colon \omega_i = \#\text{boules dans la boîte } a_i,\ \sum_{i=1}^{M} \omega_i = n\Big\}.$$

Pour calculer la cardinalité de Ω_e on utilise la méthode de la section 3.3. On considère des boules distinguables dans des boîtes ordonnées et on introduit la relation d'équivalence : deux rangements sont équivalents si et seulement s'ils se distinguent uniquement par une permutation des n boules. Comme les boîtes sont ordonnées, chaque classe d'équivalence a $n!$ éléments et donc

$$|\Omega_e| = \frac{(M+n-1)!}{n!(M-1)!} = \binom{M+n-1}{M-1} = \binom{M+n-1}{n} \,.$$

En physique ce cas correspond à distribuer n bosons dans M niveaux d'énergie distints. Le nombre ω_k est le *nombre d'occupation du niveau d'énergie* a_k.

Exemple 3.5 On a trois pièces de monnaie identiques (indistinguables), et tous les lancers de ces trois pièces sont également probables. Quelle est la probabilité d'obtenir deux Faces et un Pile ?

Lancer trois pièces est équivalent à ranger trois boules dans deux boîtes indexées par 0 et 1. Les placements de trois objets indistinguables dans deux boîtes sont codés par les nombres de boules ω_0 et ω_1 dans les boîtes 0 et 1. L'espace fondamental est

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

L'événement « deux Faces et un Pile » est l'événement élémentaire $E = \{(2,1)\}$. Empiriquement on constate que les événements élémentaires ne sont pas également probables et que $P(E) \neq 1/4$. Le bon modèle est celui de l'espace de probabilité de la section 3.1 pour trois boules distinguables où

$$\Omega' = \{\omega = (\omega_1, \omega_2, \omega_3) \colon \omega_i \in \{0,1\}\,,\ i = 1,2,3\}$$

et où chaque événement élémentaire de Ω' est également probable. Comme ici les boules sont indistinguables, l'algèbre de Boole naturelle pour cette expérience n'est pas la collection de tous les événements de Ω', mais c'est l'algèbre $\mathcal{F}$ formée par les événements $\emptyset$, $F_1 = \{(1,1,1)\}$, $F_2 = \{(0,0,0)\}$, $F_3 = \{(1,0,0),(0,1,0),(0,0,1)\}$, $F_4 = \{(0,1,1),(1,0,1),(1,1,0)\}$, ainsi que les unions de ces événements. La probabilité d'obtenir deux Faces et un Pile est $P(F_3) = 3/8$. Il est possible de choisir Ω comme espace fondamental, mais dans ce cas la mesure de probabilité est définie par

$$P(\{(3,0)\}) = P(\{(0,3)\}) = \frac{1}{8} \quad \text{et} \quad P(\{(2,1)\}) = P(\{(1,2)\}) = \frac{3}{8}\,.$$

Remarque 3.1 Le cas de tirages non ordonnés avec remise de n jetons (toujours distinguables) est similaire à l'exemple 3.5. L'espace fondamental est celui de la section 3.1 pour les tirages *ordonnés* avec remise de n jetons, et ce sont ces tirages qui sont considérés comme également probables, car on peut toujours donner l'ordre avec lequel les jetons sont tirés si on le désire. Lorsque les tirages sont non ordonnés, l'algèbre de Boole $\mathcal{F}$ naturelle est celle des sous-ensembles $E \subset \Omega_a$ avec la propriété : si $\omega = (\omega_1, \ldots, \omega_n) \in E$, alors tous les ω' obtenus par une permutation des coordonnées de ω sont dans E. □

Remarque 3.2 Si dans l'exemple 3.5 on remplace les pièces par trois bosons que l'on place dans deux niveaux d'énergie distincts, alors la mesure de probabilité sur Ω est

$$P(\{(3,0)\}) = P(\{(0,3)\}) = P(\{(2,1)\}) = P(\{(1,2)\}) = \frac{1}{4}\,.$$

C'est la physique, non des arguments logiques, qui a montré l'importance du modèle discret où l'espace fondamental est Ω_e et chaque événement élémentaire est également probable. L'indistinguabilité des bosons est fondamentale ; on ne peut pas «marquer» des bosons, contrairement aux pièces de monnaie. □

Exemple 3.6 Le nombre de p-uples $(n_1, \ldots, n_p)$, tels que $n_i \geq 0$ et $n_1 + \cdots + n_p = n$, est égal à

$$\binom{n+p-1}{p-1}\,.$$

En effet, chaque p-uple correspond à un élément de Ω_e si $M = p$. Si l'on impose que $n_i \geq 1$ pour tout i, ce nombre est

$$\binom{n-1}{p-1}\,.$$

On obtient ce résultat en posant $n_i = 1+m_i$ et en se reportant au cas précédent avec $m_1 + \cdots + m_p = n - p$. □

Exemple 3.7 On considère un système de n bosons et M niveaux d'énergie distincts.
1. Pour un niveau fixé, par exemple le niveau a_1, quelle est la probabilité que ce niveau contienne k particules ? D'après la remarque 3.2, chaque configuration possible du système est également probable. Le nombre de configurations du système avec k particules dans un niveau fixé est égal au nombre de configurations pour $(n-k)$ bosons répartis sur les $M-1$ autres niveaux ; la probabilité cherchée est

$$P_k = \frac{\binom{(M-1)+(n-k)-1}{n-k}}{\binom{M+n-1}{n}} = \frac{(M-1)n(n-1)\cdots(n-k+1)}{(M+n-1)\cdots(M+n-k-1)}\,.$$

2. Quelle est la probabilité qu'exactement m niveaux d'énergie restent vides ? Il y a $\binom{M}{m}$ choix de m niveaux parmi les M niveaux d'énergie ; le nombre de configurations avec m niveaux spécifiés vides, les autres contenant au moins une particule, est $\binom{n-1}{M-m-1}$ (voir exemple 3.6). La probabilité cherchée est

$$\frac{\binom{M}{m}\binom{n-1}{M-m-1}}{\binom{M+n-1}{n}}\,.$$

3. On considère la limite des P_k, lorsque $M \to \infty$ et $n \to \infty$ de sorte que la densité des particules par niveau est fixée, i.e. $\frac{n}{M} = \lambda > 0$. En remplaçant n par λM dans l'expression de P_k on obtient, lorsque $M \to \infty$,

$$P_k = \frac{(M-1)(\lambda M)^k\left(1-\frac{1}{\lambda M}\right)\cdots\left(1-\frac{k-1}{\lambda M}\right)}{M^{k+1}\left(1+\frac{\lambda M-1}{M}\right)\cdots\left(1+\frac{\lambda M-k-1}{M}\right)} \to P'_k = \lambda^k \frac{1}{(1+\lambda)^{k+1}}\,.$$

Noter que $P'_0 > P'_1 > P'_2 > \cdots$. □

On termine cette section en énonçant la formule de Stirling (1692-1770) qui joue un rôle important par la suite lorsqu'on étudie le comportement asymptotique d'expressions faisant intervenir des coefficients binomiaux.

Lemme 3.1 (Formule de Stirling) *Pour tout $n \in \mathbb{N}$,*

$$\sqrt{2\pi n}\, n^n\, \mathrm{e}^{-n} \exp\frac{1}{12n+1} < n! < \sqrt{2\pi n}\, n^n\, \mathrm{e}^{-n} \exp\frac{1}{12n}\,.$$

Preuve Pour la démonstration complète du lemme voir [Fe1]. Voir aussi l'exemple 12.3 de la section 12.3. On démontre des inégalités un peu plus faibles,

$$\sqrt{\mathrm{e}n}\, n^n\, \mathrm{e}^{-n} \le n! \le \mathrm{e}\sqrt{n}\, n^n\, \mathrm{e}^{-n} \qquad \forall\, n \ge 1\,. \tag{3.2}$$

Pour $n \geq 2$,

$$n \ln n - n + 1 = \sum_{k=1}^{n-1} \int_k^{k+1} \ln x \, dx \,. \tag{3.3}$$

En utilisant la concavité de $\ln x$ on estime chaque intégrale du côté droit de (3.3).

$$\begin{aligned} \frac{\ln(k+1) + \ln k}{2} &\leq \int_k^{k+1} \ln x \, dx \leq \int_0^1 \Big(\ln(k+1) + \frac{x-1}{k+1} \Big) dx \\ &= \ln(k+1) - \frac{1}{2} \frac{1}{k+1} \,. \end{aligned}$$

La borne inférieure est l'aire sous la corde dont les extrémités sont $(k, \ln k)$ et $(k+1, \ln(k+1))$. L'équation de la tangente du logarithme en $k+1$ vérifie

$$\ln(k+1+u) \leq \ln(k+1) + \frac{u}{k+1} \,;$$

on pose $u = x - 1$ et on intègre par rapport à x sur l'intervalle $[0,1]$ pour obtenir la borne supérieure. De ces inégalités, si $n \geq 2$,

$$\begin{aligned} n \ln n - n + 1 &= \sum_{k=1}^{n-1} \int_k^{k+1} \ln x \, dx \\ &\geq \sum_{k=2}^{n} \ln k - \frac{1}{2} \ln n = \ln n! - \frac{1}{2} \ln n \end{aligned}$$

et

$$\begin{aligned} n \ln n - n + 1 &= \sum_{k=1}^{n-1} \int_k^{k+1} \ln x \, dx \leq \sum_{k=2}^{n} \ln k - \frac{1}{2} \sum_{k=2}^{n} \frac{1}{k} \\ &= \ln n! - \frac{1}{2} \sum_{k=1}^{n} \frac{1}{k} + \frac{1}{2} \\ &\leq \ln n! - \frac{1}{2} \int_1^n \frac{1}{x} \, dx + \frac{1}{2} = \ln n! - \frac{1}{2} \ln n + \frac{1}{2} \,. \end{aligned}$$

Ceci prouve les inégalités (3.2) pour $n \geq 2$ en prenant l'exponentielle de ces expressions. Ces inégalités restent vraies si $n = 1$. □

Le lemme 3.1 donne des estimations très précises de l'erreur relative lorsqu'on remplace $n!$ par la formule de Stirling ; cette erreur relative tend rapidement vers 0 :

$$e(n) := \frac{n! - \sqrt{2\pi n}\, n^n \, \mathrm{e}^{-n}}{n!} \,; \qquad \begin{array}{ccccccc} n & : & 10 & & 50 & & 100 \\ e(n) & : & 0{,}008 & & 0{,}001 & & 0{,}00008 \end{array} \,.$$

Exemple 3.8 . On lance une pièce de monnaie équilibrée $2n$ fois. Quelle est la probabilité qu'on obtienne exactement n fois Piles ? Ce problème est équivalent

à placer au hasard $2n$ boules distinguables dans deux boîtes et de calculer la probabilité que chaque boîte contienne n boules, ou pour la marche aléatoire de l'exemple 2.1 de calculer la probabilité qu'on retourne à l'origine après $2n$ pas. Cette probabilité est

$$u_{2n} = 2^{-2n}\binom{2n}{n} \simeq \frac{\sqrt{2\pi 2n}\,(2n)^{2n}\,\mathrm{e}^{-2n}}{\left(\sqrt{2\pi n}\,n^n\,\mathrm{e}^{-n}\right)^2} 2^{-2n} = \frac{1}{\sqrt{\pi n}}\,. \tag{3.4}$$

Pour $n = 50$, $u_{100} \approx 0{,}08$. Noter que la probabilité qu'une boîte contienne $n-k \neq n$ boules et l'autre $n+k$ boules vaut

$$2^{-2n}\binom{2n}{n-k} < 2^{-2n}\binom{2n}{n} = u_{2n}\,.$$

Pour la marche aléatoire, cela signifie que

$$P(\ell_{2n} = 2k) = P(\ell_{2n} = -2k) < P(\ell_{2n} = 0) \equiv u_{2n}\,. \tag{3.5}$$

3.6 Exercices

Exercice 3.1 On considère m boîtes dans lesquelles on met à des temps discrets $t_1 = 1, t_2 = 2, \ldots, t_n = n$ une boule en choisissant la boîte au hasard.
a) Modéliser ce problème en définissant un espace fondamental Ω et une mesure de probabilité sur Ω qui traduit le processus de remplissage décrit ci-dessus.
b) Exprimer mathématiquement l'événement A_r « au temps t_r pour la première fois il y a deux boules dans une même boîte ». Calculer la probabilité de cet événement.
c) Exprimer mathématiquement l'événement B_r « il faut attendre au moins jusqu'au temps t_r pour avoir deux boules dans une même boîte ».
d) Exprimer mathématiquement l'événement C « après le remplissage aucune boîte ne contient plusieurs boules ».

Exercice 3.2 Vérifier les trois identités suivantes.

$$\binom{m+n}{r} = \sum_{j=0}^{r}\binom{n}{j}\binom{m}{r-j} \quad \text{(convention : } \tbinom{n}{j} = 0 \text{ si } j > n\text{).} \tag{1}$$

Indication : considérer un ensemble de n boules (distinguables) noires et de m boules (distinguables) blanches.

$$\binom{n}{r} = \sum_{j=r}^{n}\binom{j-1}{r-1}\,, \quad n \geq r\,. \tag{2}$$

Indication : compter les sous-ensembles de $\{1, \ldots, n\}$ ayant r éléments tels que le plus grand élément est j.

$$\sum_{j=1}^{n} j\binom{n}{j} = n \cdot 2^{n-1}\,. \tag{3}$$

Indication : un *ensemble pointé* est un ensemble non vide dont un des éléments a été distingué. Compter de deux façons différentes le nombre de sous-ensembles pointés d'un ensemble de cardinalité n.

Exercice 3.4 Soit Ω un ensemble de cardinalité n. Montrer que

$$|\mathcal{P}(\Omega)| = 2^n \,.$$

Indication : utiliser le binôme de Newton.

Exercice 3.5 Considérer les vecteurs $(x_1, \ldots, x_r)$ avec composantes x_i non négatives et entières. Combien y a-t-il de tels vecteurs

1. si $\sum_{i=1}^r x_i \leq n$;
2. si $\sum_{i=1}^r x_i \leq n$ et $x_i \geq 1$ pour tout i ;
3. si $\sum_{i=1}^r x_i = n$ et il y a exactement k composantes égales à zéro.

Exercice 3.6 On considère un système de n particules pouvant être chacune dans r états d'énergie E_k distincts. Le nombre de particules dans l'état E_k est n_k ($\sum_{k=1}^r n_k = n$). Calculer la probabilité d'obtenir une configuration des particules avec $n_1, \ldots, n_r$ fixés, dans les trois cas qui suivent.

1) Les particules obéissent à la statistique de Maxwell-Boltzmann, i.e. les particules sont distinguables, plusieurs particules peuvent être dans le même état d'énergie.
2) Les particules obéissent à la statistique de Bose-Einstein, i.e. les particules sont indistinguables, plusieurs particules peuvent être dans le même état d'énergie.
3) Les particules obéissent à la statistique de Fermi-Dirac, i.e. les particules sont indistinguables, au plus une particule par niveau d'énergie.

Exemple numérique : $n = 3$, $r = 5$ avec $n_1 = n_3 = n_4 = 1$ et $n_2 = n_5 = 0$.

Exercice 3.7 Quelle est la valeur minimale de r pour que dans un groupe de r personnes la probabilité de l'événement « au moins deux personnes ont leur anniversaire le même jour » soit supérieure à $1/2$? On fait la modélisation suivante. Une année a 365 jours ; il y a équiprobabilité pour les anniversaires ; les individus sont distinguables.

Que se passe-t-il si les personnes sont remplacées par des bosons et les dates d'anniversaire par des niveaux d'énergie ?

Exercice 3.8 Loterie « 6 de 36 ». Chaque participant choisit 6 nombres différents parmi les 36 nombres $\{1, 2, \ldots, 36\}$. On tire au hasard 6 nombres différents. On gagne si 3 des nombres au moins de son choix coïncident avec la série des 6 nombres tirés. Calculer la probabilité de gagner.

Exercice 3.9 Dans le cadre du modèle de l'exemple 3.4 on considère un texte de n pages contenant chacune N symboles. On suppose qu'il y a r fautes de frappe dans tout le texte.
a) Calculer la probabilité que le nombre de fautes de frappe à la page 1 est r_1, à la page 2 est r_2, ..., à la page n est r_n.
b) Montrer que si N devient grand cette probabilité converge vers

$$\frac{1}{n^r}\frac{r!}{r_1!\cdots r_n!}$$

qui est la probabilité de placer r boules distinguables dans n boîtes avec r_1 boules dans la boîte 1, r_2 boules dans la boîte 2, ..., r_n boules dans la boîte n.
Indication : utiliser la formule de Stirling.

Exercice 3.10 On place au hasard n boules distinguables dans M boîtes.
a) Montrer en utilisant l'inégalité

$$\mathrm{e}^{-x-x^2} \leq 1-x \quad \text{si } 0 \leq x \leq \frac{1}{2} \qquad (\star)$$

qu'il existe une constante c telle que si $n \leq c\sqrt{M}$, alors la probabilité qu'aucune boîte ne contienne plus d'une boule est au moins $1/2$.
b) Montrer l'inégalité $(\star)$.
Indication : utiliser

$$\ln \frac{1}{1-x} = x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots \quad \text{si } 0 \leq x < 1.$$

Chapitre 4

Probabilité conditionnelle et indépendance

A la place de considérer un événement A il est souvent préférable de considérer la partition $\{A, A^c\}$ de Ω. Cette partition correspond à une *question simple* dont la réponse est *oui* ou *non* : si $\omega \in A$ la réponse est *oui* et si $\omega \in A^c$ la réponse est *non*. Une *partition* Q de Ω en p événements est une décomposition de Ω en p sous-ensembles, $Q = \{A_1, \ldots, A_p\}$, disjoints deux à deux et tels que $\bigcup_i A_i = \Omega$. Une partition correspond à une *question à choix multiple*. Si l'on connaît $P(A_i)$ pour chaque $i = 1, \ldots, p$, on connaît $P(B)$ pour chaque B appartenant à l'*algèbre de Boole engendrée* par la partition. Cette algèbre est par définition la plus petite algèbre de Boole contenant tous les A_i. C'est la collection des sous-ensembles

$$\mathcal{A} = \left\{ B \subset \Omega \colon \exists J \subset \{1, \ldots, p\}\,,\ B = \bigcup_{i \in J} A_i \right\}$$

(on fait la convention que si $J = \emptyset$, alors $B = \emptyset$). La vérification est facile (voir lemme 4.1). Comme les A_i sont disjoints, $P(B) = \sum_{i \in J} P(A_i)$. A partir de deux partitions $Q_1 = \{A_1, \ldots, A_p\}$ et $Q_2 = \{B_1, \ldots, B_q\}$ on peut former une nouvelle partition

$$Q_1 \vee Q_2 := \{A_i \cap B_j \colon 1 \leq i \leq p\,,\ 1 \leq j \leq q\}\,.$$

Exemple 4.1 Soit trois urnes : U_1 contenant 2 jetons noirs et 3 blancs, U_2 contenant 1 jeton noir et 1 blanc, U_3 contenant 1 jeton noir et 4 blancs. On choisit au hasard une urne et puis on tire au hasard un jeton de l'urne choisie. Cette expérience est analysée à l'aide des questions $Q_1 = \{C_1, C_2, C_3\}$ où C_i est l'événement « l'urne choisie est U_i », et $Q_2 = \{A, A^c\}$ où A est l'événement « le jeton tiré est noir ». Ce qui nous intéresse, c'est la réponse à la question $Q_1 \vee Q_2$ car elle donne le résultat final, quelle est l'urne choisie et quelle est la couleur du jeton. □

L'analyse d'une expérience aléatoire par des questions successives $Q_1 = \{A_1, \ldots, A_p\}$, $Q_2 = \{B_1, \ldots, B_q\}$ ... est parfois schématisée par un *diagramme en arbre* qui est un graphe orienté. La *racine* de l'arbre est un sommet S_0 qui représente Ω ; à partir de S_0 sont issus p liens orientés $\langle S_0, S_1\rangle, \ldots, \langle S_0, S_p\rangle$ vers

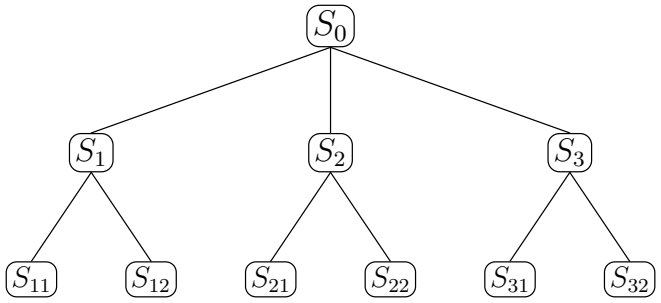


Figure 4.1 – Diagramme en arbre de l'exemple 4.1.

les p sommets $S_1, \ldots, S_p$ qui représentent respectivement les p événements de la question Q_1. A partir de chaque sommet S_i sont issus q liens orientés $\langle S_i, S_{i1}\rangle$, $\ldots$, $\langle S_i, S_{iq}\rangle$ vers les sommets $S_{i1}, \ldots, S_{iq}$ qui représentent respectivement les événements $A_i \cap B_j$, $j = 1, \ldots, q$. Les sommets S_{ij}, $i = 1, \ldots, p$ et $j = 1, \ldots, q$, représentent les événements de la question $Q_1 \vee Q_2$. Cette construction est itérée autant de fois que nécessaire. Les sommets du graphe dont ne sont issus aucun lien sont les *feuilles* de l'arbre. Si l'arbre est construit à partir de m questions $Q_1, \ldots, Q_m$, ces feuilles représentent les événements de la question $Q_1 \vee \cdots \vee Q_m$. Pour analyser de cette manière une expérience aléatoire on utilise la notion de probabilité conditionnelle.

4.1 Probabilité conditionnelle

Soit $(\Omega, \mathcal{F}, P)$ un espace de probabilité. Si l'on sait que $A \in \mathcal{F}$ est réalisé, comment définir une nouvelle mesure de probabilité qui tienne compte de cette information ? Si $B \in \mathcal{F}$ est réalisé, alors on sait que $B \cap A$ est réalisé. Si Ω est fini et tous les résultats sont également probables, la probabilité de B, sachant que A est réalisé, est donnée par

$$\frac{\text{nombre de cas favorables à la réalisation des événements } A \text{ et } B}{\text{nombre de cas favorables à la réalisation de l'événement } A}\,.$$

Dans le cas général on remarque que l'application $B \mapsto P(B \cap A)$ est σ-additive sur $\mathcal{F}$; pour obtenir une mesure de probabilité il suffit de la normaliser en divisant par $P(\Omega \cap A) = P(A)$.

Définition 4.1 *Soit $A \in \mathcal{F}$ tel que $P(A) > 0$. Pour tout $B \in \mathcal{F}$,*

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

est la probabilité conditionnelle de B sachant A. *Si* $P(A) = 0$, $P(B|A)$ *n'est pas défini.*

Si l'on connaît $P(B|A)$, on connaît aussi $P(B^c|A) = 1 - P(B|A)$. Mais de $P(B|A)$ on ne peut rien déduire sur $P(B|A^c)$ sauf dans le cas où les événements A et B sont indépendants (voir section 4.2). Noter l'identité

$$P(B \cap A) = P(B|A)P(A) = P(A|B)P(B) \tag{4.1}$$

dont la validité peut être étendue aux cas $P(A) = 0$ ou $P(B) = 0$. Lorsque $P(A) > 0$ et $P(B) > 0$ cette identité peut s'écrire sous la forme

$$P(A|B) = P(A)\,\frac{P(B|A)}{P(B)}\,. \tag{4.2}$$

Proposition 4.1 *Soit* $\{A_1, \ldots, A_k\}$ *une partition de* Ω.

1. Formule des probabilités totales: *si* $B \in \mathcal{F}$,

$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)\,.$$

2. Formule de Bayes (1702-1761): *si* $P(B) > 0$,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}\,.$$

3. Formule de multiplication: *si* $B_1 \in \mathcal{F}, \ldots, B_m \in \mathcal{F}$,

$$P(B_1 \cdots B_m) = P(B_1)P(B_2|B_1)P(B_3|B_1B_2)\cdots P(B_m|B_1\cdots B_{m-1})\,.$$

Preuve B est l'union disjointe des $B \cap A_i$; l'identité (4.1) implique

$$P(B) = \sum_{i=1}^{k} P(BA_i) = \sum_{i=1}^{k} P(B|A_i)P(A_i)\,.$$

La formule de Bayes est une conséquence de (4.2) pour $A = A_i$ et de la formule des probabilités totales. Enfin, par itération de (4.1) on obtient

$$\begin{aligned} P(B_1 \cdots B_m) &= P(B_1 \cdots B_{m-1})P(B_m|B_1 \cdots B_{m-1}) \\ &= P(B_1 \cdots B_{m-2})P(B_{m-1}|B_1 \cdots B_{m-2})P(B_m|B_1 \cdots B_{m-1}) \\ &\cdots \\ &= P(B_1)P(B_2|B_1)P(B_3|B_1B_2)\cdots P(B_m|B_1\cdots B_{m-1})\,. \end{aligned}$$

□

Remarque 4.1 Si $\{B_1, \ldots, B_q\}$ est une partition de Ω, les $P(B_i)$ définissent sur l'algèbre de Boole $\mathcal{B}$ engendrée par cette partition une mesure de probabilité

qui est appelée *mesure de probabilité a priori* sur $\mathcal{B}$. Si l'on sait que l'événement A est réalisé, alors la mesure de probabilité sur $\mathcal{B}$ qui tient compte de cette information est celle donnée par les $P(B_i|A)$; elle est appelée *mesure de probabilité a posteriori* sur $\mathcal{B}$. □

Lorsqu'on analyse une expérience aléatoire à l'aide d'un diagramme en arbre, on donne un poids à chaque *lien* du diagramme : si le lien est $\langle S, S' \rangle$, alors le poids est $P(A'|A)$; A est l'événement représenté par S et A' celui par S'. Noter que si S est la racine, $P(A'|A) = P(A'|\Omega) = P(A')$. Pour trouver la probabilité de l'événement représenté par une feuille de l'arbre on applique la formule de multiplication de la proposition 4.1.

$$P(B_1 \cdots B_k | B_1 \cdots B_{k-1}) = P(B_k | B_1 \cdots B_{k-1}) \,,$$

la probabilité est donnée par le produit des poids le long de l'unique chemin allant de la racine à la feuille.

Exemple 4.2 Suite de l'exemple 4.1. Le choix au hasard de l'urne à la première étape est modélisé par $P(C_1) = P(C_2) = P(C_3) = 1/3$. Sachant qu'on a choisi l'urne U_i on détermine de la même manière $P(A|C_i)$: $P(A|C_1) = 2/5$, $P(A|C_2) = 1/2$ et $P(A|C_3) = 1/5$. On peut alors calculer les probabilités $P(C_i A) = P(A|C_i)P(C_i)$ et $P(C_i A^c) = P(A^c|C_i)P(C_i)$. Par la formule de Bayes on obtient par exemple la probabilité que l'urne U_1 a été choisie si l'on a tiré un jeton noir,

$$P(C_1|A) = \frac{P(A|C_1)P(C_1)}{P(A|C_1)P(C_1) + P(A|C_2)P(C_2) + P(A|C_3)P(C_3)} = \frac{4}{11} \,.$$

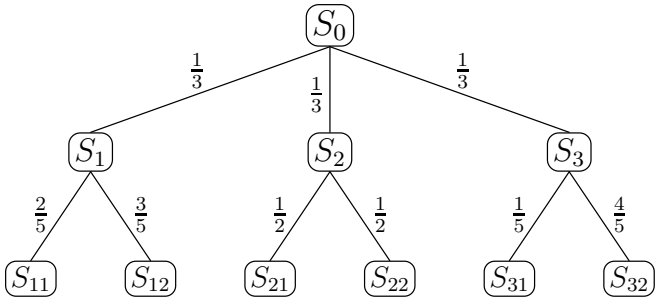


Figure 4.2 Diagramme en arbre avec poids de l'exemple 4.1.

Exemple 4.3 Un laboratoire veut exprimer l'efficacité d'un test pour détecter un virus. Il le fait en donnant deux probabilités conditionnelles $P(E|V^c)$ et $P(E|V)$, où E et V sont respectivement les événements « le test est positif » et « la personne est porteuse du virus ». Ces probabilités conditionnelles sont

déterminées en procédant à des expériences sur une population saine, respectivement porteuse du virus. On teste un grand nombre d'individus pour obtenir un résultat fiable. L'efficacité du test est formulée par exemple ainsi :

$$P(E|V^c) = 0{,}001 \qquad P(E|V) = 0{,}99\,.$$

L'interprétation de ces probabilités est donnée par la loi des grands nombres (voir chap. 10) : $P(E|V^c) = 0{,}001$ signifie (approximativement) qu'une fois sur mille le test est positif, sachant que la personne est saine. Lorsqu'on fait une campagne de dépistage ce sont les probabilités $P(V \cap E)$ ou $P(V|E)$ qui nous intéressent ; $P(V|E)$ est la probabilité de porter le virus, sachant que le test est positif. Pour calculer ces probabilités à partir des données fournies par le laboratoire, il faut connaître $P(V)$, i.e. l'état sanitaire de la population. □

Exemple 4.4 Un système permettant la transmission d'information est un *canal*. Le canal accepte des données qui sont représentées dans le cas le plus simple par les éléments d'un ensemble fini $\mathbb{A} := \{a_1, \ldots, a_m\}$ qu'on nomme *alphabet d'entrée*. Les données de sortie du canal sont représentées par les éléments de $\mathbb{B} := \{b_1, \ldots, b_n\}$, l'*alphabet de sortie*. Le canal est aléatoire si la lettre de sortie n'est pas univoquement déterminée par la lettre d'entrée. Un exemple simple, mais important, est celui d'un *canal symétrique binaire* dont les alphabets sont $\mathbb{A} = \mathbb{B} = \{0, 1\}$. Avec probabilité $1 - p$ la lettre de sortie coincide avec la lettre d'entrée, et avec probabilité p elle diffère. Ici p est petit et représente la probabilité d'une erreur de transmission.

Le caractère aléatoire du canal est décrit par les probabilités conditionnelles $P(\,\cdot\,|a_i)$ indexées par $a_i \in \mathbb{A}$: $P(b_k|a_j)$ est la probabilité de recevoir b_k si l'on a envoyé a_j à travers le canal. Les propriétés du canal sont spécifiées par la matrice $\mathbf{M}$,

$$\mathbf{M}_{jk} := P(b_k|a_j) \quad j = 1, \ldots, m\,,\ k = 1, \ldots, n\,.$$

Attention à l'ordre des indices ! $\mathbf{M}$ est une *matrice stochastique*, i.e.

$$\mathbf{M}_{jk} \geq 0 \quad \text{et} \quad \sum_{k=1}^{n} \mathbf{M}_{jk} = 1 \quad \forall\, j\,.$$

Dans le cas du canal symétrique binaire $P(0|0) = P(1|1) = 1 - p$ et $P(1|0) = P(0|1) = p$.

En théorie de l'information la source qui alimente le canal est souvent elle-même aléatoire. Dans ce cas on donne aussi pour chaque j la probabilité $P(A_j)$ de l'événement « la lettre a_j est envoyée par le canal ».

Pour l'espace fondamental de cette expérience aléatoire, on choisit $\Omega = \mathbb{A} \times \mathbb{B}$ et $\mathcal{F} = \mathcal{P}(\Omega)$;

$$A_j := \{a_j\} \times \mathbb{B} \equiv \{\text{la lettre } a_j \text{ est envoyée}\}$$
$$B_k := \mathbb{A} \times \{b_k\} \equiv \{\text{la lettre } b_k \text{ est reçue}\}\,.$$

La mesure de probabilité est donnée par

$$P(A_j \cap B_k) = P(B_k|A_j)\, P(A_j) = P(A_j)\, \mathbf{M}_{jk}\,.$$

On suppose qu'on reçoive la lettre b_k. Quelle est la lettre qui a été envoyée dans le canal ? Une réponse à cette question est formalisée par le choix d'une *fonction de décision*, $\varphi : \{1, \ldots, n\} \to \{1, \ldots, m\}$, dont l'interprétation est

$$a_{\varphi(k)} \text{ est la lettre d'entrée conjecturée si l'on reçoit } b_k\,.$$

Quel est le « meilleur choix » pour φ ? Une possibilité est de déterminer φ de sorte que la probabilité de faire une erreur soit minimale. On fait une erreur chaque fois qu'on reçoit b_k et que la lettre envoyée est différente de $a_{\varphi(k)}$. La fonction de décision φ doit minimiser l'expression

$$\begin{aligned} Z_\varphi :&= \sum_{\substack{i,k:\\ \varphi(k)\neq i}} P(A_i \cdot B_k) \\ &= \sum_{k=1}^{n} \big(P(B_k) - P(A_{\varphi(k)} \cdot B_k)\big) \\ &= \sum_{k=1}^{n} P(B_k)\big(1 - P(A_{\varphi(k)}|B_k)\big)\,. \end{aligned}$$

Si $P(B_k) = 0$ on peut définir $\varphi(k)$ n'importe comment. Si $P(B_k) > 0$ il faut choisir $\varphi(k)$ de sorte que

$$P(A_{\varphi(k)}|B_k) = \max_{j=1}^{m} P(A_j|B_k)\,.$$

Pour résoudre ce problème il faut connaître $P(A_j)$ et utiliser la formule de Bayes pour calculer $P(A_j|B_k)$. Le résultat obtenu est plausible : si l'on reçoit b_k, on choisit $\varphi(k)$ de sorte que la lettre d'entrée $a_{\varphi(k)}$ a la plus grande probabilité d'être envoyée, sachant que b_k est la lettre reçue.

Si l'on n'a aucune information sur la source, une manière naturelle de choisir $\varphi(k)$ est de prendre le maximum de la fonction $j \mapsto P(B_k|A_j)$ de sorte que

$$P(B_k|A_{\varphi(k)}) \geq P(B_k|A_i) \quad \forall\, i \in \{1, \ldots m\}\,.$$

Si l'on suppose que les lettres envoyées sont également probables, en calculant $P(A_j|B_k)$ et en choisissant l'indice j qui donne le maximum de cette expression, on retrouve cette dernière fonction de décision. □

Exemple 4.5 On modifie l'exemple 4.1 en choisissant l'urne U_1 avec probabilité 0,85, l'urne U_2 avec probabilité 0,05 et l'urne U_3 avec probabilité 0,1. Une fois l'urne choisie, on effectue des tirages avec remise de cette urne en indiquant seulement le nombre de fois qu'un jeton noir a été tiré, par exemple, le résultat des tirages est l'événement E « 1 jeton noir a été tiré sur 10 tirages ».

Sachant que l'événement E est réalisé, quelle est l'urne qui a été utilisée pour les tirages ? On calcule $P(E|C_i)$ pour $i = 1, 2, 3$. Pour faire ce calcul, on utilise le modèle de la remarque 3.1 section 3.5. Par exemple, pour l'urne U_1 les jetons a_1 et a_2 sont noirs et les jetons a_3, a_4 et a_5 sont blancs ; l'espace fondamental est celui des tirages ordonnés avec remise et les événements élémentaires de cet espace sont également probables. Dans le cas de l'urne U_1, $|E| = 10 \cdot 2 \cdot 3^9$ (le jeton noir peut être tiré au premier tirage ou au deuxième etc, et à chaque fois il y a 2 possibilités d'obtenir un jeton noir ; il y a 3^9 possibilités pour tirer les jetons blancs). On obtient

$$P(E|C_1) = 0{,}040 \qquad P(E|C_2) = 0{,}009 \qquad P(E|C_3) = 0{,}268\,.$$

Si l'on n'avait pas d'information supplémentaire, la réponse la plus vraisemblable serait « l'urne U_3 ». Cependant on connaît $P(C_i)$, et on peut donc calculer

$$P(C_1|E) = 0{,}556 \qquad P(C_2|E) = 0{,}007 \qquad P(C_3|E) = 0{,}435\,.$$

On voit que l'information supplémentaire change complètement notre perception de la situation. □

Exemple 4.6 On considère une urne avec n jetons et on procède à un tirage ordonné sans remise des n jetons. L'espace fondamental est celui de la section 3.2 avec $n = M$. On tire les jetons les uns après les autres. Sachant que les j premiers jetons tirés sont $b_1, \ldots, b_j$, quelle est la probabilité que les m suivants sont $b_{j+1}, \ldots, b_{j+m}$? Soit $\{b_1, \ldots, b_j\} \cap \{b_{j+1}, \ldots, b_{j+m}\} = \emptyset$ et

$$E := \{\omega \colon \omega_i = b_i\,,\ i = 1, \ldots, j\}\,,\ F := \{\omega \colon \omega_i = b_i\,,\ i = j+1, \ldots, j+m\}\,.$$

On obtient

$$P(F|E) = \frac{|E \cap F|}{|E|} = \frac{[n-j-m]_{n-j-m}}{[n-j]_{n-j}} = \frac{1}{[n-j]_m}\,.$$

Cette probabilité conditionnelle est égale à la probabilité d'un tirage ordonné sans remise de m jetons d'une urne contenant $n - j$ jetons.

On suppose que les jetons $a_1, \ldots, a_k$ sont noirs et les autres blancs. On introduit les applications

$$X_\ell(\omega) := \begin{cases} 1 & \text{si } \omega_\ell \text{ est un jeton noir} \\ 0 & \text{si } \omega_\ell \text{ est un jeton blanc.} \end{cases}$$

On note l'événement « le $j^{\text{ième}}$ jeton tiré est noir » par $\{X_j = 1\}$. On a

$$P(\{X_1 = 1\}) = \frac{k}{n} \qquad P(\{X_1 = 0\}) = \frac{n-k}{n}$$

et

$$P(\{X_2 = 1\}|\{X_1 = 1\}) = \frac{k-1}{n-1} \qquad P(\{X_2 = 1\}|\{X_1 = 0\}) = \frac{k}{n-1}\,.$$

On tire le premier jeton ; sans connaître sa couleur, quelle est la probabilité que le deuxième jeton tiré soit noir ?

$$P(\{X_2 = 1\}) = P(\{X_2 = 1\}|\{X_1 = 1\})\, P(\{X_1 = 1\}) \\ + P(\{X_2 = 1\}|\{X_1 = 0\})\, P(\{X_1 = 0\}) = \frac{k}{n}\,.$$

De même, si l'on tire les $j-1$ premiers jetons sans prendre connaissance des couleurs de ces jetons, quelle est la probabilité que le $j^{\text{ième}}$ jeton tiré soit noir ? Les événements

$$E(b_1, \cdots, b_{j-1}) := \{\omega \colon \omega_i = b_i\,,\ b_i \in \mathtt{A}\,,\ i = 1, \ldots, j-1\}$$

forment une partition de Ω_b. On peut calculer comme ci-dessus la probabilité $P(\{X_j = 1\})$ par la formule des probabilités totales, mais il est plus simple de calculer directement $P(\{X_j = 1\})$. En effet, la cardinalité de $\{X_j = 1\}$ est $k \cdot [n-1]_{n-1}$: il y a k possibilités pour avoir un jeton noir au $j^{\text{ième}}$ tirage et le nombre de choix pour les $n-1$ autres tirages est $[n-1]_{n-1}$; par conséquent

$$P(\{X_j = 1\}) = \frac{k \cdot [n-1]_{n-1}}{[n]_n} = \frac{k}{n}\,.$$

Même si $j = n$, tant que l'on n'a pas *pris connaissance du résultat* des $n-1$ tirages précédents, $P(\{X_n = 1\}) = k/n$, bien qu'au dernier tirage il ne reste plus qu'une boule dans l'urne. Ce qui importe, ce n'est pas que les $n-1$ premiers tirages aient eu lieu, mais ce que nous savons de ces $n-1$ premiers tirages. □

4.2 Indépendance

L'indépendance est un concept-clé de la théorie. On considère dans cette section l'indépendance d'événements. Un point essentiel à retenir est que le concept d'indépendance d'événements *ne fait sens que si l'on a fixé une mesure de probabilité*. Ce n'est pas une propriété intrinsèque des événements.

Exemple 4.7 On considère deux expériences aléatoires distinctes qui sont décrites par des espaces de probabilité discrets $(\Omega_1, \mathcal{F}_1, P_1)$ et $(\Omega_2, \mathcal{F}_2, P_2)$. La cardinalité de Ω_i est n_i et tous les événements élémentaires de chaque expérience sont également probables. On considère ces deux expériences comme une seule « grande » expérience dont l'espace fondamental est $\Omega = \Omega_1 \times \Omega_2$, et la mesure de probabilité est P. Comment définir P, si les *expériences sont indépendantes*, dans le sens que si l'une est effectuée son résultat n'est pas modifié par le fait que l'autre soit effectuée ou non ? Soit $A \in \mathcal{F}_1$; les événements

$$A \times \{\eta_k\} \subset \Omega\,, \quad \text{où } \eta_k \text{ parcourt } \Omega_2,$$

sont mutuellement disjoints et leur réunion est l'ensemble $A \times \Omega_2$. Les sous-ensembles $A \subset \Omega_1$ et $A \times \Omega_2 \subset \Omega$ représentent le même événement. Si l'événement A est réalisé, les événements $A \times \{\eta\}$ et $A \times \{\eta'\}$ sont également probables

puisque, par hypothèse, leurs réalisations ne dépendent que des réalisations des événements élémentaires $\{\eta\}$ et $\{\eta'\}$ de la deuxième expérience. Par conséquent, pour tout $\{\eta\}$,

$$P_1(A) = P(A \times \Omega_2) = \sum_{\eta' \in \Omega_2} P(A \times \{\eta'\}) \implies P(A \times \{\eta\}) = \frac{P_1(A)}{n_2}\,.$$

Si $B \in \mathcal{F}_2$,

$$P(A \times B) = \sum_{\eta' \in B} P(A \times \{\eta'\}) = P_1(A)P_2(B)\,.$$

Cette identité définit complètement la mesure de probabilité P. □

Dans l'exemple 4.7 l'indépendance des expériences est en relation avec une propriété de produit de l'espace fondamental Ω et de la mesure de probabilité P : $P(A \times B) = P_1(A)P_2(B)$ pour tout $A \in \mathcal{F}_1$ et $B \in \mathcal{F}_2$.

Définition 4.2 *Soit $(\Omega, \mathcal{F}, P)$ un espace de probabilité. Deux* événements *A et B sont* indépendants pour la mesure de probabilité *P si et seulement si* $P(A \cap B) = P(A)P(B)$.

Remarques 4.2 a) Lorsque la mesure de probabilité P est *fixée et connue*, on dit simplement que A et B sont indépendants, sinon on dit que A et B sont *indépendants sous* P.

b) Si $P(B) > 0$, A et B sont indépendants si et seulement si $P(A|B) = P(A)$.

c) Si $P(A) = 0$, respectivement $P(A) = 1$, alors pour tout B, $P(A \cap B) = P(A)P(B)$; les événements A et B sont indépendants.

d) Des événements disjoints *ne sont pas* indépendants (sauf dans les cas de la remarque précédente).

e) Les événements A et B sont indépendants si et seulement si A^c et B, ou A^c et B^c, ou A et B^c sont indépendants. *L'indépendance de A et B* est en fait *une affirmation sur les algèbres de Boole* $\mathcal{A}_A := \{\Omega, \emptyset, A, A^c\}$ et $\mathcal{A}_B := \{\Omega, \emptyset, B, B^c\}$ qui sont engendrées par les partitions $\{A, A^c\}$ et $\{B, B^c\}$.

f) Soit $0 < P(B) < 1$; les événements A et B sont indépendants si et seulement si $P(A|B) = P(A|B^c)$. En effet, si les événements sont indépendants c'est évident. Inversement, si $P(A|B) = P(A|B^c)$, par la formule des probabilités totales

$$\begin{aligned} P(A)P(B) &= \big(P(A|B)P(B) + P(A|B^c)P(B^c)\big)P(B) \\ &= P(A|B)P(B) = P(A \cap B)\,. \end{aligned}$$

□

La remarque e) suggère la définition suivante.

Définition 4.3 *a)* k algèbres de Boole $\mathcal{F}_1, \ldots, \mathcal{F}_k$ *sont* indépendantes sous P *si et seulement si*

$$P(E_1 \cdots E_k) = P(E_1) \cdots P(E_k) \quad \forall\, E_1 \in \mathcal{F}_1, \ldots, \forall\, E_k \in \mathcal{F}_k\,.$$

b) Même définition pour des σ*-algèbres de Boole.*
c) k événements B_j, $j = 1, \ldots, k$, *sont* indépendants sous P *si et seulement si les* k *sous-algèbres* $\mathcal{A}_{B_j}$ *sont indépendantes.*

Lemme 4.1 *Une partition* $\{A_1, \ldots, A_k\}$ *de* Ω *engendre une algèbre de Boole finie* $\mathcal{A}$ *qui est*

$$\mathcal{A} = \Big\{ B_J \subset \Omega \colon \exists J \subset \{1, \ldots, k\}\,,\ B_J = \bigcup_{i \in J} A_i \Big\}\,.$$

Inversement, chaque algèbre de Boole $\mathcal{A}$ *qui est finie est engendrée par une partition dont les éléments sont dans* $\mathcal{A}$*. Les éléments de cette partition sont appelés les* atomes de l'algèbre $\mathcal{A}$*.*

Preuve Soit une partition $\{A_1, \ldots, A_k\}$. La collection $\mathcal{A}$ donnée dans le lemme est une algèbre de Boole (on fait la convention que si $J = \emptyset$, alors $B_J = \emptyset$). En effet, $B_J^c = B_{J^c}$ et $B_{J_1} \cup B_{J_2} = B_{J_1 \cup J_2}$. N'importe quelle algèbre de Boole contenant les A_i contient aussi leurs unions et donc contient $\mathcal{A}$. L'algèbre $\mathcal{A}$ est donc la plus petite algèbre qui contient chaque élément de la partition.

Soit $\mathcal{A}$ une algèbre de Boole qui est finie. On énumère tous ses éléments $B_1, B_2, \ldots, B_n$. On définit

$$B_i^1 := B_i \quad \text{et} \quad B_i^{-1} := B_i^c$$

de sorte que pour tout $\omega \in \Omega$ soit $\omega \in B_i^1$, soit $\omega \in B_i^{-1}$. On définit aussi pour tout $b = (b_1, \ldots, b_n)$, $b_i = \pm 1$,

$$C^b := \cap_{i=1}^n B_i^{b_i}\,.$$

(On peut avoir $C^b = \emptyset$). Par définition des C^b, pour tout ω il existe b tel que $\omega \in C^b$. D'autre part, $C^b \cap C^{b'} = \emptyset$ si $b \neq b'$; en effet, dans ce cas il existe j tel que $b_j = 1$ et $b'_j = -1$ ou vice-versa ; $b_j = 1$ et $b'_j = -1$ impliquent $C^b \subset A_j$ et $C^{b'} \subset A_j^c$. Par conséquent les $C^b \neq \emptyset$ forment une partition de Ω qui engendre $\mathcal{A}$ et

$$B_i = \bigcup_{b\colon\, b_i = 1} C^b\,.$$

□

Proposition 4.2 *Soit trois algèbres de Boole, $\mathcal{A}$ engendrée par la partition $\{A_1, \ldots, A_r\}$, $\mathcal{B}$ par la partition $\{B_1, \ldots, B_s\}$ et $\mathcal{C}$ par la partition $\{C_1, \ldots, C_t\}$. Alors les algèbres $\mathcal{A}$, $\mathcal{B}$ et $\mathcal{C}$ sont indépendantes sous P si et seulement si*

$$P(A_i B_j C_k) = P(A_i)P(B_j)P(C_k)$$

pour tout $i \in \{1, \ldots, r\}$, pour tout $j \in \{1, \ldots, s\}$ et pour tout $k \in \{1, \ldots, t\}$.

Preuve C'est clairement nécessaire. Un élément de $\mathcal{A}$ s'écrit comme une union disjointe : il existe $I \subset \{1, \ldots, r\}$ tel que $A = \cup_{i \in I} A_i$; de même pour $B \in \mathcal{B}$ et $C \in \mathcal{C}$.

$$A \cap B \cap C = \Big(\bigcup_{i \in I} A_i\Big) \cap \Big(\bigcup_{j \in J} B_j\Big) \cap \Big(\bigcup_{k \in K} C_k\Big) = \bigcup_{i \in I} \bigcup_{j \in J} \bigcup_{k \in K} A_i \cap B_j \cap C_k \,.$$

$A \cap B \cap C$ est exprimé comme une union disjointe et par conséquent

$$\begin{aligned} P(A \cap B \cap C) &= \sum_{i \in I} \sum_{j \in J} \sum_{k \in K} P(A_i \cap B_j \cap C_k) \\ &= \sum_{i \in I} \sum_{j \in J} \sum_{k \in K} P(A_i)P(B_j)P(C_k) \\ &= P(A)P(B)P(C) \,. \end{aligned}$$

□

Proposition 4.3 *Pour que k événements B_i, $i = 1, \ldots, k$, soient indépendants sous P il faut et il suffit que pour toute sous-famille $\{B_{j_1}, \ldots, B_{j_m}\}$ de $\{B_1, \ldots, B_k\}$*

$$P(B_{j_1} \cap \cdots \cap B_{j_m}) = P(B_{j_1}) \cdots P(B_{j_m}) \,.$$

Ces conditions sont équivalentes aux conditions : pour toute sous-famille $\{B_{j_1}, \ldots, B_{j_m}\}$ de $\{B_1, \ldots, B_k\}$ telle que $P(B_{j_1} \cdots B_{j_m}) > 0$,

$$P(B_i | B_{j_1} \cdots B_{j_m}) = P(B_i) \quad \forall\, i \notin \{j_1, \ldots, j_m\} \,.$$

Preuve Les premières conditions sont clairement nécessaires. Soit $\mathcal{A}_{B_i} := \{\emptyset, B_i, B_i^c, \Omega\}$; si ces conditions sont vérifiées, il faut vérifier pour tout $C_i \in \mathcal{A}_{B_i}$, $i = 1, \ldots, k$,

$$P(C_1 \cap \cdots \cap C_k) = P(C_1) \cdots P(C_k) \,.$$

Il suffit de montrer que les conditions postulées sont encore vraies si l'on remplace certains des B_j par B_j^c. Par exemple,

$$\begin{aligned} P(B_{j_1}^c B_{j_2} \cdots B_{j_m}) &= P(B_{j_2} \cdots B_{j_m}) - P(B_{j_1} B_{j_2} \cdots B_{j_m}) \\ &= (1 - P(B_{j_1})) \prod_{n=2}^{m} P(B_{j_n}) \,. \end{aligned}$$

La preuve de l'équivalence des deuxièmes conditions avec les premières est laissée au lecteur (utiliser la formule de multiplication). □

Exemple 4.8 Deux dés équilibrés, un blanc et un noir, sont lancés (équiprobabilité des résultats). Soit les événements A « la somme des dés vaut 7 », B « le dé blanc donne 4 » et C « le dé noir donne 3 ». On a

$$P(AB) = P(A)P(B) \quad P(AC) = P(A)P(C) \quad P(BC) = P(B)P(C)\,.$$

Mais

$$P(A(BC)) = \underbrace{P(A|BC)}_{1} P(BC) = \frac{1}{36} \neq \frac{1}{6^3}\,.$$

Les événements A, B, C ne sont pas indépendants.

Exemple 4.9 On considère des tirages ordonnés avec remise de n jetons d'une urne contenant les jetons $a_1, \ldots, a_M$. On pose $\mathtt{A} = \{a_1, \ldots, a_M\}$. Chaque tirage est également probable. L'espace fondamental est Ω_a de la section 3.1. Grâce à la notion d'indépendance on peut donner une description plus simple de cette expérience. On introduit les applications $X_j \colon \Omega_a \to \mathtt{A}$,

$$X_j(\omega) := \omega_j \quad (\text{le } j^{\text{ième}} \text{ jeton tiré est } \omega_j)\,.$$

Les événements $\{X_1 = b_1\}$, …, $\{X_n = b_n\}$, quels que soient les $b_j \in \mathtt{A}$, sont indépendants,

$$P\big(\{X_1 = b_1\} \cdots \{X_n = b_n\}\big) = \prod_{i=1}^{n} P\big(\{X_i = b_i\}\big)\,.$$

De même, les partitions

$$Q_i := \Big\{\{X_i = b\} \colon b \in \mathtt{A}\Big\}\,,\ i = 1, \ldots, n\,,$$

sont indépendantes. Par exemple, la partition Q_3 correspond à la question à choix multiple "quel est le jeton tiré lors du troisième tirage". L'expérience aléatoire considérée est équivalente à répéter n expériences indépendantes (voir exemple 4.7), chacune consistant simplement à tirer un jeton de l'urne.

4.3 Exercices

Exercice 4.1 A une interrogation on pose une question à choix multiple avec m possibilités de réponse, une seule étant correcte. Si l'étudiant ne connaît pas la réponse il donne une réponse au hasard. Sachant que 65% des étudiants ne connaissent pas la réponse, si l'on interroge un étudiant au hasard et qu'il donne la bonne réponse, quelle est la probabilité qu'il connaissait la réponse ?

Exercice 4.2 On considère deux événements A et B. Ne sachant rien sur la probabilité de l'événement A, on choisit comme probabilité a priori $P(A) = 1/2$. Par contre on sait que

$$P(B|A) = 0{,}1 \quad \text{et} \quad P(B|A^c) = 0{,}99\,.$$

Si l'on observe l'événement B, quelle est la probabilité a posteriori de l'événement A ?

Exercice 4.3 On considère deux pièces de monnaie équilibrées qu'on lance successivement ; les résultats de l'expérience aléatoire sont équiprobables. A est l'événement « la première pièce donne Face », B « la deuxième pièce donne Face », et C « une et une seule des pièces donne Face ». Les événements A, B et C sont-ils indépendants ?

Exercice 4.4 Trois joueurs a, b et c jouent tour à tour un jeu pour lequel il n'y a pas de partie nulle. On procède de la façon suivante : lors de la première partie c ne joue pas ; la partie suivante est jouée entre le gagnant de la partie et la personne qui n'a pas joué la partie précédente et ainsi de suite. On suppose que chaque participant gagne sa partie avec probabilité $1/2$. Les parties s'arrêtent dès qu'un des joueurs gagne successivement deux fois.

a) Modéliser cette expérience aléatoire.
Indication : on peut décrire les résultats possibles de cette expérience aléatoire en indiquant le gagnant de chaque partie.

b) Montrer qu'on ne joue pas indéfiniment. Calculer les probabilités que les parties s'arrêtent parce que a a gagné deux fois de suite, respectivement b a gagné deux fois de suite, respectivement c a gagné deux fois de suite.

Exercice 4.5 On fait le jeu suivant. On a trois boîtes 1, 2 et 3. La boîte 1 contient 100 francs et les deux autres sont vides. Le jeu se joue en trois temps.
a) A qui ne connaît pas le contenu des boîtes choisit au hasard une des boîtes et ne l'ouvre pas.
b) B qui connaît le contenu des boîtes ouvre intentionnellement une boîte vide parmi les boîtes restantes.
c) A a la possibilité d'échanger sa boîte choisie avec celle qui reste. On considère les trois stratégies suivantes pour ce deuxième choix de A :
1) A garde toujours la boîte choisie initialement ;
2) A échange toujours la boîte choisie initialement avec la boîte restante ;
3) A joue à Pile ou Face et échange sa boîte s'il obtient Face.

Analyser ce jeu en utilisant des diagrammes en arbre. Soit

$$G_j := \{\text{au début } A \text{ choisit la boîte } j\}$$

$$F := \{B \text{ ouvre une boîte vide}\} \quad \text{et} \quad F_2 := \{B \text{ ouvre la boîte } 2\}\,.$$

Calculer la probabilité de F et de F_2. De même calculer les probabilités conditionnelles de F sachant G_1 et de F_2 sachant G_1.

Est-ce que G est indépendant de F, respectivement de F_2 ?
Calculer la probabilité de l'événement « A gagne les 100 francs » pour les différentes stratégies.

Exercice 4.6 On examine une autre version du jeu présenté dans l'exercice 4.5. A la deuxième étape B ouvre, au hasard, une des deux boîtes restantes. Sinon le jeu et les statégies pour le deuxième choix de A sont les mêmes qu'avant. Répondre aux mêmes questions que celles de l'exercice 4.5.

Exercice 4.7 On considère un test pour détecter un virus (exemple 4.3). L'événement E est « le test est positif » et l'événement V « la personne est porteuse du virus ». Les propriétés statistiques de ce test $P(E|V)$ et $P(E|V^c)$ sont connues. Pour détecter si une personne est porteuse du virus, on fait le même test dans deux laboratoires. On pose $E_i := \{\text{test fait dans le labo } i \text{ est positif}\}$, $i = 1, 2$. Les deux laboratoires travaillent de façon indépendante, ce qui peut s'exprimer ainsi.
1) *Etant donné la personne testée et le résultat du laboratoire* 1*, le résultat du laboratoire* 2 *ne dépend que de la personne testée*,

$$P(E_2|E_1V) = P(E_2|V) \quad \text{et} \quad P(E_2|E_1V^c) = P(E_2|V^c)\,.$$

2) *Etant donné la personne testée, les résultats des laboratoires sont indépendants*,

$$P(E_2E_1|V) = P(E_2|V)P(E_1|V) \quad \text{et} \quad P(E_2E_1|V^c) = P(E_2|V^c)P(E_1|V^c)\,.$$

Montrer que 1) est équivalent à 2).

Exercice 4.8 Suite de l'exercice 4.7. On suppose que $P(E|V) = 0{,}99$ et $P(E|V^c) = 0{,}001$ et que la proportion des porteurs du virus dans la population est de 0,1%. On choisit une personne au hasard.
a) Calculer la probabilité de détecter le virus si les deux tests sont positifs.
b) Est-ce que les événements $E_i := \{\text{le test fait dans le labo } i \text{ est positif}\}$, $i = 1, 2$, sont indépendants ?

Exercice 4.9 On considère l'espace de probabilité discret $\Omega = \{1, 2, \ldots, 100\}$ avec la mesure de probabilité uniforme.
a) Donner un exemple de trois événements non indépendants A, B et C tels que $P(ABC) = P(A)P(B)P(C)$.
b) Donner un exemple de trois événements non indépendants A, B et C tels que $P(AB) = P(A)P(B)$, $P(AC) = P(A)P(C)$ et $P(CB) = P(C)P(B)$.

Exercice 4.10 Ω est un ensemble, $\mathcal{A} \subset \mathcal{P}(\Omega)$ et $\mathcal{B} \subset \mathcal{P}(\Omega)$ sont deux algèbres de Boole.
a) Est-ce que la famille des sous-ensembles $\mathcal{A} \cup \mathcal{B}$ est une algèbre de Boole ?
b) Si $\mathcal{A}_n \subset \mathcal{P}(\Omega)$ sont des algèbres de Boole et si $\mathcal{A}_n \subset \mathcal{A}_{n+1}$ pour tout n, est-ce que la collection des sous-ensembles, qui est formée par l'union des $\mathcal{A}_n$, est une algèbre de Boole ?

Chapitre 5

Espaces de probabilité sur $\mathbb{R}$ et $\mathbb{R}^k$

Contrairement au cas des espaces de probabilité discrets, pour lesquels une description complète a été donnée, on ne considère ici que les cas $\Omega = \mathbb{R}$, $\Omega = I$, un intervalle, ou $\Omega = \mathbb{R}^k$. Plusieurs aspects mathématiques sont passés sous silence dans le cadre de ce livre. Cependant, si l'on s'intéresse principalement à des questions de nature non théorique, cela ne constitue pas un handicap majeur. Grâce à la notion de variable aléatoire introduite au chapitre 6, les exemples d'espace de probabilité traités dans ce chapitre suffisent pour exposer une part substantielle de la théorie. Certains points théoriques importants sont exposés sans démonstration.

L'exemple suivant est un exemple générique de construction d'un espace de probabilité non discret.

Exemple 5.1 Un *générateur de nombres aléatoires* (GNA) est un procédé qui sélectionne au hasard (i.e. de manière également probable) un nombre réel dans $(0, 1)$. On décrit ici l'espace de probabilité $(\Omega, \mathcal{F}, P)$ associé à un GNA. Le modèle mathématique d'un GNA (idéal) est clair ; du point de vue pratique la construction d'un GNA est délicate.

L'espace fondamental est $\Omega = (0, 1)$ puisque le résultat est un $t \in (0, 1)$. Une mesure de probabilité est définie sur la collection des événements. Un événement simple pour lequel on a une bonne intuition est un intervalle $I = (a, b)$: « le nombre sélectionné par le GNA est dans I ». Intuitivement la condition d'équiprobabilité se traduit par

$$P(I) = \text{longueur de l'intervalle } I \equiv \ell(I) = \int_a^b dt\,. \tag{5.1}$$

La propriété d'additivité de P pour une famille finie d'intervalles disjoints est immédiate. Pour compléter la définition de l'espace de probabilité on est confronté au problème suivant :

Peut-on trouver une σ-algèbre $\mathcal{F}$ qui contient tous les intervalles et une mesure de probabilité P sur $\mathcal{F}$ telle que

$$P(I) = \ell(I) = \int_a^b dt \quad \textit{si } I \textit{ est un intervalle ?}$$

Il s'agit d'étendre l'application P, qui est définie sur les intervalles de façon naturelle, à une σ-algèbre de sorte que P soit σ-additive. Ce problème a été résolu par H. Lebesgue (1875-1941) dans sa thèse de doctorat intitulée *Intégrale, longueur, aire* (1902).

1. Il existe une (plus petite) σ-algèbre contenant tous les intervalles, notée $\mathcal{B}((0,1))$. Cette collection de sous-ensembles de $(0,1)$ est celle des *ensembles boréliens*. Elle ne contient pas tous les sous-ensembles de $(0,1)$, mais elle est suffisamment riche pour toutes les situations concrètes rencontrées dans la pratique. Elle contient par exemple tous les sous-ensembles ouverts et fermés.
2. Pour chaque E de la collection $\mathcal{B}((0,1))$ on peut définir la longueur

$$\ell(E) = \int_E dt$$

qui a la propriété de σ-additivité. Le résultat principal est qu'*il n'y a qu'une et une seule façon* de définir $\ell(E)$ si la longueur des intervalles est donnée par (5.1).

On prend donc pour modèle mathématique d'un GNA l'espace de probabilité $(\Omega, \mathcal{F}, P)$ avec $\Omega = (0,1)$, $\mathcal{F} = \mathcal{B}((0,1))$ et $P(E) = \ell(E)$. □

Remarque 5.1 Définir la longueur $\ell(E)$ pour $E \in \mathcal{B}((0,1))$ revient à définir l'intégrale $\int_E dt$ (intégrale de Lebesgue de E). Si $\int_E dt$ existe comme intégrale de Riemann, alors on montre que $E \in \mathcal{B}((0,1))$ et que l'intégrale de Riemann est égale à $\ell(E)$. Dans tous les cas concrets de ce livre on est dans cette situation. Cette remarque, ainsi que l'existence et l'unicité de l'espace de probabilité pour un GNA, font que dans ce livre on n'a pas besoin de connaître d'autres propriétés de la σ-algèbre $\mathcal{B}((0,1))$.

Remarque 5.2 Dans la modélisation mathématique d'un GNA on admet que le résultat de l'expérience est un nombre réel déterminé t. Par contre, si $t \in (0,1)$, l'événement « on observe le résultat t » correspond au sous-ensemble

$$\{t\} = \bigcap_{n \geq n_0} \Big(t - \frac{1}{n}, t + \frac{1}{n}\Big) \subset (0,1)\,.$$

(n_0 est suffisamment grand). La propriété de continuité monotone séquentielle de P implique que $P(\{t\}) = 0$. L'interprétation de ce résultat est la suivante. Pour *observer* un nombre t donné, i.e. l'événement $\{t\}$, il faut connaître toutes les décimales de t, ce qui requiert une précision infinie qu'on ne peut pas atteindre expérimentalement. $P(\{t\}) = 0$ signifie que cet événement n'est pas observable. Par contraste, si $\varepsilon > 0$, l'événement $(t-\varepsilon, t+\varepsilon)$ est observable avec une précision finie, et

$$P\big((t-\varepsilon, t+\varepsilon)\big) = \ell\big((t-\varepsilon, t+\varepsilon) \cap (0,1)\big) > 0\,.$$

Il est important de distinguer $t \in (0,1)$ et $\{t\} \subset (0,1)$.

5.1 Mesure de probabilité sur $\mathbb{R}$

On considère le cas d'un espace de probabilité dont l'espace fondamental $\Omega = \mathbb{R}$. On introduit comme dans l'exemple 5.1 la σ-algèbre des ensembles boréliens $\mathcal{F} := \mathcal{B}(\mathbb{R})$; $\mathcal{B}(\mathbb{R})$ est la plus petite σ-algèbre contenant tous les intervalles $(a, b]$. Tous les événements $E \subset \mathbb{R}$ considérés dans ce livre sont clairement dans cette σ-algèbre. Les ensembles $(-\infty, t] \in \mathcal{B}(\mathbb{R})$, quel que soit $t \in \mathbb{R}$. Si P est une mesure de probabilité sur $\mathcal{F}$, la *fonction de répartition de* P est, par définition, la fonction définie sur $\mathbb{R}$ par

$$t \mapsto F(t) := P((-\infty, t]) .$$

Proposition 5.1 *Soit F la fonction de répartition de P.*

1) $0 \leq F(t) \leq 1$, F est monotone non décroissante.

2) $\lim_{t \to -\infty} F(t) = 0$ et $\lim_{t \to \infty} F(t) = 1$.

3) F est continue à droite, i.e. pour tout $t \in \mathbb{R}$ $\lim_{s \downarrow t} F(s) = F(t)$.

4) $P(\{x\}) = F(x) - \lim_{t \uparrow x} F(t)$; F a un saut *en x si et seulement si $P(\{x\}) > 0$. $P(\{x\})$ est la* hauteur du saut *en x.*

Preuve Le point 1) suit du lemme 2.1, et le point 3) des propriétés de continuité monotone. Si $s_n > t$ et $s_n \downarrow t$,

$$(-\infty, t] = \bigcap_n (-\infty, s_n] .$$

(Si $x > t$, il existe n tel que $t \leq s_n < x$ de sorte que $x \notin (-\infty, s_n]$). Par continuité monotone de P

$$F(t) = P((-\infty, t]) = \lim_{s_n \downarrow t} P((-\infty, s_n]) = \lim_{s_n \downarrow t} F(s_n) .$$

Le point 2) est traité de façon similaire à partir de

$$(-\infty, t_n] \downarrow \emptyset \quad \text{si } t_n \downarrow -\infty \quad \text{et} \quad (-\infty, t_n] \uparrow \mathbb{R} \quad \text{si } t_n \uparrow \infty .$$

Pour le point 4), soit $t_n < x$ et $t_n \uparrow x$; on a $\{x\} = \bigcap_n (t_n, x]$ et

$$\begin{aligned} P(\{x\}) &= \lim_n P((t_n, x]) \\ &= \lim_n (P((-\infty, x]) - P((-\infty, t_n])) = F(x) - \lim_n F(t_n) . \end{aligned}$$

Par conséquent, F n'est pas continue à gauche en x si et seulement si elle fait un saut en x de hauteur $P(\{x\}) > 0$. □

Le théorème 5.1 est un des théorèmes importants de la théorie (voir par exemple A. A. Kirillov, A. D. Gvishiani, *Theorems and Problems in Functional*

Analysis[1] chapitre II, en particulier le théorème 4). Il permet de définir *toutes les mesures de probabilité sur* $\mathbb{R}$.

Définition 5.1 *Une* fonction de répartition *est une fonction définie sur* $\mathbb{R}$ *à valeur dans* $[0,1]$*, vérifiant les propriétés 1, 2 et 3 de la proposition 5.1.*

Théorème 5.1 *Soit* $F\colon \mathbb{R} \to [0,1]$ *une fonction de répartition. Alors* F définit de façon unique *sur* $\mathcal{B}(\mathbb{R})$ *une mesure de probabilité telle que pour tout intervalle* $(a,b]$

$$P((a,b]) = F(b) - F(a)\,.$$

Dans ce livre on n'a besoin que des deux cas suivants qu'on nomme respectivement le cas continu et le cas discret.

1) On suppose que la fonction de répartition F est dérivable et que sa dérivée $f := F'$ vérifie pour tout a et b

$$F(b) - F(a) = \int_a^b f(t)\,dt\,.$$

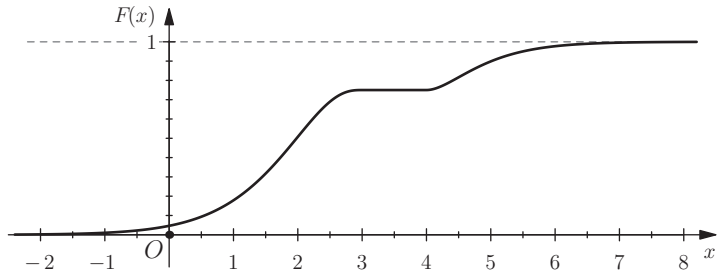


Figure 5.1 – Fonction de répartition continue.

En particulier, si $f = F'$ est continue, cette identité est vraie. La mesure de probabilité définie par F est appelée *mesure de probabilité continue* ; pour tout $A \in \mathcal{B}(\mathbb{R})$ (voir remarque 5.1)

$$P(A) = \int_A f(t)\,dt\,. \tag{5.2}$$

Inversement, si $f : \mathbb{R} \to \mathbb{R}^+$ vérifie $\int_{\mathbb{R}} f(t)\,dt = 1$, alors la fonction

$$t \mapsto F(t) := \int_{-\infty}^{t} f(s)\,ds$$

1. Comme c'est souvent le cas dans la littérature russe, la fonction F dans cette référence est définie par $F(t) := P\big((-\infty,t)\big)$ à la place de $F(t) = P\big((-\infty,t]\big)$; par conséquent cette fonction est continue à gauche à la place d'être continue à droite.

est une fonction de répartition. Par conséquent f définit une mesure de probabilité P telle que (5.2) est vérifié. La fonction f est appelée *densité de probabilité* pour la raison suivante : (si f continue)

$$\lim_{\varepsilon\downarrow 0} \frac{P\big((t-\varepsilon, t+\varepsilon)\big)}{2\varepsilon} = \lim_{\varepsilon\downarrow 0} \frac{1}{2\varepsilon} \int_{t-\varepsilon}^{t+\varepsilon} f(s)\, ds = f(t)\,.$$

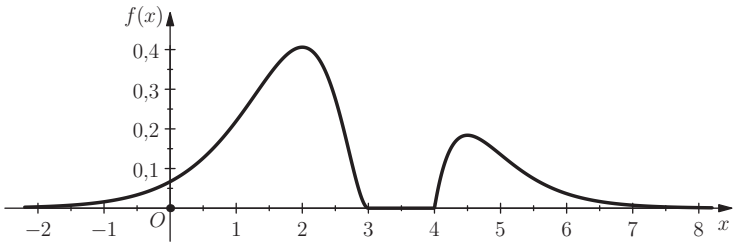


Figure 5.2 – Densité de probabilité de la fonction de répartition de la figure 5.1.

En résumé, on spécifie une mesure de probabilité continue en donnant F ou en donnant $f = F'$ qui est la densité de probabilité de la mesure.

2) On suppose que la fonction de répartition F est constante par morceaux. L'ensemble D des sauts de F est fini ou dénombrable. En effet, pour tout $m \geq 1$ il y a au plus m sauts de hauteur plus grande que $1/m$ puisque $F(\infty) = 1$; par conséquent la proposition I.1 implique que D est fini ou dénombrable. Si $a \in D$, la hauteur du saut en a est

$$q(a) := F(a) - \lim_{s\uparrow a} F(s)\,.$$

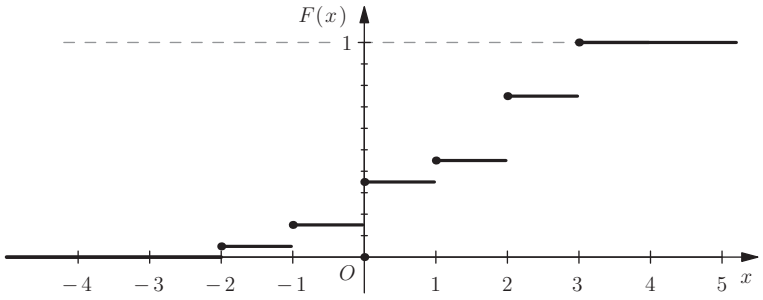


Figure 5.3 – Fonction de répartition discrète.

La mesure de probabilité associée est appelée *mesure de probabilité discrète* ; elle est donnée par

$$P(A) = \sum_{a\in A\cap D} q(a)\,.$$

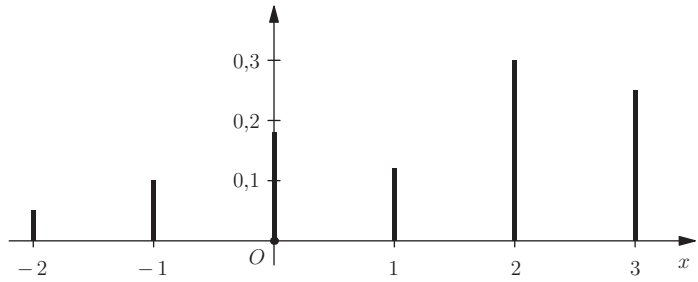


Figure 5.4 – Mesure de probabilité discrète spécifiée par la position et la hauteur des sauts de la fonction de répartition de la figure 5.3.

Par définition $q(a) > 0$ et $\sum_{a\in D} q(a) = 1$. La mesure de probabilité est entièrement déterminée par la donnée de l'ensemble dénombrable D et de la fonction $q\colon D \to [0,1]$ telle que $\sum_{a\in D} q(a) = 1$. C'est essentiellement la définition d'une mesure de probabilité sur un espace de probabilité discret (sect. 2.3).

Exemple 5.2 Soit $\lambda > 0$. Une *mesure de probabilité de Poisson (1781-1840) sur* $\mathbb{R}$, *de paramètre* λ, est une mesure de probabilité discrète notée π_λ, telle que l'ensemble des sauts $D = \{0, 1, 2, \ldots\}$ et

$$\pi_\lambda(k) := \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}\,, \quad k \in D\,.$$

Voir figures 6.2 et 6.3.

Exemple 5.3 Soit $m \in \mathbb{R}$ et $0 < \sigma^2 < \infty$. Une *mesure de probabilité gaussienne sur* $\mathbb{R}$, *de paramètres* m *et* σ^2, est donnée par la densité de probabilité

$$f_{m,\sigma^2}(t) := \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{(t-m)^2}{2\sigma^2}\Big)\,.$$

Cette mesure est notée $N(m, \sigma^2)$.

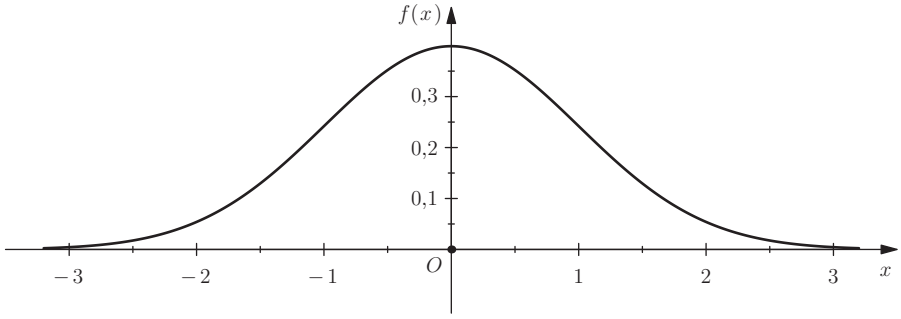


Figure 5.5 – Densité de probabilité de la mesure gaussienne $N(0, 1)$.

Exemple 5.4 Soit $x > 0$ et $\lambda > 0$. Une *mesure de probabilité gamma, de paramètres* (x, λ), est donnée par la densité de probabilité notée $\gamma_{x,\lambda}$,

$$\gamma_{x,\lambda}(t) := \Gamma(x)^{-1}\lambda \mathrm{e}^{-\lambda t}(\lambda t)^{x-1} \quad \text{si } t > 0,\ \gamma_{x,\lambda}(t) = 0 \text{ sinon.}$$

$\Gamma(x)$ est la fonction Gamma évaluée en x. □

Remarque 5.3 Par définition, si $x > 0$, la *fonction Gamma* est

$$\Gamma(x) := \int_0^\infty \mathrm{e}^{-t}\, t^{x-1}\, dt\,.$$

Par intégration par parties on obtient la relation fondamentale

$$\Gamma(x+1) = x\Gamma(x)\,.$$

$\Gamma(1) = 1$ et par récurrence $n! = \Gamma(n+1)$. Par le changement de variable $t \to u^2$ on obtient l'*intégrale de Gauss* (1777-1855)

$$\Gamma(x) = \int_0^\infty \mathrm{e}^{-u^2}\, u^{2(x-1)} 2u\, du = 2\int_0^\infty \mathrm{e}^{-u^2}\, u^{2x-1}\, du\,.$$

Lorsque $x = 1/2$ (voir (5.6)),

$$\Gamma(1/2) = 2\int_0^\infty \mathrm{e}^{-t^2}\, dt = \int_{-\infty}^\infty \mathrm{e}^{-t^2}\, dt = \sqrt{\pi}\,.$$

Soit $p > 0$ et $q > 0$; en passant aux coordonnées polaires

$$\begin{aligned}\Gamma(p)\Gamma(q) &= 4\int_0^\infty\int_0^\infty \mathrm{e}^{-(u^2+v^2)} u^{2p-1} v^{2q-1}\, du dv \\ &= \underbrace{2\int_0^\infty \mathrm{e}^{-r^2} r^{2(p+q)-1}\, dr}_{=\Gamma(p+q)} \cdot \underbrace{2\int_0^{\pi/2} \cos^{2p-1}\theta \sin^{2q-1}\theta\, d\theta}_{\equiv B(p,q)}\,.\end{aligned}$$

En posant $\cos^2\theta = t$ on obtient pour $B(p,q)$ l'expression

$$B(p,q) = B(q,p) = \int_0^1 t^{p-1}(1-t)^{q-1}\, dt = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}\,. \tag{5.3}$$

□

Les considérations de cette section sont aussi valables pour un intervalle I. Il suffit de considérer des densités de probabilité qui sont nulles en dehors de I ou de supposer que les fonctions de répartition sont nulles à gauche de I et égales à un à droite de I. Dans le cas des mesures de probabilité discrètes l'ensemble des sauts D est un sous-ensemble de I.

5.2 Mesure de probabilité sur $\mathbb{R}^k$

Lorsque $\Omega = \mathbb{R}^k$ la σ-algèbre $\mathcal{F}$ est celle des ensembles boréliens $\mathcal{B}(\mathbb{R}^k)$ qui est engendrée par les produits de k intervalles $I_1 \times \cdots \times I_k$. Dans la suite, dans les situations concrètes, on a besoin seulement des deux cas suivants qui ne nécessitent pas d'explications supplémentaires. Ces deux cas n'épuisent de loin pas les exemples de mesures de probabilité sur $\mathbb{R}^k$.

1) Une *mesure de probabilité discrète sur* $\mathbb{R}^k$ est spécifiée par un ensemble dénombrable $D \subset \mathbb{R}^k$ et une application $q\colon D \to [0,1]$,

$$\sum_{\mathbf{a}\in D} q(\mathbf{a}) = 1\,.$$

La mesure de probabilité μ est définie par

$$\mu(A) := \sum_{\mathbf{a}\colon\, \mathbf{a}\in A\cap D} q(\mathbf{a})\,.$$

2) Une *mesure de probabilité continue sur* $\mathbb{R}^k$ est spécifiée par une densité de probabilité $f\colon \mathbb{R}^k \to \mathbb{R}^+$,

$$\int_{\mathbb{R}^k} f(\mathbf{x})\, d\mathbf{x} = 1\,.$$

La mesure de probabilité μ est définie par

$$\mu(A) := \int_A f(\mathbf{x})\, d\mathbf{x}\,.$$

Dans la formule ci-dessus on a utilisé une notation concise. On écrit aussi $\mathbf{x} = (x_1, \ldots, x_k)$ et $d\mathbf{x} = dx_1 \cdots dx_k$. Avec ces notations

$$\mu(A) = \int \cdots \int_A f(x_1, \ldots, x_k)\, dx_1 \cdots dx_k\,.$$

5.2.1 Mesure gaussienne

Un exemple central de mesure de probabilité sur $\mathbb{R}^k$ est celui d'une mesure gaussienne. Les mesures gaussiennes (non nécessairement normalisées) jouent également un rôle important dans d'autres domaines, en particulier en physique.

Une *mesure gaussienne sur* $\mathbb{R}^k$ est spécifiée par la donnée d'une matrice $\mathbf{A}$ réelle de type $k \times k$, symétrique, définie positive et d'un vecteur $\mathbf{m} \in \mathbb{R}^k$; sa densité est par définition

$$g(\mathbf{x}) \equiv g(x_1, \ldots, x_k) := \exp\Big(-\frac{1}{2}\,\langle(\mathbf{x}-\mathbf{m})|\mathbf{A}(\mathbf{x}-\mathbf{m})\rangle\Big)\,.$$

Cette mesure n'est pas normalisée,

$$\int_{\mathbb{R}^k} g(\mathbf{x})\, d\mathbf{x} = \sqrt{(2\pi)^k \det \mathbf{A}^{-1}}\,. \tag{5.4}$$

Cette identité est un cas particulier de (5.7). Lorsque la mesure est normalisée, on a une *mesure de probabilité gaussienne* $N(\mathbf{m}, \mathbf{A})$; sa densité est

$$f_{\mathbf{m},\mathbf{A}}(\mathbf{x}) := \frac{1}{\sqrt{(2\pi)^k \det \mathbf{A}^{-1}}}\, g(\mathbf{x})\,. \tag{5.5}$$

Un cas simple est celui où $\mathbf{m} = \mathbf{0}$ et $\mathbf{A}$ diagonale, $\mathbf{A}_{ii} = 1/\sigma^2$ pour tout i,

$$f_{\mathbf{0},\mathbf{A}}(x_1, \ldots, x_k) = \frac{1}{(2\pi\sigma^2)^{\frac{k}{2}}} \exp\Big(-\frac{x_1^2 + \cdots + x_k^2}{2\sigma^2}\Big)\,.$$

Exemple 5.5 Calculer

$$I := \int_{\mathbb{R}^2} \exp(-5x^2 - 6y^2 + 4xy)\, dxdy\,.$$

C'est une intégrale gaussienne avec

$$\mathbf{A} = \begin{pmatrix} 10 & -4 \\ -4 & 12 \end{pmatrix}\,.$$

La matrice est symétrique et définie positive puisque $\det\mathbf{A} = 104$. En utilisant (5.4) on obtient

$$I = \frac{2\pi}{\sqrt{\det \mathbf{A}}} = \frac{\pi}{\sqrt{26}}\,.$$

Proposition 5.2 *Soit* $\mathbf{m} \in \mathbb{R}^{\mathbf{k}}$ *et* $\mathbf{A}$ *une matrice réelle de type* $k \times k$*, symétrique et définie positive.*

$$\int_{-\infty}^{\infty} \mathrm{e}^{-t^2}\, dt = \sqrt{\pi}\,. \tag{5.6}$$

$$\int_{\mathbb{R}^k} \exp(\langle \mathbf{z}|\mathbf{x}\rangle) g(\mathbf{x})\, d\mathbf{x} = \frac{\sqrt{(2\pi)^k}}{\sqrt{\det \mathbf{A}}} \exp\Big(\langle \mathbf{z}|\mathbf{m}\rangle + \frac{1}{2}\,\langle \mathbf{z}|\mathbf{A}^{-1}\mathbf{z}\rangle\Big)\,. \tag{5.7}$$

Preuve On établit tout d'abord l'identité (5.6) qui est l'identité fondamentale dans ce contexte. En passant aux coordonnées polaires,

$$\Big(\int_{-\infty}^{\infty} \mathrm{e}^{-t^2}\, dt\Big)^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-t^2-s^2}\, dtds = 2\pi \int_0^{\infty} \mathrm{e}^{-r^2}\, rdr = \pi\,.$$

A partir de (5.6) on dérive (5.7). Soit $\mathbf{U}$ la matrice orthogonale qui diagonalise la matrice $\mathbf{A}$, $\mathbf{D} = \mathbf{U}\,\mathbf{A}\,\mathbf{U}^\top$ avec $\mathbf{D}$ diagonale ; les éléments diagonaux de $\mathbf{D}$ sont les valeurs propres λ_j de $\mathbf{A}$, qui sont toutes strictement positives puisque $\mathbf{A}$ est définie positive. On fait le changement de variables $\mathbf{x} - \mathbf{m} = \mathbf{U}^\top\mathbf{y}$ (jacobien égal à 1) ; en utilisant $\big\langle \mathbf{z}|\mathbf{U}^\top\mathbf{y}\big\rangle = \langle \mathbf{U}\mathbf{z}|\mathbf{y}\rangle$ on obtient

$$\int \exp(\langle \mathbf{z}|\mathbf{x}\rangle) g(\mathbf{x})\, d\mathbf{x} = \exp\big(\langle \mathbf{z}|\mathbf{m}\rangle\big) \int \exp\Big(\,\langle \mathbf{U}\mathbf{z}|\mathbf{y}\rangle - \frac{1}{2}\,\langle \mathbf{y}|\mathbf{D}\mathbf{y}\rangle\,\Big)\, d\mathbf{y}\,.$$

L'intégrale du membre de droite factorise en k intégrales sur $\mathbb{R}$. On pose $\mathbf{u} := (\mathbf{U}\mathbf{z})$, $\mathbf{u} = (u_1, \dots, u_k)$, et on écrit

$$u_j y_j - \frac{1}{2}\lambda_j y_j^2 = -\frac{1}{2}\Big(\sqrt{\lambda_j}y_j - \frac{u_j}{\sqrt{\lambda_j}}\Big)^2 + \frac{u_j^2}{2\lambda_j}\,.$$

Ceci permet de faire le changement de variables

$$v_j := \sqrt{\lambda_j}y_j - \frac{u_j}{\sqrt{\lambda_j}}$$

pour se ramener à l'intégrale (5.6).

$$\begin{aligned}\int_{\mathbb{R}} \exp\Big(u_j y_j - \frac{1}{2}\lambda_j y_j^2\Big)\,dy_j &= \exp\Big(\frac{u_j^2}{2\lambda_j}\Big)\int_{\mathbb{R}} \exp\Big(-\frac{1}{2}\Big(\sqrt{\lambda_j}y_j - \frac{u_j}{\sqrt{\lambda_j}}\Big)^2\Big)\,dy_j \\ &= \exp\Big(\frac{u_j^2}{2\lambda_j}\Big)\frac{1}{\sqrt{\lambda_j}}\int_{\mathbb{R}} \exp\Big(-\frac{1}{2}v_j^2\Big)\,dv_j \\ &= \exp\Big(\frac{u_j^2}{2\lambda_j}\Big)\frac{\sqrt{2\pi}}{\sqrt{\lambda_j}}\,.\end{aligned}$$

Par conséquent

$$\begin{aligned}\int \exp(\langle \mathbf{z}|\mathbf{x}\rangle) g(\mathbf{x})\,d\mathbf{x} &= \frac{(2\pi)^{k/2}}{\sqrt{\det \mathbf{A}}}\exp\Big(\langle \mathbf{z}|\mathbf{m}\rangle + \frac{1}{2}\langle \mathbf{z}|\mathbf{U}^\top \mathbf{D}^{-1}\mathbf{U}\mathbf{z}\rangle\Big) \\ &= \frac{(2\pi)^{k/2}}{\sqrt{\det \mathbf{A}}}\exp\Big(\langle \mathbf{z}|\mathbf{m}\rangle + \frac{1}{2}\langle \mathbf{z}|\mathbf{A}^{-1}\mathbf{z}\rangle\Big) \\ &= \sqrt{(2\pi)^k \det \mathbf{A}^{-1}}\,\exp\Big(\langle \mathbf{z}|\mathbf{m}\rangle + \frac{1}{2}\langle \mathbf{z}|\mathbf{A}^{-1}\mathbf{z}\rangle\Big)\,.\end{aligned}$$

□

Soit $i_1, \dots, i_m$ une famille de m indices, $i_j \in \{1, \dots, k\}$. Par exemple, si $k = 3$ et $m = 6$, $i_1 = i_2 = i_3 = i_4 = 1$ et $i_5 = i_6 = 2$. On rappelle (voir exemple 3.3) qu'un appariement de $2n$ objets est une partition de ces objets en n paires (non ordonnées), chaque objet figurant dans une et une seule paire. Dans l'exemple ci-dessus le nombre d'appariements est donné par (3.1), soit 15. La proposition suivante est appelée *théorème de Wick* dans la littérature physique.

Proposition 5.3 *Soit $i_1, \dots, i_n$ une famille de n indices, $i_j \in \{1, \dots, k\}$.*
Si $n = 2m - 1$, $m \geq 1$,

$$\int_{\mathbb{R}^k} f_{\mathbf{0},\mathbf{A}}(x_1, \dots, x_k)\Big(\prod_{j=1}^{2m-1} x_{i_j}\Big)\,dx_1 \cdots dx_k = 0\,.$$

Si $n = 2m$, $m \geq 1$,

$$\int_{\mathbb{R}^k} f_{\mathbf{0},\mathbf{A}}(x_1, \dots, x_k)\Big(\prod_{j=1}^{2m} x_{i_j}\Big)\,dx_1 \cdots dx_k = \sum_{\substack{\text{appariements} \\ \text{de } i_1, \dots, i_{2m}}} \prod_{\substack{\text{chaque} \\ \text{paire}}} \mathbf{A}^{-1}_{i_a i_b}\,.$$

Preuve Par symétrie, en faisant le changement de variables $\mathbf{x}$ en $-\mathbf{x}$, on obtient le cas n impair. Si n est pair, l'identité (5.7) s'écrit ici

$$\int_{\mathbb{R}^k} f_{\mathbf{0},\mathbf{A}} \exp\Big(\sum_i z_i x_i\Big)\, d\mathbf{x} = \exp\Big(\frac{1}{2}\sum_{i,j} z_i \mathbf{A}^{-1}_{ij} z_j\Big)\,.$$

On fait agir sur cette identité les opérateurs différentiels

$$\frac{\partial}{\partial z_{i_1}}, \dots, \frac{\partial}{\partial z_{i_{2m}}}\,,$$

i.e. on dérive cette identité par rapport à $z_{i_1}, \dots, z_{i_{2m}}$, puis on pose $z_i = 0$ pour tout i. Le côté gauche donne

$$\int_{\mathbb{R}^k} f_{\mathbf{0},\mathbf{A}}(x_1, \dots, x_k)\Big(\prod_{j=1}^{2m} x_{i_j}\Big)\, dx_1 \cdots dx_k\,.$$

En développant l'exponentielle du côté droit en série, le seul terme qui contribue est le terme

$$\frac{1}{m!\,2^m}\Big(\sum_{i,j} z_i \mathbf{A}^{-1}_{ij} z_j\Big)^m = \frac{1}{m!\,2^m} \underbrace{\Big(\sum_{i,j} z_i \mathbf{A}^{-1}_{ij} z_j\Big) \cdots \Big(\sum_{i,j} z_i \mathbf{A}^{-1}_{ij} z_j\Big)}_{m \text{ facteurs}}\,. \quad (5.8)$$

En effet, les termes de puissances inférieures sont nuls par différentiation et ceux de puissances supérieures sont nuls lorsqu'on pose $z_i = 0$ pour tout i. De même, lorsque les opérateurs différentiels agissent sur le terme de droite de (5.8), on obtient une contribution non nulle seulement si sur chaque facteur agissent deux opérateurs différentiels. On peut décrire tous ces cas en considérant les appariements des opérateurs différentiels $\frac{\partial}{\partial z_{i_1}}, \dots, \frac{\partial}{\partial z_{i_{2m}}}$. Pour un appariement donné, on peut faire agir chaque paire d'opérateurs de l'appariement sur un facteur différent du produit du membre de droite de (5.8) ; il y a $m!$ manières différentes d'associer les m paires de l'appariement et les m facteurs. Comme $\mathbf{A}^{-1}$ est symétrique

$$\frac{\partial}{\partial z_k}\frac{\partial}{\partial z_\ell}\Big(\sum_{i,j} z_i \mathbf{A}^{-1}_{ij} z_j\Big) = 2\mathbf{A}^{-1}_{k\ell}\,.$$

Par conséquent, pour un appariement fixé des opérateurs différentiels, l'action des opérateurs différentiels sur le terme de droite de (5.8) donne la contribution

$$\prod_{\substack{\text{chaque}\\ \text{paire}}} \mathbf{A}^{-1}_{i_a i_b}\,.$$

□

Exemple 5.6 Soit I l'expression suivante

$$I := \frac{\int_{\mathbb{R}^3} x_1^4 x_2^2 \exp\big(-(x_1^2 + x_1x_2 + 2x_2^2 + x_3^2)\big)\, dx_1 dx_2 dx_3}{\int_{\mathbb{R}^3} \exp\big(-(x_1^2 + x_1x_2 + 2x_2^2 + x_3^2)\big)\, dx_1 dx_2 dx_3}\,.$$

Cela correspond au cas

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 4 & 0 \\ 0 & 0 & 2 \end{pmatrix} \qquad \mathbf{A}^{-1} = \frac{1}{14} \begin{pmatrix} 8 & -2 & 0 \\ -2 & 4 & 0 \\ 0 & 0 & 7 \end{pmatrix} .$$

La matrice $\mathbf{A}$ est symétrique et définie positive car les mineurs principaux de $\mathbf{A}$ sont positifs (proposition I.6). Il y a 15 appariements des indices $i_1 = 1$, $i_2 = 1$, $i_3 = 1$, $i_4 = 1$, $i_5 = 2$ et $i_6 = 2$; 3 appariements donnent la contribution $(\mathbf{A}_{11}^{-1})^2 \mathbf{A}_{22}^{-1}$ et les 12 autres la contribution $\mathbf{A}_{11}^{-1} (\mathbf{A}_{12}^{-1})^2$. Le résultat final est $I = 144/343$.

5.3 Exercices

Exercice 5.1 On considère une mesure de probabilité P définie sur $\mathbb{R}^2$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^2)$. On définit la *fonction de répartition de* P sur $\mathbb{R}^2$ par la formule

$$(t_1, t_2) \mapsto F(t_1, t_2) := P(X_1 \leq t_1, X_2 \leq t_2) .$$

Démontrer les affirmations suivantes.

1) $0 \leq F(t_1, t_2) \leq 1$ pour tout $(t_1, t_2) \in \mathbb{R}^2$.
2) F_X est monotone non décroissante en chaque variable.
3) F_X est continue à droite : si $(t_1(n), t_2(n))$, $t_i(n) \geq t_i$, $i = 1, 2$, est une suite dans $\mathbb{R}^2$ telle que $\lim_n t_i(n) = t_i$, alors

$$\lim_{n \to \infty} F(t_1(n), t_2(n)) = F(t_1, t_2) .$$

4) $\lim_{t_1 \to -\infty} F(t_1, t_2) = 0$ et $\lim_{t_1 \to \infty} F(t_1, t_2)$ est une fonction de répartition d'une mesure de probabilité sur $\mathbb{R}$. Idem si t_1 est remplacé par t_2.

Exercice 5.2 On considère la mesure de probabilité gaussienne $N(0, \sigma^2)$ sur $\mathbb{R}$. Calculer tous les *moments* de cette mesure de probabilité, i.e.

$$\int_{-\infty}^{\infty} x^n \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{x^2}{2\sigma^2}} \, dx \quad n = 1, 2, \ldots .$$

Exercice 5.3 Calculer l'expression suivante

$$\frac{\int_{\mathbb{R}^3} x_1^2 x_2 x_3^3 \exp(-2x_1^2 - 2x_1 x_2 + x_2 x_3 - x_2^2 - 2x_3^2) \, dx_1 dx_2 dx_3}{\int_{\mathbb{R}^3} \exp(-2x_1^2 - 2x_1 x_2 + x_2 x_3 - x_2^2 - 2x_3^2) \, dx_1 dx_2 dx_3} .$$

Exercice 5.4 On considère une suite réelle a_n, $n \geq 1$, et une mesure de probabilité sur $\mathbb{R}$. Soit $A_n := (-\infty, a_n)$.
1) Si la suite a_n est décroissante de limite a, est-ce que la limite $\lim_n P(A_n)$ existe ? Si oui, de quel événement $\lim_n P(A_n)$ est-elle la probabilité ?

2) Même question si a_n est croissante et de limite a.
3) Même question si a_n est une suite de limite a.

Exercice 5.5 Calculer l'intégrale

$$\int_{\mathbb{R}^2} \exp(-4x^2 + 3xy - 2y^2)\, dxdy\,.$$

Exercice 5.6 : Soit A et B deux événements, $P(A) = p$ et $P(B) = q$. Donner une borne supérieure et une borne inférieure pour la probabilité de $P(A \cap B)$ si

$$\text{(a) p = 0,3 et q = 0,5} \qquad \text{(b) p = 0,4 et q = 0,7}\,.$$

Combien de fois faut-il lancer un dé (équilibré) pour que la probabilité d'obtenir un 1 soit plus grande que 0,9 ?

Exercice 5.7 Soit $h\colon [0,x] \to \mathbb{R}^+$ une fonction positive et intégrable sur $[0,x]$. Montrer l'identité

$$\int_0^x dt_1\, h(t_1) \int_0^{t_1} dt_2\, h(t_2) \cdots \int_0^{t_{k-1}} dt_k\, h(t_k) = \frac{1}{k!}\Big(\int_0^x dt\, h(t)\Big)^k$$

par un argument probabiliste.
Indication : introduire un espace de probabilité sur $\Omega = [0,x]^k$ et ramener la question au calcul de la probabilité d'un événement approprié. Le calcul de la probabilité de cet événement peut se faire sans calcul explicite.

Exercice 5.8 En mécanique statistique un gaz idéal de N particules classiques identiques dans une boîte $\Lambda \subset \mathbb{R}^3$ de volume V, à la température T et à l'équilibre, est décrit par une mesure de Gibbs sur l'espace des phases

$$\Omega = \Lambda^N \times \mathbb{R}^{3N}\,.$$

Un élément $(\mathbf{x_1}, \ldots, \mathbf{x_N}, \mathbf{p_1}, \ldots, \mathbf{p_N}) \in \Omega$ donne la position et l'impulsion de chaque particule. La mesure de Gibbs sur Ω est spécifiée par la densité de probabilité

$$\frac{\exp\Big(-\beta \sum_{j=1}^N \frac{\langle \mathbf{p_j} | \mathbf{p_j} \rangle}{2m}\Big)}{Z}$$

où $\beta = (k_B T)^{-1}$ et Z est la fonction de partition (voir exemple 2.4).
a) Calculer Z.
b) Si l'on a une mole de gaz, l'équation des gaz parfaits s'écrit $PV = RT$. Donner l'expression de la constante R des gaz parfait à partir de la relation

$$P = \frac{\partial}{\partial V}(k_B T \ln Z)\,.$$

Exercice 5.9 On considère la mesure de probabilité uniforme sur la boule unité $\Omega = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq 1\}$. Déterminer le rayon r de la boule $B(0,r) := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq r\}$ de sorte que $P(B(0,r)) = 10^{-3}$. Estimer r pour $n = 100$. Indication : écrire $r = 1 - \varepsilon$.

Exercice 5.10 Soit $\mathbf{A}$ une matrice symétrique, définie positive, de type 3×3. Vérifier l'identité

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\mathbf{0},\mathbf{A}}(x_1, x_2, x_3)\, dx_2 dx_3 = \frac{1}{\sqrt{2\pi \mathbf{A}_{11}^{-1}}} \exp\Big(- \frac{x_1^2}{2\mathbf{A}_{11}^{-1}} \Big)$$

où $\mathbf{A}_{11}^{-1}$ est le coefficient 11 de la matrice inverse de $\mathbf{A}$.

Chapitre 6

Variable aléatoire

La notion de variable aléatoire est la notion principale de la théorie des probabilités et de la statistique.

Soit $f : A \to B$ une application définie sur un ensemble A à valeur dans un ensemble B. Soit $C \subset B$; par définition *l'image inverse de C par f*, ou *préimage de C*, est le sous-ensemble de A

$$f^{-1}(C) := \{x \in A \colon f(x) \in C\} \equiv \{f \in C\}\,.$$

Si $C = (a, b] \subset \mathbb{R}$ on écrit aussi $\{a < f \leq b\}$ à la place de $\{f \in C\}$. Le sous-ensemble $f^{-1}(C) \subset A$ est toujours défini, quel que soit f et quel que soit le sous-ensemble C ; $f^{-1}(C) = \emptyset$ est possible. Les identités

$$f^{-1}\big(\bigcap_n C_n\big) = \bigcap_n f^{-1}(C_n)$$
$$f^{-1}\big(\bigcup_n C_n\big) = \bigcup_n f^{-1}(C_n)$$

découlent directement de la définition de l'image inverse.

6.1 Variable aléatoire réelle

Définition 6.1 *Soit $(\Omega, \mathcal{F}, P)$ un espace de probabilité. Une* variable aléatoire (v.a.) réelle *est une application $X \colon \Omega \to \mathbb{R}$ telle que*

$$\forall\, t \in \mathbb{R} \colon \quad \{X \leq t\} \in \mathcal{F}\,.$$

La fonction de répartition de X *est la fonction $F_X \colon \mathbb{R} \to [0, 1]$ définie par*

$$F_X(t) := P(X \leq t)\,.$$

Le résultat ω de l'expérience est aléatoire (imprédictible) et par conséquent la valeur de X, $X(\omega)$, est aussi aléatoire. De là vient la terminologie « variable aléatoire ». Mais pour ω donné le nombre réel $X(\omega)$ est parfaitement défini puisque X est une application définie sur Ω.

Remarque 6.1 La condition, pour tout t $\{X \leq t\} \in \mathcal{F}$, est nécessaire pour pouvoir définir la fonction de répartition F_X. On peut formuler d'autres conditions équivalentes ; par exemple il suffit de vérifier cette condition pour $t \in \mathbb{Q}$ ou de montrer que $\{X < t\} \in \mathcal{F}$. De ces conditions on déduit

$$\{a < X \leq b\} = \{X \leq b\} \backslash \{X \leq a\} \in \mathcal{F}\,.$$

Il découle des identités pour les images inverses que $X^{-1}(A) \in \mathcal{F}$ pour tout $A \in \mathcal{B}(\mathbb{R})$. C'est un exercice de routine en théorie de la mesure. □

Exemple 6.1 Les applications X_j introduites dans les exemples 4.6 et 4.9 sont des v.a.. Plus généralement, si $(\Omega, \mathcal{F}, P)$ est un espace de probabilité discret, *toute application* $X\colon \Omega \to \mathbb{R}$ est une v.a. réelle puisque pour ces espaces $\mathcal{F} = \mathcal{P}(\Omega)$. □

La fonction de répartition F_X de la v.a. X est une fonction de répartition au sens de la définition 5.1. Elle est monotone non décroissante et continue à droite. En effet, si $s_n \downarrow t$

$$\{X \leq t\} = \bigcap_n \{X \leq s_n\}\,;$$

par continuité monotone de P

$$F_X(t) = \lim_{s_n \downarrow t} F_X(s_n)\,.$$

De même

$$F_X(t) \to 0 \text{ si } t \to -\infty \quad \text{et} \quad F_X(t) \to 1 \text{ si } t \to \infty\,.$$

Par conséquent F_X définit univoquement une mesure de probabilité sur $\mathbb{R}$ (théorème 5.1) qui est désignée par μ_X.

Définition 6.2 *La* loi d'une v.a. X *est la mesure de probabilité* μ_X.

Les événements, qu'on peut exprimer à l'aide de la v.a. X, sont les événements de $\mathcal{F}$ qui s'écrivent sous la forme $\{X \in A\}$ où $A \in \mathcal{B}(\mathbb{R})$ (on observe la *valeur* de X). Connaissant la loi de X, on peut calculer la probabilité de ces événements :

$$\forall A \in \mathcal{B}(\mathbb{R})\colon \quad \mu_X(A) = P(X \in A)\,. \tag{6.1}$$

Pour alléger l'écriture on écrit $P(X \in A)$ à la place de $P(\{X \in A\})$. Lorsqu'on définit une v.a., la loi est automatiquement définie par (6.1) ; il est utile cependant de considérer une v.a. comme le couple :

1) $X\colon \Omega \to \mathbb{R}$, une application ;
2) μ_X, la loi de X qui est une mesure de probabilité sur $\mathbb{R}$.

Pour indiquer la loi de X on utilise la notation

$$X \sim \mu_X \qquad \text{(la loi de } X \text{ est } \mu_X\text{)}.$$

Une *variable aléatoire discrète* est une v.a. dont la loi est une mesure de probabilité discrète. Une telle v.a. prend un nombre fini ou dénombrable de valeurs. La loi est spécifée en donnant les probabilités

$$P(X = x) \quad (x \text{ une valeur de } X)\,.$$

Cela revient à faire une partition de l'espace fondamental Ω en sous-ensembles $\{X = x\}$ et à donner les probabilités des ensembles de la partition. Si $A \subset \Omega$, *l'indicatrice de* A est la fonction I_A,

$$I_A(\omega) := 1 \text{ si } \omega \in A \quad \text{et} \quad I_A(\omega) := 0 \text{ si } \omega \notin A.$$

Une v.a. discrète X est une application de la forme

$$X(\omega) = \sum_{x \in D} x\, I_{\{X=x\}}(\omega)\,.$$

Ces v.a. sont aussi appelées *variables aléatoires étagées*. Une *variable aléatoire continue* est une v.a. dont la loi est une mesure de probabilité continue.

Souvent, c'est la loi μ_X qui nous intéresse, et non directement l'application $X : \Omega \to \mathbb{R}$, car cette loi permet de calculer les probabilités des événements $\{X \in A\}$ via l'identité (6.1). En effet, en statistique en particulier, ce qui est important ce sont les probabilités des événements qu'on peut observer à l'aide de X. Ce changement de perspective, où l'accent est mis sur les valeurs des v.a. et où l'espace de probabilité $(\Omega, \mathcal{F}, P)$ est à l'arrière plan, permet pour les besoins de ce livre de se contenter de spécifier seulement la loi des v.a. continues. C'est pourquoi on considère seulement les espaces de probabilité définis sur $\mathbb{R}$ ou $\mathbb{R}^k$ à côté des espaces de probabilité discrets.

Une *copie* (ou *représentation*) *d'une variable aléatoire* X est une v.a. Y telle que la loi de Y est égale à la loi de X. On n'exige pas que Y soit définie sur l'espace de probabilité sur lequel X est définie. On écrit $X \stackrel{\mathcal{L}}{=} Y$, si $\mu_X = \mu_Y$.

Exemples 6.2 Exemples de variables aléatoires discrètes.

a) Soit B un événement ; la loi de I_B est la mesure de probabilité discrète donnée par $D = \{0, 1\}$, $q(0) = P(B^c)$ et $q(1) = P(B)$. Par définition, une *v.a. de Bernoulli de paramètre* p est une v.a. X telle que

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

b) Soit B un événement, $P(B) = p$. Si B est réalisé, on parle de succès, sinon d'échec. On répète l'expérience n fois, de façon indépendante. On introduit la v.a. X_j, telle que $X_j = 1$ s'il y a un succès lors de la $j^{\text{ième}}$ expérience ; sinon on pose $X_j = 0$. La v.a. qui compte le nombre de succès est

$$Y := \sum_{j=1}^{n} X_j\,.$$

Une réalisation ω avec k succès est équivalente à un rangement de n boules numérotées dans deux boîtes a_0 et a_1, avec k boules dans la boîte a_1, de sorte

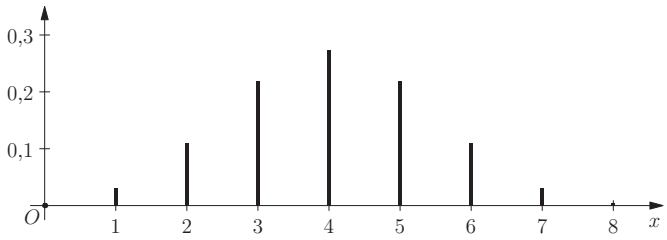


Figure 6.1 Loi binomiale $\mathcal{B}_i(8,\ 0{,}5)$ spécifiée par la position et la hauteur des sauts. Les paramètres sont $n = 8$ et $p = 0{,}5$.

que la probabilité de mettre une boule dans la boîte a_1 est p. La probabilité d'un rangement correspondant à $Y = k$ est $p^k(1-p)^{n-k}$; il y a $\binom{n}{k}$ rangements distincts. Par conséquent

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}\,.$$

Y est une *v.a. binomiale de paramètres* (n, p). La loi de Y est notée $\mathcal{B}_i(n, p)$. Elle est définie sur $D = \{0, \ldots, n\}$ et $\mathcal{B}_i(n, p)(k) = P(Y = k)$.

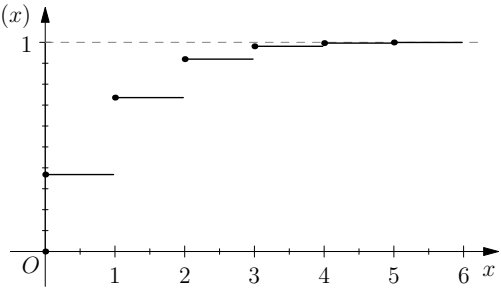


Figure 6.2 Fonction de répartition de la loi de Poisson π_λ. Cette fonction de répartition a une infinité de sauts. Le paramètre est $\lambda = 1$.

c) X est une *v.a. de Poisson de paramètre* λ si $\lambda > 0$ et

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

La loi de cette v.a. est la mesure de Poisson π_λ. □

Exemples 6.3 Exemples de variables aléatoires continues.

a) Soit $m \in (-\infty, \infty)$ et $\sigma \in (0, \infty)$. X est une *v.a. gaussienne de paramètres* (m, σ^2), ou *v.a. normale* $N(m, \sigma^2)$, si la loi de X est une mesure de probabilité gaussienne de densité

$$f_{m,\sigma^2}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(m-t)^2}{2\sigma^2}\Big)\,.$$

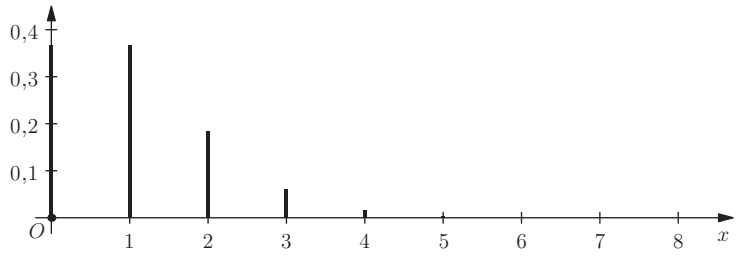


Figure 6.3 – Loi de Poisson π_λ spécifiée par la position et la hauteur des sauts. Le paramètre est $\lambda = 1$.

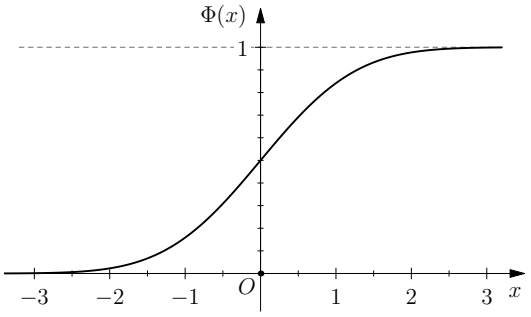


Figure 6.4 – Fonction de répartition de la loi gaussienne réduite $N(0,1)$.

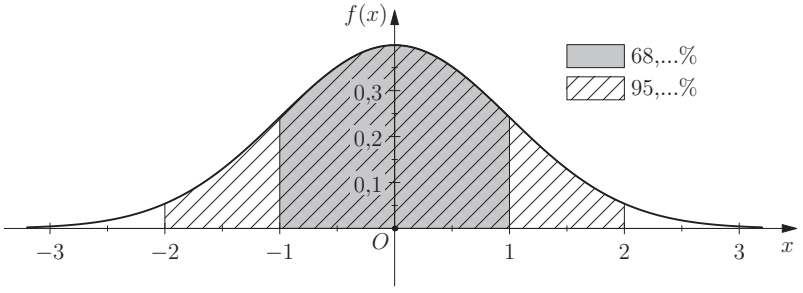


Figure 6.5 – Densité de probabilité de la loi gaussienne réduite $N(0,1)$.

Si $m = 0$ et $\sigma = 1$, on parle de *loi normale standard* ou *loi normale réduite*. Si $X \sim N(m, \sigma^2)$, par changement de variable, on peut se ramener au cas $N(0,1)$. Les lois normales sont parmi les plus importantes. Si $X \sim N(0,1)$,

$$P(-1 \le X \le 1) \simeq 0{,}6826 \qquad P(-2 \le X \le 2) \simeq 0{,}9544$$
$$P(-3 \le X \le 3) \simeq 0{,}9975 \qquad P(|X| > 4) \simeq 0\,.$$

La fonction de répartition de la loi $N(0,1)$ est notée

$$\Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}\, dt\,.$$

Deux valeurs à retenir,

$$\Phi(1) = 0{,}8413 \quad \text{et} \quad \Phi(1{,}96) = 0{,}975\,.$$

Inégalités importantes :

$$\text{si } x > 0\text{:} \quad (x^{-1} - x^{-3}) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} \le 1 - \Phi(x) \le x^{-1} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}\,.$$

Pour les vérifier il suffit d'écrire

$$1 - \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \mathrm{e}^{-\frac{t^2}{2}}\, dt = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} t^{-1} \frac{d}{dt}\big(-\mathrm{e}^{-\frac{t^2}{2}} \big)\, dt\,.$$

La borne supérieure découle de l'inégalité $t^{-1} \le x^{-1}$, puis d'une intégration. La borne inférieure est obtenue en faisant d'abord une intégration par parties, puis on procède comme pour la borne supérieure.

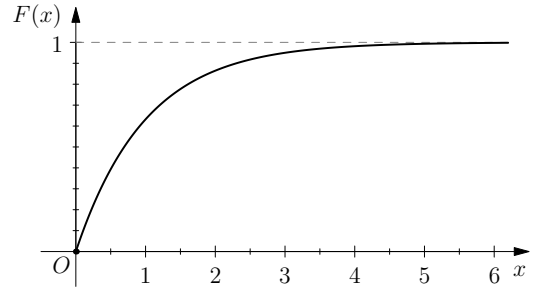


Figure 6.6 – Fonction de répartition de la loi exponentielle de paramètre $\lambda = 1$.

b) Soit $x > 0$ et $\lambda > 0$. X est une *v.a. gamma de paramètres* (x, λ) si la loi de X est la mesure de probabilité $\gamma_{x,\lambda}$; sa densité est

$$\gamma_{x,\lambda}(t) = \Gamma(x)^{-1} \lambda \mathrm{e}^{-\lambda t} (\lambda t)^{x-1} I_{\mathbb{R}^+}(t)\,.$$

Un cas particulier important est celui où $x = 1$. Lorsque $X \sim \gamma_{1,\lambda}$, X est une *v.a. exponentielle de paramètre* λ ; la densité de la loi de X est

$$\lambda \mathrm{e}^{-\lambda t} I_{\mathbb{R}^+}(t)\,.$$

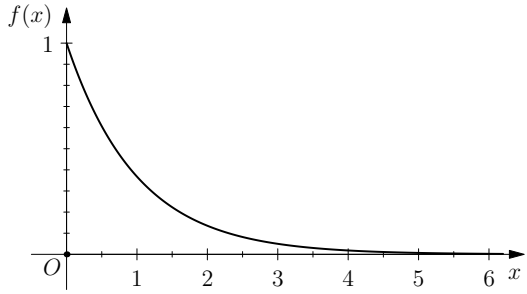


Figure 6.7 – Densité de probabilité de la loi exponentielle de paramètre $\lambda = 1$.

c) Dans le plan $\mathbb{R}^2$ on considère les points A de coordonnées $(0,1)$ et B de coordonnées $(0,0)$. Un rayon de lumière est émis en A et coupe l'axe horizontal $y = 0$ au point X. On désigne par θ l'angle de sommet A du triangle formé par les sommets X, A et B. Par convention, $-\pi/2 < \theta < \pi/2$ et $\theta < 0$ si $X < 0$. Si chaque angle θ est également probable,

$$P(X \leq t) = P(\tan\theta \leq t) = P(\theta \leq \tan^{-1} t) = \frac{1}{\pi}\big(\tan^{-1} t - (-\pi/2)\big)\,.$$

La densité de probabilité de X est donnée par

$$\frac{d}{dt}\Big(\frac{1}{2} + \frac{1}{\pi}\tan^{-1} t\Big) = \frac{1}{\pi(1+t^2)}\,.$$

X est une *v.a. de Cauchy*. De manière plus générale, *la loi de Cauchy de paramètre* $a > 0$ est donnée par la densité

$$\frac{a}{\pi(a^2+t^2)}\,, \quad t \in \mathbb{R}\,.$$

La densité de la loi de Cauchy apparaît dans plusieurs disciplines, en particulier en spectroscopie sous le nom de *fonction lorentzienne* (H. Lorentz (1853-1928)). Sa transformée de Fourier est

$$\int_{-\infty}^{\infty} \frac{a}{\pi(a^2+t^2)} \mathrm{e}^{i\omega t}\, dt = \mathrm{e}^{-a|\omega|}\,.$$

d) Soit $[a,b]$ un intervalle. La *loi uniforme sur* $[a,b]$ est donnée par la densité

$$f(t) = \frac{1}{b-a} I_{[a,b]}(t)\,.$$

X est une *v.a. uniforme sur* $[a,b]$ si la loi de X est la loi uniforme sur $[a,b]$. □

Soit $\varphi\colon \mathbb{R} \to \mathbb{R}$ une fonction réelle [1] et $X\colon \Omega \to \mathbb{R}$ une v.a. réelle. On définit une nouvelle v.a. réelle en posant $Y := \varphi \circ X \equiv \varphi(X)$. D'une manière générale

1. Il faut que $\varphi^{-1}((-\infty,t]) \in \mathcal{B}(\mathbb{R})$ pour tout $t \in \mathbb{R}$. Cette condition n'est pas restrictive ; elle implique que $\varphi^{-1}(A) \in \mathcal{B}(\mathbb{R})$ pour tout $A \in \mathcal{B}(\mathbb{R})$. Toutes les fonctions continues vérifient cette condition.

le calcul de la loi de la v.a. $Y = \varphi(X)$ n'est pas aisé. Dans le cas d'une v.a. discrète, qui prend ses valeurs dans l'ensemble fini ou dénombrable D, la v.a. Y est aussi discrète et

$$P(Y = y) = \sum_{x \in D\,:\, \varphi(x) = y} P(X = x)\,.$$

Exemple 6.4 Suite de l'exemple 2.4. L'aimantation (par spin) est une v.a. M_n définie sur Ω_n et qui prend ses valeurs dans l'ensemble

$$D := \{x_k = \frac{2k}{n} - 1\,,\ k = 0, 1, \ldots, n\}\,.$$

L'énergie du système (2.3) est une v.a. qui s'écrit aussi

$$H_n = -n\frac{M_n^2}{2} - nhM_n\,.$$

La loi de M_n est discrète ; elle est spécifiée par D et

$$P(M_n = x_k) = \binom{n}{k} \frac{\exp\left(n\beta\frac{x_k^2}{2} + n\beta h x_k\right)}{Z_n}\,.$$

Z_n est la fonction de partition. Pour évaluer $P(M_n = x_k)$ il faut évaluer Z_n et le coefficient binomial $\binom{n}{k}$ qui donne le nombre de configurations de Ω_n telles que $M_n = x_k$. Le coefficient binomial est estimé à l'aide de la fonction s définie sur $[-1, 1]$ par (avec la convention $0 \ln 0 = \lim_{x \downarrow 0} x \ln x = 0$)

$$s(x) := -\frac{1+x}{2} \ln \frac{1+x}{2} - \frac{1-x}{2} \ln \frac{1-x}{2}\,.$$

Il existe des constantes C_1 et C_2 telles que

$$\frac{C_1}{\sqrt{n}} \exp(ns(x_k)) \le \binom{n}{k} \le C_2 \exp(ns(x_k)) \quad \text{si } k = 0, 1, \ldots, n. \tag{6.2}$$

Ce résultat s'obtient pour $k = 1, \ldots, n-1$ à partir des inégalités 3.2, puis on note que les inégalités (6.2) sont encore vraies si $k = 0$ ou $k = n$. Par exemple,

$$\begin{aligned}\frac{n!}{k!(n-k)!} &\ge \frac{\sqrt{\mathrm{e}}\sqrt{n} n^n \mathrm{e}^{-n}}{\mathrm{e}^2 \sqrt{k}\sqrt{n-k}\, k^k \mathrm{e}^{-k}\, (n-k)^{n-k} \mathrm{e}^{-(n-k)}} \\ &\ge \frac{C_1}{\sqrt{n}} \exp\Big[- n\Big(\frac{k}{n} \ln \frac{k}{n} + \frac{(n-k)}{n} \ln \frac{(n-k)}{n}\Big)\Big]\,.\end{aligned}$$

(Pour C_2 on peut prendre $C_2 = 1$, voir exercice 8.1).

Remarque 6.2 La fonction s est non-négative (somme de deux termes non-négatifs) et concave ; son maximum est en $x = 0$ et $s(0) = \ln 2$. Elle représente l'entropie d'une v.a. de Bernoulli telle que

$$P(X = 1) = \frac{1+x}{2} \quad \text{et} \quad P(X = -1) = \frac{1-x}{2}\,.$$

De manière plus générale, si X est une v.a. discrète prenant ses valeurs dans D, l'*entropie de la v.a. X* (*entropie de Shannon* (1916-2001)) est la quantité

$$H(X) := -\sum_{x\in D} P(X=x)\ln P(X=x)\,. \tag{6.3}$$

$H(X)$ est une mesure de l'*incertitude de la v.a. X*. L'*incertitude sur la valeur x de la v.a. X* est par définition $-\ln P(X=x)$. L'entropie de X est maximale lorsque chaque valeur de X est également probable (voir exercice 6.9). Par contre, pour une v.a. de Bernoulli de paramètre p proche de 1, l'entropie est petite puisque $X=1$ avec grande probabilité. □

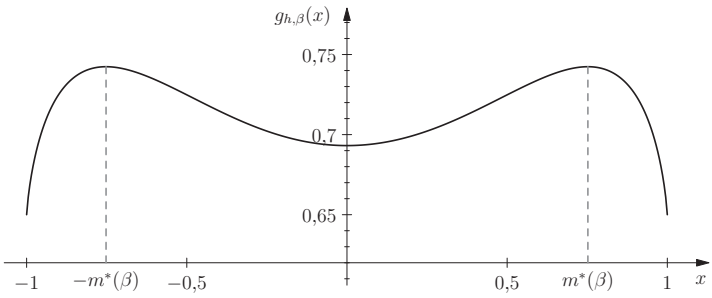


Figure 6.8 – La fonction $x \mapsto g_{h,\beta}(x)$ pour $h=0$ et $\beta = 1.3$.

A partir des inégalités (6.2) il est facile d'estimer Z_n. Soit

$$g_{h,\beta}(x) := \beta\frac{x^2}{2} + \beta h x + s(x)\,. \tag{6.4}$$

En effet,

$$\frac{C_1}{\sqrt{n}Z_n}\,\exp(n g_{h,\beta}(x_k)) \le P(M_n = x_k) \le \frac{C_2}{Z_n}\exp(n g_{h,\beta}(x_k))\,. \tag{6.5}$$

Il suffit de sommer cette équation sur les $(n+1)$ valeurs de M_n pour obtenir la borne supérieure dans l'expression ci-dessous,

$$\frac{C_1}{\sqrt{n}}\exp\big(n\max_{x_k\in D} g_{h,\beta}(x_k)\big) \le Z_n \le C_2(n+1)\exp\big(n\max_{x_k\in D} g_{h,\beta}(x_k)\big)\,.$$

On vérifie par un calcul de routine les propriétés suivantes :

1. si $\beta \le 1$, $g_{h,\beta}$ est concave sur $[-1,1]$;
2. si $\beta > 1$, $g_{h,\beta}$ est convexe sur $\Big[-\dfrac{\sqrt{\beta-1}}{\sqrt{\beta}}, \dfrac{\sqrt{\beta-1}}{\sqrt{\beta}}\Big]$, sinon elle est concave (voir figure 6.8 pour le cas $h=0$) ;
3. un point critique x de $g_{h,\beta}$ vérifie l'équation

$$g'_{h,\beta}(x) = 0 \iff 2\beta(h+x) = \ln\frac{1+x}{1-x} \iff x = \tanh(\beta x + \beta h)\,. \tag{6.6}$$

Les points critiques de la fonction $g_{h,\beta}$ peuvent être déterminés graphiquement à l'aide de la troisième équation (6.6), qui peut être mise sous la forme

$$\tanh(u) = \frac{1}{\beta}u - h \quad \text{avec} \quad x = \frac{u}{\beta} - h\,.$$

On les obtient en trouvant les points d'intersection des graphes de $u \mapsto \tanh(u)$ et de la droite $u \mapsto \beta^{-1}u - h$. Si $h \neq 0$ ou si $\beta \leq 1$, il y a un maximum global unique de $g_{h,\beta}$, qui est noté $m^*(h,\beta)$; lorsque $h > 0$, $m^*(h,\beta)$ est la solution positive x de (6.6). L'analyse graphique permet d'étudier le comportement de $m^*(h,\beta)$ en fonction de h. Si $\beta \leq 1$, $m^*(h,\beta)$ est une fonction continue de h, alors que si $\beta > 1$, $m^*(h,\beta)$ présente une discontinuité en $h = 0$,

$$m^*(\beta) := \lim_{h \downarrow 0} m^*(h,\beta) > 0\,. \tag{6.7}$$

(Voir figure 10.1). Par symétrie, lorsque $\beta > 1$ et $h = 0$, il y a deux maxima globaux de $g_{h,\beta}$ en $\pm m^*(\beta)$.

La dérivée de $g_{h,\beta}$ étant continue, pour tout sous-ensemble fermé $E \subset (-1,1)$, il existe une constante C telle que $|g'_{h,\beta}(x)| \leq C$ pour tout $x \in E$. Par conséquent

$$|g_{h,\beta}(x) - g_{h,\beta}(y)| \leq C|x-y| \quad \forall\, x, y \in E\,. \tag{6.8}$$

Pour tout β, il existe un intervalle fermé $I \subset (-1,1)$ qui contient les maxima globaux de $g_{h,\beta}$. En prenant $E = I$ on obtient

$$\Big|\max_{x\in[-1,1]} g_{h,\beta}(x) - \max_{x_k \in D} g_{h,\beta}(x_k)\Big| \leq \frac{2C}{n}\,. \tag{6.9}$$

Ces résultats donnent immédiatement l'énergie libre par spin (2.6) dans la limite thermodynamique,

$$f(h,\beta) = -\lim_{n\to\infty} \frac{1}{\beta n} \ln Z_n = -\max_{x\in[-1,1]} \Big(\frac{x^2}{2} + hx + k_B T\, s(x)\Big)\,. \tag{6.10}$$

En utilisant la deuxième équation (6.6) on peut aussi écrire l'énergie libre sous la forme

$$f(h,\beta) = \frac{(m^*(h,\beta))^2}{2} + \frac{k_B T}{2} \ln \frac{1 - (m^*(h,\beta))^2}{4}\,.$$

Suite à l'exemple 8.4 section 8.2. □

Si X est continue et si φ est C^1 et strictement monotone, alors la loi de $Y = \varphi(X)$ est donnée par la densité de probabilité

$$f_Y(y) = f_X(\varphi^{-1}(y)) \Big| \frac{d}{dy} \varphi^{-1}(y) \Big|\,. \tag{6.11}$$

Vérification de l'identité (6.11) dans le cas monotone décroissant.

$$\begin{aligned} P(\varphi(X) \le t) &= P(X \ge \varphi^{-1}(t)) \\ &= 1 - P(X \le \varphi^{-1}(t)) \\ &= 1 - F_X(\varphi^{-1}(t))\,. \end{aligned}$$

En dérivant par rapport à t on obtient

$$f_Y(t) = -f_X(\varphi^{-1}(t))\frac{d}{dt}\varphi^{-1}(t) = f_X(\varphi^{-1}(t))\Big|\frac{d}{dt}\varphi^{-1}(t)\Big|\,.$$

Exemple 6.5 Particule dans un puit de potentiel de barrière d'énergie E positive. La loi d'Arrhénius (1859-1927) donne le temps de sortie (par fluctuations thermiques) :

$$\tau(E) = \tau_0 \mathrm{e}^{E/k_B T}\,.$$

La constante τ_0 est un temps de référence caractéristique, T est la température absolue et k_B est la constante de Boltzmann. On suppose que la barrière est décrite par une v.a. X de loi exponentielle de paramètre $\lambda = 1/E_0$ où E_0 est une énergie de référence. Le temps de sortie est aussi une v.a.,

$$Y := \tau(X) \equiv \varphi(X)$$

avec

$$\varphi\colon \mathbb{R}^+ \to [\tau_0, \infty)\,,\ x \mapsto \varphi(x) := \tau_0 \mathrm{e}^{x/k_B T}\,.$$

On peut inverser φ sur $[\tau_0, \infty)$,

$$\varphi^{-1}(t) = k_B T \ln(t/\tau_0) \quad t \ge \tau_0\,.$$

La densité de probabilité de la v.a. Y est

$$f_Y(t) = I_{[\tau_0,\infty)}(t) f_X(\varphi^{-1}(t))\Big|\frac{d}{dt}\varphi^{-1}(t)\Big|\,.$$

Si $t \ge \tau_0$:

$$\begin{aligned} f_Y(t) &= \frac{1}{E_0}\exp\Big(-\frac{1}{E_0}k_B T \ln\frac{t}{\tau_0}\Big)\cdot k_B T\frac{\tau_0}{t}\frac{1}{\tau_0} \\ &= \frac{k_B T}{E_0}\Big(\frac{\tau_0}{t}\Big)^{\frac{k_B T}{E_0}}\frac{1}{t} \\ &\equiv \alpha\frac{\tau_0^\alpha}{t^{\alpha+1}}\,. \end{aligned}$$

La loi de Y est une *loi de Pareto* (1848-1923). Si α est petit, la densité de probabilité tend lentement vers 0 lorsque t tend vers l'infini. Ceci a des conséquences importantes sur le comportement du système. □

6.2 Construction d'une variable aléatoire réelle

Chaque mesure de probabilité μ sur $\mathbb{R}$ est caractérisée par sa fonction de répartition F. Dans ce paragraphe on répond à la question : peut-on construire un espace de probabilité Ω et une application $X : \Omega \to \mathbb{R}$ de sorte que $X \sim \mu$?

Une première solution est de poser $(\Omega, \mathcal{F}, P) := (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$ et de définir la v.a. $X : \Omega \to \mathbb{R}$ par

$$X(t) := t \quad \text{(application identité)}\,.$$

On a bien sûr $P(X \in A) = \mu(A)$. On dit que X est la *représentation canonique d'une v.a. de loi* μ. Pour chaque mesure de probabilité μ on a besoin d'un espace de probabilité différent pour définir X.

Une autre solution est de choisir l'espace de probabilité d'un GNA : $\Omega = (0,1)$, $\mathcal{F} := \mathcal{B}((0,1))$ et P est la mesure de probabilité uniforme sur $(0,1)$. L'intérêt de ce choix est que l'espace de probabilité ne dépend pas de μ. On peut ainsi construire des v.a. de lois différentes sur le même espace de probabilité. Soit F la fonction de répartition de la loi μ. Considérons d'abord le cas simple où F est continue, $F(t) = 0$ si $t \le a$, strictement croissante sur $[a,b]$, et $F(t) = 1$ si $t \ge b$. Dans ce cas F applique l'intervalle (a,b) de façon bijective sur $(0,1)$; on pose

$$\forall\, \omega \in (0,1)\colon \quad X(\omega) := F^{-1}(\omega)\,.$$

La fonction de répartition de la v.a. ainsi construite est F. En effet

$$P(X \le s) = P(\{\omega\colon F^{-1}(\omega) \le s\}) = P(\{\omega\colon \omega \le F(s)\}) = F(s)\,.$$

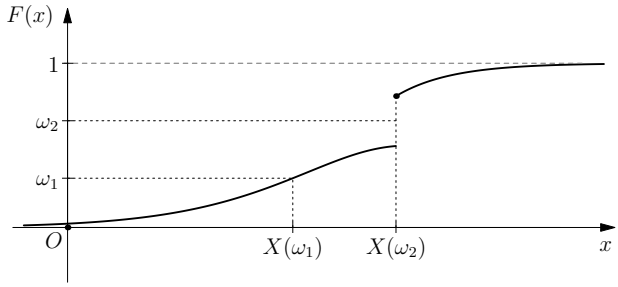


Figure 6.9 – Construction d'une v.a. X sur l'espace de probabilité d'un GNA à partir de la fonction de répartition de X.

Exemple 6.6 Donner n valeurs d'une v.a. exponentielle $X \sim \gamma_{1,\lambda}$ (n nombres aléatoires tirés selon une loi exponentielle). A l'aide d'un GNA on génère n nombres aléatoires $a_1, \ldots, a_n$, puis on évalue $F^{-1}(a_i)$ pour chaque i, où $F(t) = 1 - \mathrm{e}^{-\lambda t}$. □

La méthode ci-dessus se généralise pour toute v.a. (voir figure 6.9). On pose

$$\forall\, \omega \in (0,1)\colon \quad X(\omega) := \min\{s \in \mathbb{R}\colon F(s) \geq \omega\}\,.$$

La continuité à droite de F garantit l'existence du minimum. Pour vérifier que $F_X = F$ il suffit de montrer que

$$\{\omega\colon X(\omega) \leq s\} = \{\omega\colon \omega \leq F(s)\}\,.$$

Si $\omega \leq F(s)$, alors $X(\omega) = \min\{t\colon F(t) \geq \omega\} \leq s$. Inversement, si $X(\omega) \leq s$, alors $\min\{t\colon F(t) \geq \omega\} \leq s$ et donc $\omega \leq F(s)$.

Exemple 6.7 Par cette construction on obtient pour la loi de Bernoulli de paramètre $0 < p < 1$

$$Y(\omega) := \begin{cases} 0 & \text{si } \omega \in (0,(1-p)] \\ 1 & \text{si } \omega \in ((1-p),1). \end{cases}$$

Pour la loi de Poisson π_λ on obtient

$$X(\omega) = \begin{cases} 0 & \text{si } \omega \in (0, \mathrm{e}^{-\lambda}] \\ n & \text{si } \omega \in \left(\mathrm{e}^{-\lambda}\sum_{k=0}^{n-1}\frac{\lambda^k}{k!}, \mathrm{e}^{-\lambda}\sum_{k=0}^{n}\frac{\lambda^k}{k!}\right], \ n \geq 1. \end{cases}$$

6.3 Plusieurs variables aléatoires réelles

Lors de l'étude d'une expérience aléatoire on utilise la plupart du temps plusieurs v.a. réelles $X_1, \ldots, X_k$ qui sont *définies sur le même espace de probabilité* puisqu'elles se rapportent à la même expérience. On désigne en général les v.a. par des lettres majuscules X, Y, Z etc. et leurs valeurs par des lettres minuscules x, y, z etc. De façon équivalente, on peut considérer que les v.a. réelles $X_1, \ldots, X_k$ définissent une v.a. $\mathbf{X}$ à valeur dans $\mathbb{R}^k$ en posant

$$\mathbf{X} := (X_1, \ldots, X_k)\,.$$

Définition 6.3 *Soit $X_1, \ldots, X_k$ des v.a. définies sur $(\Omega, \mathcal{F}, P)$. La* loi conjointe *de $X_1, \ldots, X_k$ est la loi de la v.a. $\mathbf{X}$, i.e. la mesure de probabilité sur $(\mathbb{R}^k, \mathcal{B}(R^k))$ définie par l'identité*

$$\mu_{\mathbf{X}}(A) \equiv \mu_{X_1,\ldots,X_k}(A) = P(\mathbf{X} \in A) \quad \forall A \in \mathcal{B}(\mathbb{R}^k)\,. \tag{6.12}$$

La loi conjointe $\mu_{\mathbf{X}}$ permet de calculer, via l'identité (6.12), toutes les probabilités des événements exprimables par $X_1, \ldots X_k$, i.e. les événements de la forme $\{\mathbf{X} \in A\}$, $A \in \mathcal{B}(\mathbb{R}^k)$. Dans le cas de v.a. continues la loi conjointe est aussi continue. *Chaque fois qu'on considère une question concernant plusieurs v.a. c'est la loi conjointe qu'il faut utiliser.*

Comme dans le cas d'une v.a. réelle, si l'on connaît la loi conjointe $\mu_{\mathbf{X}}$ de k v.a. réelles, alors il existe une *représentation canonique* de ces variables : l'espace de probabilité est $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k), \mu_{\mathbf{X}})$ et

$$X_i \colon \mathbb{R}^k \to \mathbb{R}\,, \quad X_i(\mathbf{x}) \equiv X_i(x_1, \ldots, x_k) := x_i\,.$$

Dans le cas discret, où chaque X_i prend ses valeurs dans un ensemble fini ou dénombrable D_i, on peut aussi utiliser une repésentation des v.a. sur l'espace de probabilité discret défini par

$$\Omega = D_1 \times \cdots \times D_k \quad \text{et} \quad q(x_1, \ldots, x_k) := P(X_1 = x_1, \ldots, X_k = x_k)\,.$$

Noter qu'on peut avoir $q(x_1, \ldots, x_k) = 0$.

Exemple 6.8 Les v.a. $X_1, \ldots, X_k$ sont *gaussiennes*, ou le *vecteur aléatoire* $\mathbf{X} = (X_1, \ldots, X_k)$ *est gaussien*, si la loi conjointe est une mesure de probabilité gaussienne $N(\mathbf{m}, \mathbf{A})$ sur $\mathbb{R}^k$. □

Lorsqu'on connaît la loi conjointe des $X_1, \ldots, X_k$ on calcule la loi de X_i, appelée aussi *$i^{ième}$ loi marginale*, par la formule suivante :

$$\begin{aligned}\mu_{X_i}(A) &= P(X_i \in A)\\ &= P(X_1 \in \mathbb{R}, \ldots, X_{i-1} \in \mathbb{R}, X_i \in A, X_{i+1} \in \mathbb{R}, \ldots, X_k \in \mathbb{R})\\ &= \mu_{\mathbf{X}}(\underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{i-1 \text{ facteurs}} \times A \times \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{k-i \text{ facteurs}})\,.\end{aligned}$$

Exemple 6.9 On connaît la loi conjointe discrète de X_1, X_2 et X_3. Cette loi conjointe est donnée par l'ensemble fini ou dénombrable $D \subset \mathbb{R}^3$ et l'application $q : D \to \mathbb{R}^+$ telle que $\sum_{\mathbf{x} \in D} q(\mathbf{x}) = 1$. La loi de X_2 est aussi discrète,

$$\begin{aligned}P(X_2 = x_2) &= \mu_{\mathbf{X}}(\mathbb{R} \times \{x_2\} \times \mathbb{R})\\ &= \sum_{\mathbf{y} \in D\,:\, y_2 = x_2} q(\mathbf{y})\\ &= \sum_{\substack{y_1, y_3\,:\\ (y_1, x_2, y_3) \in D}} q(y_1, x_2, y_3)\,.\end{aligned}$$

Exemple 6.10 On connaît la loi conjointe continue de X_1, X_2 et X_3. Cette loi conjointe est donnée par une densité de probabilité $f(x_1, x_2, x_3)$ définie sur $\mathbb{R}^3$. La loi de X_2 est aussi continue avec densité g,

$$\begin{aligned}P(X_2 \in A) &= \int_{\mathbb{R}} \int_A \int_{\mathbb{R}} f(x_1, x_2, x_3)\, dx_1 dx_2 dx_3\\ &= \int_A dx_2 \underbrace{\int_{\mathbb{R}} dx_1 \int_{\mathbb{R}} dx_3 f(x_1, x_2, x_3)}_{\equiv g(x_2)} = \int_A g(x_2)\, dx_2\,.\end{aligned}$$

Par exemple, si la loi conjointe est donnée par la densité de probabilité

$$f(x_1, x_2, x_3) := \frac{1}{(2\pi\sigma^2)^{\frac{3}{2}}} \exp\Big(- \frac{x_1^2 + x_2^2 + x_3^2}{2\sigma^2} \Big) ,$$

la loi de X_2 est la loi $N(0, \sigma^2)$. □

Exemple 6.11 On choisit au hasard un point dans un disque de rayon 1. Pour modéliser cette expérience on utilise

$$\Omega := \{\omega = (x, y) \in \mathbb{R}^2 : \, x^2 + y^2 \leq 1\}$$

et la mesure uniforme sur Ω. On définit deux v.a.

$$X(x, y) := x \quad \text{et} \quad Y(x, y) := y \, .$$

La loi conjointe de X et Y est donnée par la densité

$$f_{X,Y}(x, y) = \frac{1}{\pi} \text{ si } x^2 + y^2 \leq 1 \quad \text{et} \quad f_{X,Y}(x, y) = 0 \text{ sinon.}$$

Exemple d'un calcul de la probabilité d'un événement :

$$P\Big(X \geq 0, Y \in \Big[-\frac{1}{2}, 1 \Big]\Big) = \int_0^1 dx \int_{-1/2}^1 dy \, f_{X,Y}(x, y) \, .$$

Calcul de la loi marginale de X. La densité de cette loi est

$$f_X(x) = \begin{cases} \displaystyle\int_{\mathbb{R}} dy f_{X,Y}(x, y) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} \, dy = \frac{2}{\pi} \sqrt{1 - x^2} & \text{si } |x| \leq 1 \\ 0 & \text{sinon.} \end{cases}$$

Expression analogue pour f_Y. Noter que $f_X(x) f_Y(y) \neq f_{X,Y}(x, y)$: *les lois marginales ne déterminent pas la loi conjointe*, sauf dans le cas de v.a. indépendantes.

Si lors de l'expérience on observe seulement la distance du point au centre du disque, la v.a. naturelle est

$$R(x, y) := \sqrt{x^2 + y^2} \, .$$

La fonction de répartition de R est par conséquent

$$F(t) = \begin{cases} 0 & \text{si } t \leq 0 \\ P(R \leq t) = t^2 & \text{si } 0 \leq t \leq 1 \\ 1 & \text{si } t \geq 1. \end{cases}$$

La densité de la loi de R est $g(t) = F'(t) = 2t I_{[0,1]}(t)$. □

6.4 Variables aléatoires indépendantes

Définition 6.4 *Les v.a. $X_1, \ldots, X_k$ définies sur le même espace de probabilité sont* indépendantes *si et seulement si pour tout $A_1 \in \mathcal{B}(\mathbb{R}), \ldots, A_k \in \mathcal{B}(\mathbb{R})$*

$$P(X_1 \in A_1, \ldots, X_k \in A_k) = P(X_1 \in A_1) \cdots P(X_k \in A_k)\,. \tag{6.13}$$

La condition (6.13) *est équivalente à*

$$\mu_{\mathbf{X}}(A_1 \times \cdots \times A_k) = \mu_{X_1}(A_1) \cdots \mu_{X_k}(A_k)\,.$$

Dans ce cas, et uniquement dans celui-ci, la loi conjointe des $X_1, \ldots, X_k$ est le produit des lois marginales.

Remarque 6.3 Dans le cas de v.a. discrètes la loi de chaque v.a. X_i est spécifiée par les probabilités $P(X_i = x_i)$; les événements $\{X_i = x_i\}$, x_i parcourant les valeurs de X_i, forment une partition de Ω. Les conditions (6.13) signifient simplement que les algèbres engendrées par ces partitions sont indépendantes. Ces conditions sont équivalentes aux conditions (proposition 4.2) :

$$\forall y_1, \ldots, \forall y_k : \quad P(X_1 = y_1, \ldots, X_k = y_k) = P(X_1 = y_1) \cdots P(X_k = y_k)\,.$$

Dans le cas général, les conditions (6.13) signifient que les σ-algèbres $\mathcal{F}_{X_i}$ sont indépendantes ; par définition, la σ-algèbre $\mathcal{F}_{X_j}$ est formée par la collection de tous les événements exprimables par X_j. □

Proposition 6.1 *Les v.a. $X_1, \ldots, X_k$ sont indépendantes si et seulement si pour tout $t_1 \in \mathbb{R}, \ldots, t_k \in \mathbb{R}$*

$$P(X_1 \leq t_1, \ldots, X_k \leq t_k) = P(X_1 \leq t_1) \cdots P(X_k \leq t_k)\,.$$

Dans le cas continu les v.a. $X_1, \ldots, X_k$ sont indépendantes si et seulement si pour tout $x_1 \in \mathbb{R}, \ldots, x_k \in \mathbb{R}$

$$f_{\mathbf{X}}(x_1, \ldots, x_k) = f_{X_1}(x_1) \cdots f_{X_k}(x_k)$$

où $f_{\mathbf{X}}$ est la densité de la loi conjointe et f_{X_i} la densité de la loi de X_i.

Preuve Vérification dans le cas continu pour $k = 2$. L'indépendance implique entre autres

$$\begin{aligned} P(X_1 \leq t, X_2 \leq s) &= \int_{-\infty}^{t} dx_1 \int_{-\infty}^{s} dx_2 f_{X_1,X_2}(x_1, x_2) \\ &= \int_{-\infty}^{t} dx_1 f_{X_1}(x_1) \int_{-\infty}^{s} dx_2 f_{X_2}(x_2)\,. \end{aligned}$$

En dérivant cette expression par rapport à t et s on obtient

$$f_{X_1,X_2}(t, s) = f_{X_1}(t) f_{X_2}(s)\,.$$

Inversement, si la densité de la loi conjointe factorise, par un calcul immédiat on vérifie les conditions (6.13). □

Exemple 6.12 a) Les v.a. X et Y de l'exemple 6.11 ne sont pas indépendantes puisque $f_X(x)f_Y(y) \neq f_{X,Y}(x,y)$.
b) On choisit au hasard un point dans le carré $[0,1]\times[0,1] \subset \mathbb{R}^2$. Les coordonnées cartésiennes du point $\omega = (x,y)$ définissent deux v.a. X et Y. Les v.a. X^2 et Y^3 sont indépendantes car pour tout t et pour tout s

$$P(X^2 \leq t, Y^3 \leq s) = P(X^2 \leq t)P(Y^3 \leq s)\,.$$

En effet, si $t < 0$ ou $s < 0$ le résultat est évident ; si $t \geq 0$ et $s \geq 0$,

$$\begin{aligned} P(X^2 \leq t, Y^3 \leq s) &= P(X \leq t^{1/2}, Y \leq s^{1/3}) = t^{1/2}\, s^{1/3} \\ &= P(X \leq t^{1/2})P(Y \leq s^{1/3}) \\ &= P(X^2 \leq t)P(Y^3 \leq s)\,. \end{aligned}$$

Par contre, les v.a. X et $X+Y$ ne sont pas indépendantes,

$$P(X \leq 1/2, X+Y \leq 1) = \frac{3}{8} \neq P(X \leq 1/2)P(X+Y \leq 1) = \frac{1}{4}\,.$$

Exemple 6.13 On considère des v.a. réelles X et Y ayant une densité de probabilité conjointe qui est une fonction de x^2+y^2. Les v.a. X et Y de l'exemple 6.11 ont cette propriété ; elles sont identiquement distribuées mais ne sont pas indépendantes. Lorsque X et Y sont identiquement distribuées *et indépendantes* on a le résultat remarquable suivant. Si pour tout x et y

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = g(x^2+y^2) > 0$$

et si f_X et g sont dérivables, alors

$$f_X'(x)\, f_Y(y) = 2xg'(x^2+y^2)$$

et

$$\frac{f_X'(x)}{2xf_X(x)} = \frac{g'(x^2+y^2)}{g(x^2+y^2)} = \frac{f_X'(y)}{2yf_X(y)}\,.$$

On en déduit que le quotient de gauche est égal à une constante.

$$\frac{f_X'(x)}{xf_X(x)} = c \implies \frac{d}{dx}(\ln f_X(x)) = cx \implies f_X(x) = k\mathrm{e}^{cx^2/2}\,.$$

Comme $f_X(x) > 0$ sur $\mathbb{R}$, la normalisation $\int f_X = 1$ implique $c < 0$, i.e. X est une v.a. gaussienne. Cet argument se généralise sans peine à plus de deux v.a. C'est essentiellement sous ces hypothèses que Maxwell (1831-1879) a obtenu sa première dérivation de ce qui est connu sous le nom de *loi de Maxwell* pour la distribution des vitesses dans un gaz à l'équilibre.

1) A l'équilibre (état stable) les différentes vitesses se produisent avec des fréquences bien déterminées : il y a une loi de probabilité des vitesses qui est la même dans toutes les directions. On applique ici un principe d'homogénéité.
2) Cette loi est uniquement une fonction de la vitesse et par isotropie c'est une fonction de l'énergie cinétique.
3) Maxwell fait une hypothèse qui est moins évidente : les composantes des vitesses selon les axes rectangulaires sont des v.a. indépendantes.

La critique de cette dernière hypothèse a conduit Maxwell à donner d'autres dérivations de cette loi fondamentale de la distribution des vitesses d'un gaz à l'équilibre. □

Proposition 6.2 *Soit les v.a.* $Y_i := \varphi_i \circ X_i$, $i = 1, \ldots, k$, *où* $\varphi_i \colon \mathbb{R} \to \mathbb{R}$. *Si les v.a.* $X_1, \ldots, X_k$ *sont indépendantes, alors les v.a.* $Y_1, \ldots, Y_k$ *sont indépendantes.*

Preuve $\{Y_i \leq t_i\} = \{X_i \in B_i\}$ où

$$B_i = \{x \in \mathbb{R} \colon \varphi_i(x) \leq t_i\}\,.$$

L'affirmation découle de la définition 6.4 et de la proposition 6.1. □

Les v.a. sont souvent spécifiées seulement par leurs lois. Si des v.a. concernent la *même expérience aléatoire*, il est essentiel de les définir sur le *même espace de probabilité* $(\Omega, \mathcal{F}, P)$ décrivant cette expérience. Lorsque les v.a. sont indépendantes cela ne pose pas de problème, car il existe toujours un espace de probabilité sur lequel des copies ou des représentations de ces v.a. peuvent être définies simultanément, puisque la loi conjointe est dans ce cas le produit des lois marginales (il suffit de prendre la représentation canonique de ces v.a.). L'exemple suivant montre que l'existence de v.a. sur un même espace de probabilité ne va pas de soi. Soit X, Y, Z des v.a. prenant les valeurs ± 1, de même loi, $Y \overset{\mathcal{L}}{=} X$ et $Z \overset{\mathcal{L}}{=} X$, de sorte que

$$P(X = 1) = P(X = -1) = \frac{1}{2}\,.$$

Si l'on impose en plus

$$P(X = Y) = \frac{3}{4}\,, \tag{6.14}$$

on peut facilement construire des v.a. X_1 et Y_1 sur le même espace de probabilité qui vérifient (6.14) et telles que $X_1 \overset{\mathcal{L}}{=} X$, $Y_1 \overset{\mathcal{L}}{=} Y$. Il suffit de poser

$$P(X_1 = -1, Y_1 = -1) = P(X_1 = 1, Y_1 = 1) = \frac{3}{8}$$
$$P(X_1 = -1, Y_1 = 1) = P(X_1 = 1, Y_1 = -1) = \frac{1}{8}\,.$$

De même, si l'on impose

$$P(X = Z) = \frac{3}{4}\,, \tag{6.15}$$

on peut construire des v.a. X_2 et Z_2 sur le même espace de probabilité qui vérifient (6.15) et telles que $X_2 \stackrel{\mathcal{L}}{=} X$ et $Z_2 \stackrel{\mathcal{L}}{=} Z$. Si l'on impose

$$P(Y = Z) = \frac{1}{4}\,, \tag{6.16}$$

on peut construire des v.a. Y_3 et Z_3 sur le même espace de probabilité,

$$P(Y_3 = -1, Z_3 = -1) = P(Y_3 = 1, Z_3 = 1) = \frac{1}{8}$$
$$P(Y_3 = -1, Z_3 = 1) = P(Y_3 = 1, Z_3 = -1) = \frac{3}{8}\,,$$

qui vérifient (6.16) et telles que $Y_3 \stackrel{\mathcal{L}}{=} Y$ et $Z_3 \stackrel{\mathcal{L}}{=} Z$. *Mais on ne peut pas construire des v.a. X, Y, Z sur le même espace de probabilité telles que*

$$(X, Y) \stackrel{\mathcal{L}}{=} (X_1, Y_1)\,,\ (X, Z) \stackrel{\mathcal{L}}{=} (X_2, Z_2)\,,\ (Y, Z) \stackrel{\mathcal{L}}{=} (Y_3, Z_3)\,.$$

Si c'était possible,

$$\begin{aligned}
\frac{1}{4} &= P(Y = Z) \geq P(Y = Z, X = Z) = P(Y = X, X = Z) \\
&= P(Y = X) - P(Y = X, X \neq Z) \geq P(Y = X) - P(X \neq Z) \\
&= P(Y = X) - \big(1 - P(X = Z)\big) = \frac{1}{2}\,.
\end{aligned}$$

Remarque 6.4 Cet exemple est inspiré d'une expérience fondamentale de la mécanique quantique mettant en évidence le phénomène d'*intrication quantique*. L'intrication quantique est un phénomème dans lequel l'état quantique de deux objets doit être décrit globalement, sans pouvoir séparer un objet de l'autre, bien qu'ils puissent être spatialement séparés. Voir H. Thorisson *Coupling, Stationarity and Regeneration*, Springer, New York (2000) pp. 27-31 pour une discussion de cet exemple et pour la signification de la non-existence des v.a. X, Y, Z. Voir le livre *Einstein et les révolutions quantiques*, A. Aspect, CNRS Editions (2019), pour une introduction à l'intrication quantique. □

6.5 Somme de variables aléatoires indépendantes

Proposition 6.3 *Soit X de densité f_X et Y de densité f_Y deux v.a. réelles indépendantes. Alors la densité de $X + Y$ est donnée par le* produit de convolution

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} dx f_X(x) f_Y(u - x) = \int_{-\infty}^{\infty} dy f_X(u - y) f_Y(y)\,.$$

Preuve Par définition

$$P(X+Y\le u)=\iint_{\{(x,y):\ x+y\le u\}} f_{X,Y}(x,y)\,dxdy=\int_{-\infty}^{\infty}dx f_X(x)\int_{-\infty}^{u-x}dy f_Y(y)\,.$$

La densité s'obtient en dérivant $P(X+Y\le u)$ par rapport à u,

$$f_{X+Y}(u)=\int_{-\infty}^{\infty}dx f_X(x)f_Y(u-x)\,.$$

□

Proposition 6.4 *Si X et Y sont* indépendantes *:*

a) $X\sim N(m_1,\sigma_1^2)$ *et* $Y\sim N(m_2,\sigma_2^2)\implies X+Y\sim N(m_1+m_2,\sigma_1^2+\sigma_2^2)$ *;*

b) $X\sim \pi_{\lambda_1}$ *et* $Y\sim\pi_{\lambda_2}\implies X+Y\sim\pi_{\lambda_1+\lambda_2}$ *;*

c) $X\sim\mathcal{B}_i(n,p)$ *et* $Y\sim\mathcal{B}_i(m,p)\implies X+Y\sim\mathcal{B}_i(n+m,p)$ *;*

d) $X\sim\gamma_{s,\lambda}$ *et* $Y\sim\gamma_{t,\lambda}\implies X+Y\sim\gamma_{s+t,\lambda}$.

e) Si X et Y sont des v.a. de Cauchy de paramètre a, respectivement b, alors $X+Y$ est une v.a. de Cauchy de paramètre $a+b$.

Preuve On démontre le cas d). Les cas a), b) et c) sont laissés en exercice. Le cas e) est un calcul assez long à partir de la proposition 6.3. Dans le cas d) la densité f_{X+Y} de $X+Y$ est égale à (voir (5.3))

$$\begin{aligned}
&\frac{1}{\Gamma(s)\Gamma(t)}\int_{-\infty}^{\infty}\lambda\mathrm{e}^{-\lambda(u-y)}(\lambda(u-y))^{s-1}I_{\mathbb{R}^+}(u-y)\lambda\mathrm{e}^{-\lambda y}(\lambda y)^{t-1}I_{\mathbb{R}^+}(y)\,dy=\\
&\frac{1}{\Gamma(s)\Gamma(t)}\mathrm{e}^{-\lambda u}\int_0^u\lambda^2(\lambda u)^{s+t-2}\Big(1-\frac{y}{u}\Big)^{s-1}\Big(\frac{y}{u}\Big)^{t-1}dy=\\
&\frac{\lambda\mathrm{e}^{-\lambda u}(\lambda u)^{s+t-1}}{\Gamma(s)\Gamma(t)}\underbrace{\int_0^1(1-x)^{s-1}x^{t-1}\,dx}_{B(s,t)}=\frac{\lambda\mathrm{e}^{-\lambda u}(\lambda u)^{s+t-1}}{\Gamma(s+t)}\quad\text{si } u>0.
\end{aligned}$$

□

Exemple 6.14 Soit n v.a. indépendantes de loi exponentielle de paramètre $\lambda>0$ (loi $\gamma_{1,\lambda}$),

$$f_X(t):=\lambda\mathrm{e}^{-\lambda t}I_{\mathbb{R}^+}(t)\,.$$

La loi de $T_n=X_1+\cdots+X_n$ est une loi gamma de paramètres n et λ,

$$f_{T_n}(x)=\lambda^n\frac{x^{n-1}}{(n-1)!}\mathrm{e}^{-\lambda x}I_{\mathbb{R}^+}(x)\,.$$

La loi de T_n est aussi appelée *loi d'Erlang* (1878-1929).

Soit une file d'attente avec un temps de service modélisé par une v.a. exponentielle de paramètre λ ; X_i est le temps nécessaire pour servir un client. Le

premier client arrive à $t = 0$, le deuxième à $t = X_1$, le troisième à $t = X_1 + X_2$ etc. Pour $t > 0$ fixé,

$$N_t := \#\text{clients servis jusqu'au temps } t = \max\{k\colon T_k \le t\}\,.$$

Calcul de la loi de N_t. Si $t > 0$,

$$P(N_t = n) = P(T_n \le t, T_{n+1} > t)\,.$$

Comme $\{T_{n+1} \le t\} \subset \{T_n \le t\}$,

$$P(N_t = n) = P(T_n \le t) - P(T_{n+1} \le t)\,.$$

Si $n > 0$,

$$\begin{aligned} P(T_n \le t) &= \frac{\lambda^n}{(n-1)!}\int_0^t \mathrm{e}^{-\lambda s} s^{n-1}\,ds \\ &= \frac{\lambda^n}{(n-1)!}\Big(\frac{t^n}{n}\mathrm{e}^{-\lambda t} + \frac{\lambda}{n}\int_0^t \mathrm{e}^{-\lambda s} s^n\,ds\Big) \\ &= \frac{(\lambda t)^n}{n!}\mathrm{e}^{-\lambda t} + P(T_{n+1} \le t)\,. \end{aligned}$$

Si $n = 0$,

$$P(N_t = 0) = P(T_1 > t) = 1 - P(T_1 \le t) = 1 - (1 - \mathrm{e}^{-\lambda t}) = \mathrm{e}^{-\lambda t}\,.$$

La loi de N_t est une loi de Poisson de paramètre λt. □

6.6 La loi binomiale et la loi de Poisson

La loi de Poisson π_λ peut être obtenue comme le cas limite de la loi binomiale $\mathcal{B}_i(n,p)$, lorsque la probabilité p d'un "succès" tend vers zéro, mais que simultanément n diverge de telle manière que $pn = \lambda$. Cette limite est appelée la *limite des événements rares* puisque $p \to 0$. A partir du lemme I.1 on montre aisément

$$\lim_{\substack{n\to\infty \\ np=\lambda}} \binom{n}{k} p^k (1-p)^{n-k} = \pi_\lambda(k) \qquad \forall\, k = 0, 1, \ldots\,.$$

On peut démontrer un résultat plus fort. On considère n événements indépendants $A_1, \ldots, A_n$, $P(A_i) \equiv p_i$, $0 < p_i < 1$. On pose $X_i := I_{A_i}$ et on s'intéresse au nombre des événements, parmi ceux de la collection $\{A_1, \ldots, A_n\}$, qui sont réalisés lors de l'expérience aléatoire. Ce nombre est donné par la v.a.

$$S_n := X_1 + \cdots + X_n\,.$$

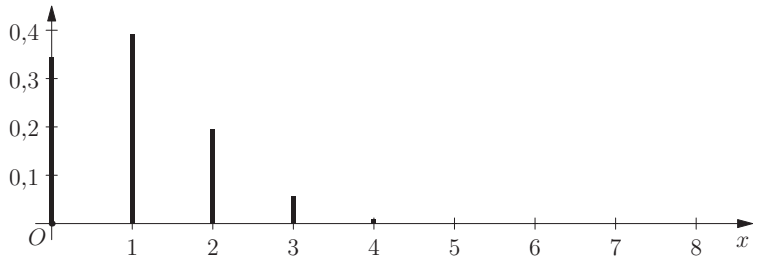


Figure 6.10 – Loi binomiale $\mathcal{B}_i(8,\ 0{,}125)$. Comparer avec la loi de Poisson de paramètre $\lambda = 1$ de la figure 6.3.

Théorème 6.1 *Soit $Z_n \sim \pi_\lambda$ une v.a. de Poisson avec $\lambda := p_1 + \cdots + p_n$. Alors pour tout $J \subset \{0, 1, 2, \ldots\}$ on a*

$$\big|P(S_n \in J) - P(Z_n \in J)\big| \leq \sum_{i=1}^{n} p_i^2 \,.$$

Remarque 6.5 L'estimation est uniforme en J. On peut même remplacer $\sum_{i=1}^n p_i^2$ par $\max_i p_i$. □

Exemple 6.15 On suppose que $p_i \equiv p$ pour tout i. Dans ce cas $S_n \sim \mathcal{B}_i(n, p)$. Si np^2 est petit, par exemple $p = q/n$ et n est grand, alors la loi binomiale est bien approximée par la loi de Poisson π_q. □

Lemme 6.1 *Soit V et W deux v.a. définies sur le même espace de probabilité. Alors*

$$\forall B \in \mathcal{B}(\mathbb{R}): \quad |P(V \in B) - P(W \in B)| \leq P(V \neq W) \,.$$

Preuve On peut supposer $P(V \in B) \geq P(W \in B)$. Les sous-ensembles $\{W \in B\}$ et $\{W \notin B\}$ forment une partition de Ω :

$$P(V \in B) = P(V \in B, W \in B) + P(V \in B, W \notin B) \,.$$

Par conséquent

$$\begin{aligned} P(V \in B) - P(W \in B) &\leq P(V \in B) - P(W \in B, V \in B) \\ &\leq P(V \in B, W \notin B) \\ &\leq P(V \neq W) \,. \end{aligned}$$

□

Preuve du théorème 6.1 On utilise les résultats de la section 2.2 et on considère d'abord le cas $n = 1$. Sur l'espace de probabilité d'un GNA on construit une v.a. Y_p de Bernoulli de paramètre p et une v.a. $X_p \sim \pi_p$. Par construction

$$Y_p = 0 \text{ sur } (0, 1-p] \quad \text{et} \quad Y_p = 1 \text{ sur } (1-p, 1)\,;$$

$$X_p = 0 \text{ sur } (0, \mathrm{e}^{-p}] \quad \text{et} \quad X_p = 1 \text{ sur } (\mathrm{e}^{-p}, (1+p)\mathrm{e}^{-p}]\,,$$

sinon $X_p > 1$. De l'inégalité élémentaire $\mathrm{e}^{-p} \geq 1 - p$ on obtient l'estimation

$$P(X_p \neq Y_p) = \big(\mathrm{e}^{-p} - (1-p)\big) + \big(1 - (1+p)\mathrm{e}^{-p}\big) = p(1 - \mathrm{e}^{-p}) \leq p^2\,.$$

Dans le cas général on fait cette construction pour chaque copie de $(0,1)$ du produit cartésien $\Omega = (0,1) \times \cdots \times (0,1)$ muni de la mesure de probabilité uniforme : pour la $i^{\text{ème}}$ copie on construit $V_i \overset{\mathcal{L}}{=} Y_{p_i}$ et $W_i \overset{\mathcal{L}}{=} X_{p_i}$,

$$V_i(\omega) := Y_{p_i}(\omega_i) \quad \text{et} \quad W_i(\omega) := X_{p_i}(\omega_i)\,.$$

On obtient ainsi une famille de v.a. indépendantes $V_1, \ldots, V_n$ et une autre famille de v.a. indépendantes $W_1, \ldots, W_n$. Par conséquent $\sum_i V_i$ est une copie de S_n et $\sum_i W_i$ est une copie de Z_n (proposition 6.4).

$$\begin{aligned}\big|P(S_n \in J) - P(Z_n \in J)\big| &\leq P(S_n \neq Z_n) \\ &\leq P\Big(\bigcup_{j=1}^{n} \{V_j \neq W_j\}\Big) \leq \sum_{j=1}^{n} p_j^2\,.\end{aligned}$$

□

Exemple 6.16 On considère le cas du rangement de n boules distinguables dans M boîtes $a_1, \ldots, a_M$ de la section 3.1. La probabilité qu'une boule soit rangée dans la première boîte est $1/M$; par conséquent la probabilité que la première boîte reste vide est

$$\Big(1 - \frac{1}{M}\Big)^n \approx \mathrm{e}^{-n/M}\,.$$

Par symétrie ce résultat s'applique à n'importe quelle boîte. La probabilité que la première boîte contienne k boules est donnée par

$$\binom{n}{k}\Big(\frac{1}{M}\Big)^k\Big(1 - \frac{1}{M}\Big)^{n-k} = \frac{1}{k!}\frac{[n]_k}{M^k}\Big(1 - \frac{1}{M}\Big)^{n-k} \approx \frac{(n/M)^k}{k!}\mathrm{e}^{-n/M}\,.$$

A gauche on a une loi binomiale $\mathcal{B}_i(n, 1/M)$ et à droite une loi de Poisson $\pi_{n/M}$. Le théorème 6.1 indique que l'approximation de la loi binomiale par la loi de Poisson est bonne si n/M^2 est petit. □

Si dans la situation de l'exemple 6.16 le nombre de boules dans la boîte a_i est donné par la v.a. X_i, la loi conjointe de ces v.a. est

$$\begin{aligned}&P(X_1 = k_1, \ldots, X_M = k_M) = \frac{n!}{k_1! \cdots k_M!}\,\frac{1}{M^n} && \text{si } k_1 + \cdots + k_M = n \\ &P(X_1 = k_1, \ldots, X_M = k_M) = 0 && \text{si } k_1 + \cdots + k_M \neq n\,.\end{aligned}$$

Les v.a. X_i ne sont pas indépendantes. A côté de ces v.a. on considère M v.a. indépendantes Y_i de loi de Poisson de paramètre n/M.

Proposition 6.5 *Dans la situation décrite ci-dessus,*

$$P(X_1 = k_1, \dots, X_M = k_M) = P(Y_1 = k_1, \dots, Y_M = k_M | Y_1 + \cdots + Y_M = n)\,.$$

Preuve La v.a. $Z := Y_1 + \cdots + Y_M$ est une v.a. de Poisson de paramètre n.

$$\begin{aligned} P(Y_1 = k_1, \dots, Y_M = k_M | Y_1 + \cdots + Y_M = n) & \\ &= \frac{P(Y_1 = k_1, \dots, Y_M = k_M, Z = n)}{P(Z = n)} \\ (k_1 + \cdots + k_M = n) \qquad &= \frac{\prod_{j=1}^{M} \mathrm{e}^{-n/M} (n/M)^{k_j} (k_j!)^{-1}}{\mathrm{e}^{-n} n^n (n!)^{-1}} \\ &= \frac{n!}{k_1! \cdots k_M!} \, \frac{1}{M^n}\,. \end{aligned}$$

□

6.7 Exercices

Exercice 6.1 Démontrer les cas b) et c) de la proposition 6.4.

Exercice 6.2 Soit X une v.a. gaussienne de loi $N(0,1)$. Calculer la densité des v.a. $|X|$ et X^2.

Exercice 6.3 On considère une suite de v.a. X_n, $n \geq 1$, indépendantes, et de loi de Bernoulli de paramètre p. On définit deux v.a. S_n, $n \in \mathbb{N}$, et T_r, $r \in \mathbb{N}$ par

$$S_n := \sum_{j=1}^{n} X_j \quad , \quad T_r := \min\{\, m : S_m \geq r \,\}\,.$$

Dessiner le graphe de S_n en fonction de n pour une réalisation donnée de l'expérience. Vérifier l'identité

$$P(T_r > n) = P(S_n < r)\,.$$

Déterminer la distribution de ces deux v.a. ; montrer en particulier qu'on a $P(T_r < \infty) = 1$.

Exercice 6.4 On considère deux urnes U_0 et U_1 contenant chacune N boules. A chaque unité de temps on choisit une des urnes au hasard et on retire une boule de l'urne choisie. On n'enregistre pas quelle est l'urne choisie. Lorsqu'on découvre pour la première fois que l'une des urnes est vide, quelle est la probabilité que l'autre urne contienne k boules ?

Indication : utiliser les résultats de l'exercice 6.3, en définissant les v.a. $X_k = 1$, si lors du $k^{\text{ième}}$ tirage on choisit l'urne U_0, et $X_k = 0$, si lors du $k^{\text{ième}}$ tirage on choisit l'urne U_1.

Exercice 6.5 a) On lance trois fois une pièce de monnaie équilibrée. La v.a. X donne le nombre de fois qu'on a obtenu Face lors des deux premiers lancers et la v.a. Y donne le nombre de fois qu'on a obtenu Face lors des deux derniers lancers. Déterminer les lois de X, Y et la loi conjointe de X et Y.
b) On tire au hasard deux nombres dans $\{-1, 1\}$. La v.a. X donne la somme de ces nombres et la v.a. Y donne le produit de ces nombres. Déterminer les lois de X, Y et la loi conjointe de X et Y.

Exercice 6.6 X et Y sont deux v.a. indépendantes et uniformément distribuées sur l'intervalle $[0, L]$. Calculer la fonction de répartition de la v.a. $|X-Y|$. Calculer la densité de probabilité de cette v.a..

Exercice 6.7 La densité de probabilité de la loi conjointe de deux v.a. X et Y est donnée par

$$f(x,y) := \begin{cases} 2\mathrm{e}^{-x}\mathrm{e}^{-2y} & \text{si } 0 < x < \infty \text{ et } 0 < y < \infty \\ 0 & \text{sinon.} \end{cases}$$

1) Calculer $P(X \geq 1, Y < 1)$.
2) Calculer $P(X < Y)$.
3) Trouver la densité de la v.a. X/Y.
Indication : calculer la fonction de répartition de X/Y.

Exercice 6.8 Est-ce que les v.a. X et Y de l'exercice 6.7 sont indépendantes ? Même question si la loi conjointe de X et Y est donnée par

$$f(x,y) = \begin{cases} 24xy & \text{si } 0 < x < 1,\ 0 < y < 1,\ 0 < x + y < 1 \\ 0 & \text{sinon.} \end{cases}$$

Dans les deux cas justifier votre réponse.

Exercice 6.9 On désigne l'ensemble des mesures de probabilité sur $\{1, \ldots, k\}$ par

$$\mathcal{M} = \left\{ \mathbf{p} = (p_1, \ldots, p_k) \colon p_i \geq 0 \text{ et } \sum_{i=1}^{k} p_i = 1 \right\}.$$

Si $\mathbf{p} \in \mathcal{M}$ et $\mathbf{q} \in \mathcal{M}$ on pose $H(\mathbf{p}) = -\sum_{j=1}^{k} p_j \ln p_j$ et

$$D(\mathbf{p}|\mathbf{q}) := \sum_{j=1}^{k} p_j \ln \frac{p_j}{q_j} \quad \left(\text{convention : } 0 \ln \frac{0}{q} = 0 \text{ et } p \ln \frac{p}{0} = \infty \right).$$

La quantité $D(\mathbf{p}|\mathbf{q})$ est l'*entropie relative de* $\mathbf{q}$ *par rapport à* $\mathbf{p}$ ou *divergence de Kullback-Leibler de* $\mathbf{q}$ *par rapport à* $\mathbf{p}$ (Kullback (1907-1994), Leibler (1915-2003)).
a) Montrer que $D(\mathbf{p}|\mathbf{q}) \geq 0$.
Indication : $-\ln$ est convexe.
b) Montrer à l'aide de a) que $H(\mathbf{p})$ est maximale si $p_i = 1/k$ pour tout i ; calculer le maximum de l'entropie.

Exercice 6.10 Utiliser l'approximation de la loi binomiale par une loi de Poisson pour résoudre le problème suivant. On fabrique des objets de façon indépendante et la probabilité qu'un objet soit défectueux lors de sa fabrication est $p = 0{,}015$. Estimer le nombre minimal d'objets qu'il faut fabriquer pour que la probabilité d'avoir au moins 100 objets non défectueux soit plus grande ou égale à 0,8 ?

Chapitre 7

Espérance d'une variable aléatoire

Pour mettre en évidence certaines propriétés de la loi d'une v.a. on définit des paramètres réels dont les plus importants sont :

1) *L'espérance de X* ou *moyenne de X* qui est un *paramètre de position*. L'espérance est notée $\mathbb{E}(X)$ en théorie des probabilités et souvent $\langle X \rangle$ en physique. On utilise dans ce livre $\mathbb{E}(X)$.
2) *La variance de X*, notée $\mathrm{Var}X$, qui est un *paramètre de dispersion*. La *déviation standard* ou *écart-type* est $DS(X) := \sqrt{\mathrm{Var}X} \equiv \sigma(X)$.

Ces paramètres ne sont pas nécessairement définis pour toutes les v.a. ! Par contre la médiane est un paramètre de position qui existe toujours, mais qui n'est pas nécessairement défini univoquement. Il est différent de celui de l'espérance. Une *médiane de X* est un nombre réel m tel que

$$P(X \leq m) \geq 1/2 \quad \text{et} \quad P(X \geq m) \geq 1/2\,.$$

Par exemple, si $X \sim \gamma_{1,\lambda}$, la médiane est donnée par la condition

$$\lambda \int_0^m \mathrm{e}^{-t\lambda}\, dt = \frac{1}{2} \implies m = \frac{\ln 2}{\lambda}\,.$$

7.1 Définition de l'espérance

La définition de l'espérance est intuitive dans le cas des v.a. discrètes. C'est la moyenne pondérée des valeurs que peut prendre la v.a., le poids de chaque valeur étant la probabilité que la v.a. prenne cette valeur :

$$\sum_{x \in D} x P(X = x)\,.$$

L'ensemble D des valeurs de la v.a. est fini ou dénombrable. Pour que l'expression ait un sens bien défini, on suppose que la somme ci-dessus est absolument convergente de sorte qu'elle ne dépende pas d'un réarrangement de ses termes (proposition I.5).

Définition 7.1 *L'*espérance (ou moyenne) d'une v.a. discrète X *est définie si et seulement si*

$$\sum_{x\in D} |x| P(X=x) < \infty\,.$$

Lorsque c'est le cas,

$$\mathbb{E}(X) \equiv \langle X\rangle := \sum_{x\in D} xP(X=x)\,.$$

Si X est une v.a. définie sur un espace de probabilité discret,

$$\begin{aligned}\sum_{x\in D} xP(X=x) &= \sum_{x\in D} x\Big(\sum_{\substack{\omega\in\Omega:\\ X(\omega)=x}} P(\omega)\Big)\\ &= \sum_{x\in D}\Big(\sum_{\substack{\omega\in\Omega:\\ X(\omega)=x}} X(\omega)P(\omega)\Big)\\ &= \sum_{\omega\in\Omega} X(\omega)P(\omega)\,.\end{aligned}$$

Exemples 7.1 a) Soit $X = I_A$ et $P(A) = p$ (v.a. de Bernoulli).

$$\mathbb{E}(X) = X(0)(1-p) + X(1)p = p\,.$$

b) Soit X une v.a. de Poisson π_λ.

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k\geq 0} k\frac{\mathrm{e}^{-\lambda}\lambda^k}{k!} = \lambda\sum_{k\geq 1}\frac{\mathrm{e}^{-\lambda}\lambda^{(k-1)}}{(k-1)!}\\ &= \lambda\mathrm{e}^{-\lambda}\sum_{k\geq 0}\frac{\lambda^k}{k!} = \lambda\,.\end{aligned}$$

c) Soit Y une v.a. binomiale de loi $\mathcal{B}_i(n,p)$.

$$\begin{aligned}\mathbb{E}(Y) &= \sum_{k=0}^{n} kP(Y=k)\\ &= \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k}\\ &= \sum_{k=1}^{n}\frac{n(n-1)!}{(k-1)!(n-k)!}p^k(1-p)^{n-k}\\ &= np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{(n-1)-(k-1)}\\ &= np\sum_{j=0}^{n-1}\binom{n-1}{j}p^j(1-p)^{(n-1)-j} = np\,.\end{aligned}$$

Une autre manière d'obtenir ce résultat est exposée dans l'exemple 7.5.

d) Dans la remarque 6.2 la quantité $H(X)$ de l'équation (6.3), qui est l'entropie de la v.a. X, est égale à l'espérance de l'incertitude sur la valeur prise par la v.a. X. □

Pour définir l'espérance d'une v.a. continue on se ramène au cas des v.a. discrètes. Ceci nécessite un passage à la limite (voir section 7.2). Le résultat final peut être mémorisé par « on remplace des sommes par des intégrales ».

Définition 7.2 L'espérance (ou moyenne) d'une v.a. continue X *est définie si et seulement si*

$$\int_{\mathbb{R}} |s| f_X(s)\, ds < \infty\,.$$

Lorsque c'est le cas

$$\mathbb{E}(X) \equiv \langle X \rangle := \int_{\mathbb{R}} s f_X(s)\, ds\,.$$

Exemple 7.2 a) Soit X une v.a. uniforme sur $(0,1)$;

$$\mathbb{E}(X) = \int_0^1 t dt = \frac{1}{2}\,.$$

b) Si X est une v.a. de Cauchy, l'espérance $\mathbb{E}(X)$ n'existe pas car

$$\int_{-\infty}^{\infty} |t| \frac{a}{\pi(a^2+t^2)}\, dt = \infty\,.$$

c) Un autre exemple est celui d'une v.a. de Pareto Y de paramètre $\alpha \leq 1$; dans ce cas la v.a. Y n'a pas d'espérance. Dans l'exemple 6.5, $\alpha = \frac{k_B T}{E_0} \leq 1$ si T est petit ou E_0 est grand. □

Théorème 7.1 *Une v.a. X possède une espérance si et seulement si*

$$\int_0^\infty P(X > t)\, dt < \infty \quad \textit{et} \quad \int_0^\infty P(X < -t)\, dt < \infty\,.$$

Si ces conditions sont vérifiées,

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

Le théorème 7.1 est vrai en toute généralité. Dans ce théorème on peut remplacer $P(X > t)$ par $P(X \geq t)$ car

$$\{t\colon P(X > t) \neq P(X \geq t)\}$$

est l'ensemble des sauts de F_X qui est au plus dénombrable ; par conséquent les intégrales $\int P(X > t)\, dt$ et $\int P(X \geq t)\, dt$ sont égales.

Toute v.a. X peut être décomposée en $X = X^+ - X^-$ où X^+ et X^- sont des v.a. non négatives appelées respectivement *partie positive de* X et *partie négative de* X

$$X^+(\omega) := \begin{cases} X(\omega) & \text{si } X(\omega) \geq 0 \\ 0 & \text{sinon} \end{cases} \qquad X^-(\omega) := \begin{cases} |X(\omega)| & \text{si } X(\omega) \leq 0 \\ 0 & \text{sinon.} \end{cases}$$

Le théorème 7.1 indique qu'une v.a. X possède une espérance si et seulement si les v.a. X^+ et X^- possèdent des espérances, et dans ce cas

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)\,.$$

Une preuve du théorème 7.1 est donnée dans la section 7.2. Dans le cas continu, si $X \geq 0$,

$$\begin{aligned} \int_0^\infty P(X > t)\, dt &= \int_0^\infty dt \int_t^\infty ds f_X(s) \\ &= \int_0^\infty ds \int_0^s dt f_X(s) = \int_0^\infty ds f_X(s)\, s = \mathbb{E}(X)\,. \end{aligned}$$

Si l'on est intéressé essentiellement à l'aspect calculatoire, on peut directement passer aux énoncés des théorèmes essentiels 7.4 et 7.5. Cependant, pour bien comprendre la démonstration du théorème 7.5, les développements de la section suivante sont importants.

7.2 Définition de l'espérance, cas général

Dans cette section on donne la définition de l'espérance d'une v.a. X dans le cadre général d'un espace de probabilité $(\Omega, \mathcal{F}, P)$. Le point de départ est de montrer que chaque v.a. X définie sur $(\Omega, \mathcal{F}, P)$ est la limite de v.a. discrètes (lemme 7.1). L'espérance de X est définie comme la limite des espérances de ces v.a. discrètes. L'existence de cette limite est une conséquence la proposition I.4. On établit en particulier que toute v.a. bornée possède une espérance, ainsi que les théorèmes importants 7.2 et 7.3.

Lemme 7.1 *Soit $(\Omega, \mathcal{F}, P)$ un espace de probabilité et X une v.a. réelle.*

1) Toute v.a. X non négative est la limite (ponctuelle) d'une suite croissante de v.a. discrètes. Si X est bornée ($\sup_\omega X(\omega) < \infty$), la suite peut être choisie de sorte que la convergence soit uniforme.
2) Si X_n, $n \geq 1$, est une suite de v.a. réelles qui converge (ponctuellement) vers X,

$$X(\omega) := \lim_{n\to\infty} X_n(\omega) \quad \forall\, \omega\,,$$

alors X est une v.a. réelle.

Preuve Pour montrer 1) on définit pour tout $n \geq 1$ (voir figure 7.1)

$$Y_n := 0 \cdot I_{\{0 \leq X \leq 2^{-n}\}} + \sum_{k=1}^{n2^n-1} \frac{k}{2^n} I_{\{k\,2^{-n} < X \leq (k+1)2^{-n}\}} + n\, I_{\{X>n\}}\,.$$

Par définition Y_n est une v.a. discrète et

$$Y_n(\omega) \leq X(\omega) \leq Y_n(\omega) + 2^{-n} \quad \text{si } X(\omega) \leq n\,.$$

Ces inégalités démontrent l'affirmation 1).

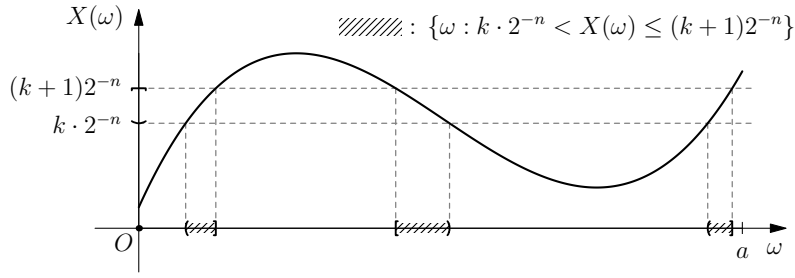


Figure 7.1 Point-clé de la construction d'une v.a. discrète approximant une v.a. positive X. Sur la partie hachurée la v.a. Y_n vaut $k2^{-n}$.

2) Soit X_n, $n \geq 1$, une suite de v.a. qui converge vers X. Soit $t \in \mathbb{R}$; $\lim_n X_n(\omega) = X(\omega) \leq t$ implique que pour tout $k \in \mathbb{N}$ il existe $n_k(\omega)$ tel que

$$X_m(\omega) \leq t + \frac{1}{k} \quad \text{si } m \geq n_k\,.$$

Pour tout $k \in \mathbb{N}$

$$\{\omega\colon X(\omega) \leq t\} \subset \bigcup_{n\geq 1} \bigcap_{m \geq n} \left\{X_m(\omega) \leq t + \frac{1}{k}\right\} \equiv E_k \in \mathcal{F}\,.$$

Par conséquent (la limite existe par hypothèse)

$$\omega \in \bigcap_{k\geq 1} E_k \equiv E \iff \lim_{n\to\infty} X_n(\omega) = X(\omega) \leq t\,.$$

Cela établit $\{X \leq t\} = E \in \mathcal{F}$. □

Pour définir l'espérance $\mathbb{E}(X)$ on définit d'abord l'espérance pour les v.a non négatives, puis on pose

$$\mathbb{E}(X) := \mathbb{E}(X^+) - \mathbb{E}(X^-)\,.$$

Pour définir $\mathbb{E}(X^+)$ on utilise le lemme 7.1 point 1) afin de se ramener au cas des v.a. discrètes. Le calcul élémentaire suivant est le point-clé. On pose

$I_k := I_{\{k\,2^{-n}<X\le\,(k+1)2^{-n}\}}$ et $I_{n2^n} := I_{\{X>n\}}$.

$$\begin{aligned}\mathbb{E}(Y_n) &= \sum_{k=1}^{n\,2^n-1} (k\,2^{-n})P(X\in I_k) + n\,P(X\in I_{n2^n})\\ &= \sum_{k=1}^{n\,2^n-1} (k\,2^{-n})\Big(P(X>k\,2^{-n}) - P(X>(k+1)\,2^{-n})\Big)\\ &\quad +\ n\,P(X\in I_{n2^n})\\ &= \sum_{k=1}^{n\,2^n} 2^{-n}P(X>k\,2^{-n})\,.\end{aligned}$$

La dernière somme est une somme de Riemann sur $[0,n]$ de la fonction monotone décroissante $t\mapsto P(X>t)$.

Pour la clarté de l'exposé, on considère d'abord les v.a. non négatives qui sont bornées. Le cas des v.a. non bornées est un peu plus technique. Soit M tel que $0\le X(\omega)\le M$. Dans ce cas la fonction $t\mapsto P(X>t)$ est Riemann intégrable sur $[0,M]$ et $P(X>t)=0$ dès que $t\ge M$ (proposition I.2). Par conséquent

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}(Y_n) &= \lim_{n\to\infty}\sum_{k=1}^{n\,2^n} 2^{-n}P(X>k\,2^{-n})\\ &= \int_0^M P(X>t)\,dt = \int_0^\infty P(X>t)\,dt\,.\end{aligned}$$

(Dans la somme de Riemann ci-dessus les termes pour $k>M2^n$ sont nuls puisque la v.a. est bornée). Pour des v.a. bornées on pose

$$\mathbb{E}(X) := \lim_n \mathbb{E}(Y_n)\,.$$

Cette limite est toujours finie. Toute v.a. bornée possède une espérance.

Lorsque X est une v.a. continue on peut exprimer $\mathbb{E}(X)$ avec la densité de probabilité f_X. Soit $n\ge M$;

$$\mathbb{E}(Y_n) = \sum_{k=0}^{n\,2^n-1} (k\,2^{-n})P(X\in I_k) = \sum_{k=0}^{n\,2^n-1} (k\,2^{-n})\int_{I_k} f_X(s)\,ds\,. \tag{7.1}$$

Si $s\in I_k$, alors $|s-k\,2^{-n}|\le 2^{-n}$; en utilisant $\int f_X(s)\,ds=1$ on obtient

$$\begin{aligned}\sum_{k=0}^{n\,2^n-1} (k\,2^{-n})\int_{I_k} f_X(s)\,ds &\le \sum_{k=0}^{n\,2^n-1}\int_{I_k} s f_X(s)\,ds \qquad (7.2)\\ &\le \sum_{k=0}^{n\,2^n-1} (k\,2^{-n})\int_{I_k} f_X(s)\,ds + 2^{-n}\,.\end{aligned}$$

Dans la limite $n \to \infty$

$$\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(Y_n) = \int_0^M s f_X(s)\, ds = \int_0^\infty s f_X(s)\, ds\,.$$

On retrouve l'expression de la définition 7.2.

Lorsque la v.a. X n'est pas bornée on se ramène au cas borné en utilisant

$$\lim_{M\to\infty} (X I_{\{X\le M\}})(\omega) = \lim_{M\to\infty} X(\omega) I_{\{X\le M\}}(\omega) = X(\omega) \quad \forall\, \omega\,.$$

La suite $\mathbb{E}(X I_{\{X\le M\}})$, $M \ge 1$, est monotone croissante (elle peut être divergente) ; le point important est le lemme 7.2.

Lemme 7.2 *Si X est une v.a. non négative,*

$$\lim_{M\to\infty} \mathbb{E}(X I_{\{X\le M\}}) = \int_0^\infty P(X > t)\, dt \le \infty\,.$$

Lorsque la limite $\lim_M \mathbb{E}(X I_{\{X\le M\}})$ est finie on définit l'espérance de X par

$$\mathbb{E}(X) := \lim_{M\to\infty} \mathbb{E}(X I_{\{X\le M\}})\,.$$

Preuve La suite $A_M := \{\omega\colon (X I_{\{X\le M\}})(\omega) > t\}$ est monotone croissante et

$$\bigcup_{M\ge 1} \{\omega\colon (X I_{\{X\le M\}})(\omega) > t\} = \{\omega\colon X(\omega) > t\}\,;$$

par conséquent $P(X I_{\{X\le M\}} > t) \le P(X > t)$ et, par continuité monotone de la mesure de probabilité P,

$$\lim_{M\to\infty} P(X I_{\{X\le M\}} > t) = P(X > t)\,. \tag{7.3}$$

On utilise la proposition I.4 pour conclure

$$\lim_{M\to\infty} \int_0^N P(X I_{\{X\le M\}} > t)\, dt = \int_0^N P(X > t)\, dt\,.$$

Par définition de l'intégrale de Riemann sur un domaine non borné

$$\begin{aligned}\int_0^\infty P(X > t)\, dt &= \lim_N \int_0^N P(X > t)\, dt \\ &\ge \lim_N \int_0^N P(X I_{\{X\le M\}} > t)\, dt = \mathbb{E}(X I_{\{X\le M\}})\,.\end{aligned}$$

Les résultats précédents montrent que

$$\begin{aligned}\int_0^\infty P(X>t)\,dt &\geq \lim_M \mathbb{E}(XI_{\{X\leq M\}}) \\ &\geq \lim_M \int_0^N P(XI_{\{X\leq M\}}>t)\,dt \\ &= \int_0^N P(X>t)\,dt \overset{N\to\infty}{\longrightarrow} \int_0^\infty P(X>t)\,dt\,.\end{aligned}$$

□

On a établit ainsi le théorème 7.1 en toute généralité. Les résultats qui suivent font partie des résultats importants de la théorie des probabilités.

Théorème 7.2 (Théorème de la convergence monotone) *Soit X_n, $n \geq 1$, une suite croissante de v.a. non négatives,*

$$X_n(\omega) \leq X_{n+1}(\omega) \quad \forall\,\omega \;\; et \;\; \forall\, n \geq 1\,.$$

Si $\lim_n X_n = X$ (ponctuellement), alors la v.a. X a une espérance si et seulement si $\lim_n \mathbb{E}(X_n) < \infty$, et dans ce cas $\mathbb{E}(X) = \lim_n \mathbb{E}_n(X)$.

Preuve Le point important de la preuve du lemme 7.2 est l'identité (7.3). De la même manière que précédemment

$$\begin{aligned}\int_0^\infty P(X>t)\,dt &\geq \lim_{n\to\infty} \int_0^\infty P(X_n>t)\,dt \\ &\geq \lim_{n\to\infty} \int_0^N P(X_n>t)\,dt \\ &= \int_0^N P(X>t)\,dt \overset{N\to\infty}{\longrightarrow} \int_0^\infty P(X>t)\,dt\,.\end{aligned}$$

□

Remarque 7.1 La preuve du lemme 7.2 et celle du théorème 7.2 sont la conséquence de la propriété de continuité monotone de la mesure de probabilité pour une suite d'événements $A_n \uparrow A$ (voir (7.3)). Inversement, si l'on pose $X_n = I_{A_n}$, alors $\mathbb{E}(X_n) = P(A_n)$ et le théorème 7.2 implique $\lim_n P(A_n) = P(A)$. La propriété de σ-additivité de la mesure de probabilité P est donc une condition nécessaire et suffisante pour le théorème 7.2 (voir remarque 2.1).

Lemme 7.3 (Lemme de Fatou) *Soit $X_n \geq 0$, $n \geq 1$, une suite de v.a. non négatives. On définit la v.a.*

$$Y(\omega) := \liminf_{n\to\infty} X_n(\omega) = \lim_{n\to\infty} \inf_{m\geq n} X_m(\omega)\,.$$

Alors

$$\liminf_{n\to\infty} \mathbb{E}(X_n) \geq \mathbb{E}(Y)\,.$$

Preuve Soit $Y_n := \inf_{m\geq n} X_m$. La suite des v.a. Y_n, $n \geq 1$, est monotone croissante et $P(Y_n > t) \leq P(X_m > t)$ si $m \geq n$. Les Y_n sont des v.a. car

$$\{Y_n < t\} = \bigcup_{m\geq n} \{X_m < t\}\,.$$

Comme $P(Y_n > t) \leq P(X_m > t)$, on obtient (voir théorème 7.1)

$$\liminf_{n\to\infty} \mathbb{E}(X_n) = \lim_{n\to\infty} \inf_{m\geq n} \mathbb{E}(X_m) \geq \lim_{n\to\infty} \mathbb{E}(Y_n) = \mathbb{E}(Y)\,.$$

La dernière égalité est une conséquence du théorème 7.2. □

Théorème 7.3 (Théorème de la convergence dominée) *Soit X_n, $n \geq 1$, une suite de v.a. qui converge ponctuellement vers la v.a. X. S'il existe une v.a. Y telle que*

$$\forall\,\omega \quad |X_n(\omega)| \leq Y(\omega) \quad \textit{et} \quad \mathbb{E}(Y) < \infty\,,$$

alors la v.a. X a une espérance et

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

Preuve On utilise deux fois le lemme 7.3 pour les suites de v.a. $Y \pm X_n \geq 0$. (On utilise l'additivité de l'espérance qui est établie au théorème 7.6).

$$\liminf_{n\to\infty} \mathbb{E}(Y + X_n) \geq \mathbb{E}(Y) + \mathbb{E}(X) \implies \liminf_{n\to\infty} \mathbb{E}(X_n) \geq \mathbb{E}(X)\,,$$

et

$$\liminf_{n\to\infty} \mathbb{E}(Y - X_n) = \mathbb{E}(Y) - \limsup_{n\to\infty} \mathbb{E}(X_n) \geq \mathbb{E}(Y) - \mathbb{E}(X)\,,$$

ce qui implique $\mathbb{E}(X) \geq \limsup_n \mathbb{E}(X_n)$. □

7.3 Propriétés de l'espérance

Les théorèmes 7.4 et 7.5 sont importants car ils permettent de calculer l'espérance de v.a. $Y = \varphi(X)$ en connaissant seulement la loi de X :

$$\mathbb{E}(Y) = \sum_{x\in D} \varphi(x) P(X = x) \quad \text{resp.} \quad \mathbb{E}(Y) = \int_{\mathbb{R}} \varphi(x) f_X(x)\,dx\,.$$

Ce résultat s'étend au cas où $Y = \varphi(X_1, \ldots, X_k)$. C'est alors la loi conjointe de $X_1, \ldots, X_k$ qu'il faut utiliser.

Exemple 7.3 Soit X une v.a. de loi exponentielle de paramètre λ. Si $\varphi(t) = t^p$, alors $\varphi(X) = X^p$ et

$$\mathbb{E}(X^p) = \lambda \int_0^\infty t^p \, \mathrm{e}^{-\lambda t} \, dt = \frac{p!}{\lambda^p} \, .$$

Le résultat est obtenu par p intégrations par parties. □

Soit k v.a. réelles $X_1, \ldots, X_k$ définies sur le même espace de probabilité et $\varphi \colon \mathbb{R}^k \to \mathbb{R}$ une application réelle [1]. On définit une nouvelle v.a.

$$Y := \varphi(X_1, \ldots, X_k) \, , \quad \omega \mapsto Y(\omega) := \varphi(X_1(\omega), \ldots, X_k(\omega)) \, .$$

Théorème 7.4 (cas discret) *Soit $X_1, \ldots, X_k$ des v.a. discrètes, $(X_1, \ldots, X_k) \in D \subset \mathbb{R}^k$, et $\varphi \colon \mathbb{R}^k \to \mathbb{R}$. L'espérance de Y est définie si et seulement si*

$$\sum_{(x_1,\ldots,x_k)\in D} |\varphi(x_1, \ldots, x_k)| P(X_1 = x_1, \ldots, X_k = x_k) < \infty \, .$$

Si cette condition est vérifiée

$$\mathbb{E}(Y) = \sum_{(x_1,\ldots,x_k)\in D} \varphi(x_1, \ldots, x_k) P(X_1 = x_1, \ldots, X_k = x_k) \, .$$

Preuve Pour simplifier l'écriture $k = 2$. L'image $\varphi(D)$ du sous-ensemble $D \subset \mathbb{R}^2$ par l'application φ est un sous-ensemble fini ou dénombrable de $\mathbb{R}$. Soit $\mathbf{x} = (x_1, x_2)$.

$$\begin{aligned} \sum_{y\in\varphi(D)} |y| P(Y = y) &= \sum_{y\in\varphi(D)} |y| \sum_{\substack{\mathbf{x}\in D: \\ \varphi(\mathbf{x})=y}} P(\mathbf{X} = \mathbf{x}) \\ &= \sum_{y\in\varphi(D)} \sum_{\substack{\mathbf{x}\in D: \\ \varphi(\mathbf{x})=y}} |\varphi(\mathbf{x})| P(\mathbf{X} = \mathbf{x}) \\ &= \sum_{\mathbf{x}\in D} |\varphi(\mathbf{x})| P(\mathbf{X} = \mathbf{x}) \, . \end{aligned}$$

On a utilisé le fait que les ensembles $E_y := \{x \in D \colon \varphi(\mathbf{x}) = y\}$ sont disjoints pour des valeurs différentes de y ; de plus l'union des ensembles E_y, lorsque y parcourt l'ensemble $\varphi(D)$, est l'ensemble D. Par un calcul analogue on obtient la formule pour $\mathbb{E}(Y)$. □

Théorème 7.5 (cas continu) *Soit $X_1, \ldots, X_k$ des v.a. avec une loi conjointe donnée par la densité de probabilité $f_{\mathbf{X}}$ et $\varphi \colon \mathbb{R}^k \to \mathbb{R}$. L'espérance de Y est définie si et seulement si*

$$\int_{\mathbb{R}^k} |\varphi(x_1, \ldots, x_k)| f_{\mathbf{X}}(x_1, \ldots, x_k) \, dx_1 \cdots dx_k < \infty \, .$$

1. Il faut que $\varphi^{-1}((-\infty, t])) \in \mathcal{B}(\mathbb{R}^k)$ pour tout $t \in \mathbb{R}$; c'est le cas par exemple si φ est continue.

Si cette condition est vérifiée

$$\mathbb{E}(Y) = \int_{\mathbb{R}^k} \varphi(x_1, \ldots, x_k) f_{\mathbf{X}}(x_1, \ldots, x_k)\, dx_1 \cdots dx_k < \infty\,.$$

Preuve On donne la preuve dans le cas où φ est bornée. C'est essentiellement le même calcul que celui de (7.1). Soit $I_k = (k2^{-n}, (k+1)2^{-n}]$. On approxime $\varphi(\mathbf{X})$ par des v.a. discrètes comme dans le lemme 7.1.

$$\begin{aligned}
\mathbb{E}\big(\varphi(\mathbf{X})\big) &= \lim_{n\to\infty} \sum_{k=0}^{n\,2^n-1} (k\,2^{-n}) P(\varphi(\mathbf{X}) \in I_k) \\
&= \lim_{n\to\infty} \sum_{k=0}^{n\,2^n-1} (k\,2^{-n}) P(\mathbf{X} \in \varphi^{-1}(I_k)) \\
&= \lim_{n\to\infty} \sum_{k=0}^{n\,2^n-1} (k\,2^{-n}) \int_{\varphi^{-1}(I_k)} f_{\mathbf{X}}(s_1\, \ldots, s_k)\, ds_1 \cdots ds_k\,.
\end{aligned}$$

Si $\mathbf{s} \in \varphi^{-1}(I_k)$, alors $|\varphi(\mathbf{s}) - k\,2^{-n}| \leq 2^{-n}$. Si φ est bornée par M et si $n > M$, les ensembles $\varphi^{-1}(I_k)$, $k = 0, \ldots, n\,2^n - 1$, forment une partition de $\{\mathbf{s} \in \mathbb{R}^k \colon \varphi(\mathbf{s}) > 0\}$. Dans la limite $n \to \infty$ (voir (7.2))

$$\mathbb{E}(\varphi(\mathbf{X})) = \int_{\mathbb{R}^k} \varphi(s_1\, \ldots, s_k) f_{\mathbf{X}}(s_1\, \ldots, s_k)\, ds_1 \cdots ds_k\,.$$

□

Exemple 7.4 Soit X_1, X_2 et X_3 des v.a. de loi conjointe gaussienne $N(\mathbf{0}, \mathbf{A})$ de paramètres $\mathbf{m} = \mathbf{0}$ et

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 4 & 0 \\ 0 & 0 & 2 \end{pmatrix}\,.$$

Si $\varphi(x_1, x_2, x_3) = x_1^4\, x_2^2$,

$$\mathbb{E}\big(\varphi(X_1, X_2, X_3)\big) = \int_{\mathbb{R}^3} x_1^4\, x_2^2\, f_{\mathbf{0},\mathbf{A}}(x_1, x_2, x_3)\, dx_1 dx_2 dx_3 = \frac{144}{343}\,.$$

(Voir exemple 5.6.) De même, si la loi conjointe de $X_1, \ldots, X_p$ est une mesure de probabilité gaussienne qui est spécifée par une matrice $\mathbf{A}$ et le vecteur $\mathbf{m} \in \mathbb{R}^p$, alors ($\mathbf{k} = (k_1, \ldots, k_p)$)

$$\mathbb{E}\Big(\exp[\sum_{j=1}^{p} k_j\, X_j]\Big) = \exp\Big(\langle \mathbf{k}|\mathbf{m}\rangle + \frac{1}{2}\,\langle \mathbf{k}|\mathbf{A}^{-1}\mathbf{k}\rangle\Big)\,. \tag{7.4}$$

(Voir proposition 5.2.) □

Le théorème 7.6 résume les propriétés de base de l'espérance. Les propriétés de l'espérance sont celles d'une intégrale : linéarité, positivité et propriété de monotonie (points 1 et 5, 2, 6). Attention, le point 7 *n'affirme pas* que si $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, alors les v.a. X et Y sont indépendantes.

Théorème 7.6 *Soit X et Y deux v.a. réelles définies sur le même espace de probabilité. On suppose que $\mathbb{E}(X)$ et $\mathbb{E}(Y)$ existent. Alors*

1) Si c est une constante, $\mathbb{E}(c \cdot X) = c \cdot \mathbb{E}(X)$.

2) Si $X \geq 0$, alors $\mathbb{E}(X) \geq 0$.

3) $\mathbb{E}(X)$ existe si et seulement si $\mathbb{E}(|X|)$ existe.

4) $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$ et $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$.

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

6) Si $X \leq Y$ alors $\mathbb{E}(X) \leq \mathbb{E}(Y)$.

7) Si X, Y sont indépendantes, alors $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Preuve On montre le point 5. Les autres points sont laissés en exercice. Si X et Y sont des v.a. discrètes, il existe des partitions $\{A_1, \ldots, A_m\}$ et $\{B_1, \ldots, B_n\}$ de Ω telles que

$$X = \sum_{i=1}^{m} x_i I_{A_i} \quad \text{et} \quad \sum_{j=1}^{n} y_j I_{B_j} .$$

On peut écrire

$$X + Y = \sum_{i,j} (x_i + y_j) I_{A_i \cap B_j} .$$

Par conséquent

$$\begin{aligned} \mathbb{E}(X+Y) &= \sum_i x_i \Big(\sum_j P(A_i \cap B_j) \Big) + \sum_j y_j \Big(\sum_i P(A_i \cap B_j) \Big) \\ &= \mathbb{E}(X) + \mathbb{E}(Y) . \end{aligned}$$

Dans le cas général X et Y sont des limites de v.a. discrètes (voir lemme 7.1). Par conséquent on a encore $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$. □

Définition 7.3 *La v.a. X possède une* variance *si et seulement si $\mathbb{E}(X)$ et $\mathbb{E}((X - \mathbb{E}(X))^2)$ existent. Par définition la variance est notée* $\mathrm{Var}X$ *et*

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

*L'*écart-type *ou* déviation standard *est $\sigma(X)$.*

Pour alléger l'écriture on écrit plus simplement $\mathrm{Var}X = \mathbb{E}(X - \mathbb{E}(X))^2$. Noter l'identité :

$$\begin{aligned} \mathrm{Var}X &= \mathbb{E}\big(X^2 - 2\mathbb{E}(X)\, X + \mathbb{E}(X)^2\big) \\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X)\, \mathbb{E}(X) + \mathbb{E}(X)^2 \\ &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 . \end{aligned}$$

La quantité $(X - \mathbb{E}(X))^2$ est le carré de l'écart entre la valeur de X et l'espérance de X. La variance représente donc l'*écart quadratique moyen* entre la valeur de X et l'espérance de X (voir aussi proposition 7.1 a) ci-dessous).

Exemples 7.5 a) Si X est une v.a. de Bernoulli de paramètre p,

$$\mathrm{Var}X = (1-p)(-p)^2 + p(1-p)^2 = p(1-p)\,.$$

b) Si $Y \sim \mathcal{B}_i(n,p)$, alors $Y \overset{\mathcal{L}}{=} \sum_{i=1}^n X_i$, où les X_i sont n v.a. indépendantes de Bernoulli de paramètre p. Par conséquent

$$\mathbb{E}(Y) = \mathbb{E}\big(\sum_{i=1}^n X_i\big) = n\mathbb{E}(X_1) = np\,.$$

$$\begin{aligned}
\mathrm{Var}Y &= \mathbb{E}\big([\sum_{i=1}^n X_i - np]^2\big) = \mathbb{E}\big([\sum_{i=1}^n (X_i - p)]^2\big) \\
&= \mathbb{E}\big([\sum_{i=1}^n (X_i - p)][\sum_{j=1}^n (X_j - p)]\big) \\
&= \sum_{i=1}^n \mathbb{E}((X_i-p)^2) + \sum_{i,j\,:\,i\neq j} \underbrace{\mathbb{E}((X_i-p)(X_j-p))}_{\substack{=0 \\ \text{(les v.a. sont indépendantes)}}} \\
&= \sum_{i=1}^n \mathrm{Var}X_i = np(1-p)\,.
\end{aligned}$$

c) Soit X une v.a. uniforme sur $(0,1)$;

$$\mathrm{Var}X = \int_0^1 \Big(t - \frac{1}{2}\Big)^2 dt = \frac{1}{12}\,.$$

d) Si $X \sim \pi_\lambda$, $\mathbb{E}(X) = \lambda$ et $\mathrm{Var}X = \lambda$.
e) Si $X \sim N(m,\sigma^2)$, $\mathbb{E}(X) = m$ et $\mathrm{Var}X = \sigma^2$.
f) Si $X \sim \gamma_{x,\lambda}$, $\mathbb{E}(X) = x/\lambda$ et $\mathrm{Var}X = x/\lambda^2$. □

La proposition suivante exprime le caractère de dispersion de la variance.

Proposition 7.1 *Soit X une v.a. possédant une espérance.*
a) Si $\mathbb{E}(X^2) < \infty$, alors

$$\forall a \in \mathbb{R}\colon\ \mathbb{E}(X-a)^2 = \mathrm{Var}X + (\mathbb{E}(X) - a)^2 \geq \mathrm{Var}X\,.$$

b) Si X est bornée et telle que $m \leq X(\omega) \leq M$, alors

$$\mathrm{Var}X \leq \frac{1}{4}(M-m)^2\,.$$

Cette inégalité est saturée pour $X = \pm 1$, $P(X=1) = P(X=-1) = 1/2$.
c) Si $\mathrm{Var}X = 0$, alors $P(X \neq \mathbb{E}(X)) = 0$.

Preuve a) On calcule

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

b) Par hypothèse $\mathbb{E}((M-X)(X-m)) \geq 0$. On développe et réarrange les termes ; on obtient

$$\mathbb{E}(X^2) - \mathbb{E}(X)^2 \leq \big(M-\mathbb{E}(X)\big)\big(\mathbb{E}(X)-m\big) \leq \frac{1}{4}(M-m)^2$$

car pour tout $x \in \mathbb{R}$,

$$\begin{aligned}(M-m)^2 &= (M-x)^2 + (x-m)^2 + 2(M-x)(x-m) \\ &= [(M-x)-(x-m)]^2 + 4(M-x)(x-m) \\ &\geq 4(M-x)(x-m)\,.\end{aligned}$$

Le point c) est démontré dans l'exemple 8.1 de la section 8.1. □

Définition 7.4 *Soit X et Y des v.a. telles que $\mathbb{E}(|XY|) < \infty$, $\mathbb{E}(|X|) < \infty$ et $\mathbb{E}(|Y|) < \infty$. La* covariance *de deux v.a. X et Y est le nombre réel*

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

X et Y sont non corrélées *si et seulement si* $\mathrm{Cov}(X,Y) = 0$.

Par définition $\mathrm{Cov}(X,X) = \mathrm{Var}X$. La covariance est linéaire dans chacun de ses arguments, par exemple

$$\mathrm{Cov}(aX_1 + bX_2, Y) = a\,\mathrm{Cov}(X_1,Y) + b\,\mathrm{Cov}(X_2,Y)\,.$$

Si $X_1, \ldots X_p$ sont des v.a. et si $a_i, b_i \in \mathbb{R}$, la vérification de l'identité suivante est une vérification de routine à partir des propriétés de l'espérance.

$$\begin{aligned}\mathrm{Var}\Big(\sum_{i=1}^p a_i(X_i+b_i)\Big) &= \mathbb{E}\Big(\sum_{i=1}^p a_i\big(X_i-\mathbb{E}(X_i)\big)\Big)^2 \qquad (7.5)\\ &= \sum_{i=1}^p a_i^2 \mathrm{Var}(X_i) + \sum_{i\neq j} a_i a_j \mathrm{Cov}(X_i,X_j)\,.\end{aligned}$$

Proposition 7.2 *Si des v.a. $X_1, \ldots, X_p$ possèdent des variances et sont indépendantes, elles sont non corrélées et*

$$\mathrm{Var}(X_1 + \cdots + X_p) = \mathrm{Var}X_1 + \cdots \mathrm{Var}X_p\,.$$

Preuve Si les v.a. X et Y sont indépendantes,

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}\big((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\big)\\ &= \mathbb{E}(X-\mathbb{E}(X))\,\mathbb{E}(Y-\mathbb{E}(Y)) = 0\,.\end{aligned}$$

La deuxième affirmation découle de (7.5). □

Si les v.a. $X_1,\dots,X_p$ ont une loi conjointe gaussienne $N(\mathbf{m},\mathbf{A})$,

$$\mathrm{Cov}(X_i,X_j) = \mathbf{A}^{-1}_{ij} = \mathbf{A}^{-1}_{ji}\,.$$

(Voir proposition 5.3). La matrice $\mathbf{A}^{-1}$ est appelée *matrice de covariance* des v.a. $X_1,\dots,X_p$. La loi conjointe de v.a. gaussiennes est complètement spécifiée par les espérances $\mathbb{E}(X_i) = m_i$ de chaque v.a. et par la matrice de covariance. Lorsque des v.a. sont non corrélées, elles ne sont pas indépendantes en général. La proposition suivante est à cet égard remarquable.

Proposition 7.3 *Des v.a. gaussiennes non corrélées sont indépendantes.*

Preuve Par hypothèse la matrice de covariance est diagonale et par conséquent la matrice inverse $\mathbf{A}$ est diagonale. La densité de la loi conjointe factorise, ce qui prouve l'indépendance des v.a. (proposition 6.1). □

On termine ce chapitre en donnant une preuve simple des formules d'inclusion-exclusion qui fait appel aux propriétés élémentaires de l'espérance.

Preuve de la proposition 2.2 $A := \bigcup_i A_i$; on a les relations

$$I_{A_j^c} = 1 - I_{A_j} \quad \text{et} \quad I_{A^c} = \prod_j I_{A_j^c} = \prod_j (1 - I_{A_j})\,.$$

On utilise l'identité $\mathbb{E}(I_B) = P(B)$ et on développe le produit :

$$\begin{aligned}P(A) &= 1 - P(A^c)\\ &= 1 - \mathbb{E}\big((1-I_{A_1})(1-I_{A_2})\cdots(1-I_{A_n})\big)\\ &= \sum_{j=1}^{n} P(A_j) - \sum_{J\subset\{1,\dots,n\}:\,|J|=2} P\Big(\bigcap_{j\in J} A_j\Big) + \cdots\\ &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{J\subset\{1,\dots,n\}:\,|J|=k} P\Big(\bigcap_{j\in J} A_j\Big)\,.\end{aligned}$$

De façon similaire, on montre la deuxième identité de la proposition 2.2, en utilisant la première identité.

$$\begin{aligned}
P(\bigcap_j A_j) &= 1 - P(\bigcup A_j^c) \\
&= 1 - \sum_{k=1}^{n} (-1)^{k+1} \sum_{J\subset\{1,\dots,n\}:\ |J|=k} P(\bigcap_{j\in J} A_j^c) \\
&= 1 - \sum_{k=1}^{n} (-1)^{k+1} \sum_{J\subset\{1,\dots,n\}:\ |J|=k} \Big(1 - P(\bigcup_{j\in J} A_j)\Big) \\
&= \underbrace{1 - \sum_{k=1}^{n} (-1)^{k+1} \sum_{J\subset\{1,\dots,n\}:\ |J|=k} 1}_{=0} \\
&\quad + \sum_{k=1}^{n} (-1)^{k+1} \sum_{J\subset\{1,\dots,n\}:\ |J|=k} P(\bigcup_{j\in J} A_j)\,.
\end{aligned}$$

□

7.4 Exercices

Exercice 7.1 Edmond Halley (1656-1742) a publié en 1693 une table de mortalité. Dans cette table on constate d'une part que le temps de vie moyen est de 26 ans, et d'autre part qu'il y a égale chance de mourir avant l'âge de 8 ans qu'après l'âge de 8 ans. Comment expliquer ces résultats ?

Exercice 7.2 Calculer l'espérance et la variance d'une v.a. X lorsque la loi de X est une
a) loi exponentielle de paramètre $\lambda > 0$;
b) loi de Poisson π_λ ;
c) loi normale $N(m, \sigma^2)$.

Exercice 7.3 On lance deux dés équilibrés simultanément. Soit X la somme des résultats des dés.
a) Déterminer la loi de X et calculer l'espérance de X.
b) Calculer la médiane de X.
c) Calculer la variance de X.

Exercice 7.4 Soit deux v.a. X et Y telles que $\mathrm{Var}X > 0$ et $\mathrm{Var}Y > 0$. La *corrélation entre X et Y* est le nombre

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

Montrer que

$$-1 \leq \rho(X,Y) \leq 1\,.$$

Exercice 7.5 a) Calculer l'espérance et la variance pour une v.a. $X \sim \gamma_{x,\lambda}$.
b) On considère une source radioactive ; le nombre d'émissions de cette source pendant l'intervalle de temps $[0,x]$ est décrit par une v.a. $N(x)$ dont la loi est une loi de Poisson $\pi_{\lambda x}$. On désigne par T_n la v.a. indiquant le temps de la $n^{\text{ième}}$ émission. Montrer

$$P(T_n \leq x) = P(N(x) \geq n)$$

et déterminer la densité de la v.a. T_n. Quelle est la loi de T_n ?

Exercice 7.6 Soit X une v.a. prenant les valeurs $0, 1, 2, \ldots$. Vérifier l'identité

$$\mathbb{E}(X) \equiv \sum_{n\geq 0} nP(\{X = n\}) = \sum_{k\geq 0} P(\{X > k\})\,.$$

Exercice 7.7 Soit X_1 et X_2 deux v.a. réelles, indépendantes et de même loi, qui prennent n valeurs $x_1, \ldots, x_n$; $P(X_1 = x_i) = p_i \geq 0$, $i = 1, \ldots, n$.
a) Calculer la probabilité $P(X_1 = X_2)$.
b) Montrer que

$$\frac{1}{n} \leq P(X_1 = X_2) \leq 1\,.$$

c) Donner des lois qui montrent que les bornes dans les inégalités ci-dessus sont saturées.

Exercice 7.8 X est une v.a. gaussienne de loi $N(0,\sigma^2)$ et Y une v.a. gaussienne de loi $N(0,\tau^2)$.
a) Si $X+Y$ est une v.a. gaussienne et si X et Y sont non corrélées, quelle est la loi de $X+Y$?
b) Si X et Y sont indépendantes, montrer que la loi de $X+Y$ est gaussienne et déterminer cette loi.
c) Si les $X_i \sim N(\mu,\sigma^2)$ sont indépendantes, montrer que $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$ est $N(\mu,\sigma^2/n)$ ou de façon équivalente que $Z := n^{\frac{1}{2}}\big(\overline{X} - \mu\big)/\sigma$ est $N(0,1)$.

Exercice 7.9 Soit $\alpha > 1$ et β tel que $1/\alpha + 1/\beta = 1$. Vérifier les inégalités de Hölder (1859-1937) et Minkowski (1864-1909) :

$$|\mathbb{E}(XY)| \leq \mathbb{E}(|X|^\alpha)^{1/\alpha}\,\mathbb{E}(|Y|^\beta)^{1/\beta}$$

et

$$\mathbb{E}(|X+Y|^\alpha)^{1/\alpha} \leq \mathbb{E}(|X|^\alpha)^{1/\alpha} + \mathbb{E}(|Y|^\alpha)^{1/\alpha}\,.$$

Indication : considérer des v.a. étagées.

Exercice 7.10 Soit $(\Omega,\mathcal{F},P)$ un espace de probabilité et $X : \Omega \to \mathbb{R}$ une v.a. étagée prenant les valeurs $x_1 < x_2 < \cdots < x_r$. On désigne par $\mathcal{A}$ l'algèbre de Boole associée à la v.a. X. Soit $Y : \Omega \to \mathbb{R}$ une autre v.a. ; on dit que Y est *$\mathcal{A}$-mesurable* si et seulement si $Y^{-1}((a,b]) \in \mathcal{A}$ pour tout a, b.
Montrer que Y est $\mathcal{A}$-mesurable si et seulement s'il existe une fonction $g : \mathbb{R} \to \mathbb{R}$ telle que $Y = g \circ X$.

Chapitre 8

Inégalités de Markov, Chebyshev et Hoeffding

Dans ce chapitre on présente trois inégalités importantes. Les inégalités de Markov et Chebyshev (1821-1894) sont autant utiles qu'elles sont élémentaires.

8.1 Inégalités de Markov et Chebyshev

Proposition 8.1 *Soit X une v.a. ;*

1) Inégalité de Markov *: si $\mathbb{E}(|X|) < \infty$ et $a > 0$, alors*

$$P(|X| \geq a) \leq \frac{\mathbb{E}(|X|)}{a}\,.$$

2) Inégalité de Chebyshev *: si $\mathrm{Var}X < \infty$ et $a > 0$, alors*

$$P(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathrm{Var}X}{a^2}\,.$$

Preuve Pour tout $a > 0$:

$$\begin{aligned}\mathbb{E}(|X|) &= \int_0^\infty P(|X| \geq t)\,dt \\ &\geq \int_0^a P(|X| \geq t)\,dt \\ &\geq aP(|X| \geq a)\,.\end{aligned}$$

On pose $Y := X - \mathbb{E}(X)$ et on applique l'inégalité de Markov :

$$P(|Y| \geq a) = P(|Y|^2 \geq a^2) \leq \frac{\mathbb{E}(Y^2)}{a^2} = \frac{\mathrm{Var}X}{a^2}\,.$$

□

On rappelle que la variance représente l'écart quadratique moyen entre la valeur de X et son espérance. L'inégalité de Chebyshev est utilisée souvent sous la forme suivante. Soit X une v.a. d'espérance m et d'écart-type σ. Alors

$$P(-k\,\sigma < X - m < k\,\sigma) \geq 1 - \frac{1}{k^2}\,.$$

En effet, si $Y := (X-m)/\sigma$, $\mathrm{Var}Y = 1$ et

$$P\Big(-k < \frac{X-m}{\sigma} < k\Big) = 1 - P\big(|Y| \geq k\big) \geq 1 - \frac{1}{k^2}\,.$$

Si $k = 2$, cette inégalité indique qu'avec une probabilité supérieure ou égale à 0.75, la valeur de X est comprise dans l'intervalle $[m - 2\sigma,\ m + 2\sigma]$. Ceci est vrai pour toute v.a. qui possède une variance. Pour une v.a. gaussienne on sait que cette probabilité est beaucoup plus grande, environ 0,95. Lorsque la variance n'existe pas la situation peut être très différente.

Exemple 8.1 On montre l'implication

$$\mathrm{Var}X = 0 \implies P(X \neq \mathbb{E}(X)) = 0$$

à l'aide de l'inégalité de Chebyshev. Par cette inégalité, pour tout $n \in \mathbb{N}$,

$$P\Big(|X - \mathbb{E}(X)| \geq \frac{1}{n}\Big) \leq n^2 \mathrm{Var}X = 0\,.$$

Par conséquent

$$P\Big(X \neq \mathbb{E}(X)\Big) \leq \sum_{n\geq 1} P\Big(|X - \mathbb{E}(X)| \geq \frac{1}{n}\Big) = 0\,.$$

Ceci prouve le point c) de la proposition 7.1. □

Une v.a. $X \geq 0$ possède une espérance si et seulement si la fonction $t \mapsto P(X > t)$ est intégrable (théorème 7.1). L'inégalité de Markov donne une borne supérieure pour $P(X > t)$ qui n'est pas intégrable. Le lemme suivant donne une amélioration de l'inégalité de Markov.

Lemme 8.1

$$\mathbb{E}(|X|) < \infty \implies \lim_{a\to\infty} aP(|X| > a) = 0\,.$$

Preuve On pose $Y_a := |X| I_{\{|X|>a\}}$. Par définition, $Y = 0$ ou $Y_a > a$; si $0 \leq t \leq a$, alors $P(Y_a > t) = P(|X| > a)$ et

$$\begin{aligned}
\mathbb{E}(Y_a) &= \int_0^a P(Y_a > t)\,dt + \int_a^\infty P(Y_a > t)\,dt \\
&= aP(|X| > a) + \int_a^\infty P(|X| > t)\,dt \\
&\geq aP(|X| > a)\,.
\end{aligned}$$

D'une part

$$\mathbb{E}(|X|) = \mathbb{E}(|X| I_{\{|X|\le a\}}) + \mathbb{E}(Y_a)\,;$$

d'autre part (théorème 7.2)

$$\lim_{a\to\infty} \mathbb{E}(|X| I_{\{|X|\le a\}}) = \mathbb{E}(|X|)$$

de sorte que $\lim_a \mathbb{E}(Y_a) = 0$. Par conséquent

$$0 \le \lim_{a\to\infty} aP(|X| > a) \le \lim_{a\to\infty} \mathbb{E}(Y_a) = 0\,.$$

□

Exemple 8.2 Soit $f\colon [0,1] \to \mathbb{R}$ une fonction continue. Un théorème important de Weierstrass (1815-1897) affirme que toute fonction réelle et continue sur $[0,1]$ peut être approchée uniformément par un polynôme, i.e.

$$\forall\, \varepsilon > 0 \ \exists \text{ un polynôme } Q \text{ tel que } \sup_{t\in[0,1]} |f(t) - Q(t)| \le \varepsilon\,.$$

L'idée de la preuve qui suit est de Bernstein (1880-1968). On introduit le *polynôme de Bernstein*

$$Q_n(t) := \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big)\binom{n}{k} t^k (1-t)^{n-k} = \mathbb{E}\Big[f\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)\Big]\,;$$

les v.a. $X_1, \dots, X_n$ sont indépendantes et sont des v.a. de Bernoulli de paramètre t. Soit $\varepsilon > 0$; comme f est bornée et uniformément continue il existe $\delta > 0$ tel que

$$|f(s) - f(t)| \le \varepsilon + \Big(2 \sup_{u\in[0,1]} |f(u)|\Big) I_{\{s' : |s'-t| \ge \delta\}}(s)\,.$$

Par conséquent

$$\begin{aligned}
|Q_n(t) - f(t)| &= \Big|\mathbb{E}\Big[f\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) - f(t)\Big]\Big| \\
&\le \mathbb{E}\Big(\Big|f\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) - f(t)\Big|\Big) \\
&\le \varepsilon + 2 \sup_{u\in[0,1]} |f(u)|\, P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - t\Big| \ge \delta\Big) \\
&\le \varepsilon + 2 \sup_{u\in[0,1]} |f(u)|\, \frac{t(1-t)}{n\delta^2} \\
&\le \varepsilon + 2 \sup_{u\in[0,1]} |f(u)|\, \frac{1}{4n\delta^2}\,.
\end{aligned}$$

Cette estimée est uniforme en t ; en prenant la limite $n \to \infty$ et en notant que $\varepsilon > 0$ est arbitraire, on obtient que les polynômes Q_n convergent uniformément vers f. □

8.2 Inégalité de Hoeffding

L'*inégalité de Hoeffding* (1914-1991) donne aussi une estimée de la probabilité qu'une v.a. Y, somme de v.a. indépendantes et bornées, diffère de son espérance $\mathbb{E}(Y)$ d'une quantité au moins égale à c.

Théorème 8.1 (Hoeffding (1963)) *Soit $X_1, X_2, \ldots$ une suite de v.a. indépendantes. On suppose que pour tout i il existe $a_i < b_i$ tels que*

$$a_i \leq X_i - \mathbb{E}(X_i) \leq b_i \,.$$

Alors pour tout $c > 0$

$$P\Big(\sum_{i=1}^n X_i - \mathbb{E}\big(\sum_{i=1}^n X_i\big) \geq c\Big) \leq \exp\Big(-\frac{2c^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big)$$

et

$$P\Big(\sum_{i=1}^n X_i - \mathbb{E}\big(\sum_{i=1}^n X_i\big) \leq -c\Big) \leq \exp\Big(-\frac{2c^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big) \,.$$

Les deux inégalités du théorème 8.1 donnent ensemble l'inégalité

$$P\Big(\Big|\sum_{i=1}^n X_i - \mathbb{E}\big(\sum_{i=1}^n X_i\big)\Big| \geq c\Big) \leq 2\exp\Big(-\frac{2c^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big) \,.$$

L'idée principale de la preuve du théorème 8.1 est simple et utile. Si Y est une v.a. telle que pour $t_0 > 0$

$$\mathbb{E}\exp(t_0 Y) < \infty \,,$$

alors à partir de l'inégalité de Markov

$$P(Y \geq a) = P\big(\mathrm{e}^{tY} \geq \mathrm{e}^{ta}\big) \leq \mathrm{e}^{-ta}\mathbb{E}\big(\mathrm{e}^{tY}\big) \quad \forall\, t \in [0, t_0] \,.$$

Dans cette inégalité t est un paramètre qui peut être choisi comme on le veut. Par conséquent

$$P(Y \geq a) \leq \inf_{t \in [0,t_0]} \mathrm{e}^{-ta}\mathbb{E}\big(\mathrm{e}^{tY}\big) \,.$$

Preuve du théorème 8.1. Il suffit de montrer la première inégalité car la deuxième suit de la première en posant $Z_i := -X_i$. Soit $Y_i := X_i - \mathbb{E}(X_i)$ de sorte que $a_i < Y_i < b_i$. Soit $t \geq 0$;

$$\begin{aligned} P\big(\sum_i Y_i \geq c\big) &= P\big(\exp\big(t\sum_i Y_i\big) \geq \mathrm{e}^{tc}\big) \\ &\leq \mathrm{e}^{-tc}\,\mathbb{E}\big(\exp\big(t\sum_i Y_i\big)\big) \\ &= \exp(-ct + \varphi(t)) \,, \end{aligned}$$

où l'on a posé

$$\varphi(t) := \ln \mathbb{E}\big(\exp\big(t \sum_i Y_i\big)\big)\,.$$

On écrit un développement de Taylor à l'ordre deux pour la fonction φ. Pour simplifier l'écriture on pose aussi

$$Z(t) := \mathbb{E}\big(\exp\big(t \sum_i Y_i\big)\big)\,.$$

Pour tout $t \geq 0$, il existe s tel que $0 \leq s \leq t$ et

$$\varphi(t) = \varphi(0) + t\varphi'(0) + \frac{t^2}{2}\varphi''(s)\,.$$

Comme $\mathbb{E}(Y_i) = 0$, on obtient immédiatement que $\varphi(0) = 0$ et $\varphi'(0) = 0$. En utilisant l'indépendance des v.a. Y_j, et en introduisant les mesures de probabilité

$$A \mapsto \nu_j(A) := \frac{\mathbb{E}(I_A e^{sY_j})}{\mathbb{E}(e^{sY_j})}\,,$$

on obtient

$$\begin{aligned}
\varphi''(s) &= \frac{\mathbb{E}\big[\sum_i Y_i \sum_j Y_j \exp\big(s\sum_k Y_k\big)\big] Z(s) - \big(\mathbb{E}\big[\sum_i Y_i \exp\big(s\sum_k Y_k\big)\big]\big)^2}{Z(s)^2} \\
&= \sum_{j=1}^n \Big[\frac{\mathbb{E}\big(Y_j^2 e^{sY_j}\big)}{\mathbb{E}(e^{sY_j})} - \Big(\frac{\mathbb{E}\big(Y_j e^{sY_j}\big)}{\mathbb{E}(e^{sY_j})}\Big)^2 \Big] \\
&= \sum_{j=1}^n \mathrm{Var}_{\nu_j}(Y_j) \leq \frac{1}{4}\sum_{j=1}^n (b_j - a_j)^2\,.
\end{aligned}$$

La dernière inégalité découle de la proposition 7.1. Par conséquent,

$$P\big(\sum_i Y_i \geq c\big) \leq \exp\big(-tc + \frac{t^2}{8}\sum_i (b_i - a_i)^2\big) \quad \forall\, c > 0\,.$$

Il suffit de choisir $t = \dfrac{4c}{\sum_i (b_i - a_i)^2}$ pour obtenir le théorème. □

Remarque 8.1 Dans la preuve du théorème 8.1 l'indépendance des v.a. n'a été utilisée que tout à la fin pour contrôler $\varphi''(s)$ uniformément en s.

$$\varphi''(s) = \sum_i \big(\sum_j \mathrm{Cov}_s(Y_i, Y_j)\big)\,,$$

où $\mathrm{Cov}_s(Y_i, Y_j)$ est calculée par rapport à la mesure de probabilité définie par

$$\mathbb{E}_s(Y) := \frac{\mathbb{E}\big(Y \exp\big[s\sum_k X_k\big]\big)}{\mathbb{E}\big(\exp\big[s\sum_k X_k\big]\big)}\,.$$

S'il existe une constante telle que

$$\varphi''(s) \le \frac{Kn}{4}\,,$$

alors

$$P\Big(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) \ge c\Big) \le \exp\Big(-\frac{2c^2 n}{K}\Big)\,.$$

Exemple 8.3 Soit 100 v.a. indépendantes de même loi, $P(X = i) = 1/6$ avec $i = 1, \ldots, 6$.

$$\mathbb{E}(X_i) = 3{,}5 \quad \text{et} \quad -2{,}5 \le X_i - \mathbb{E}(X_i) \le 2{,}5\,.$$

Est-ce qu'il est probable que $\sum_i X_i \ge 500$? La réponse est non, car

$$\begin{aligned} P\Big(\sum_i X_i \ge 500\Big) &= P\Big(\sum_i (X_i - \mathbb{E}(X_i)) \ge 150\Big) \\ &\le \exp\Big(-\frac{2\cdot(150)^2}{100\cdot 5^2}\Big) \approx 1{,}5\cdot 10^{-8}. \end{aligned}$$

On peut comparer cette estimée avec celle obtenue en utilisant l'inégalité de Chebyshev. Dans cet exemple $\mathrm{Var} X_i = 35/12$ et la probabilité de l'événement est majorée par 0,013. □

Si les X_i sont indépendantes et de même loi, alors $a_i = a$ et $b_i = b$ et l'inégalité s'écrit

$$\forall \varepsilon > 0\colon \quad P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\big(X_i - \mathbb{E}(X_i)\big)\Big| \ge \varepsilon\Big) \le 2\exp\Big(-\frac{2\varepsilon^2 n}{(b-a)^2}\Big)\,.$$

Dans le cas de v.a. $X_i = \pm 1$, $P(X = 1) = P(X = -1) = 1/2$, on peut prendre $-a = b = 1$ et

$$P\Big(\Big|\sum_{i=1}^{n} X_i\Big| \ge \varepsilon n\Big) \le 2\exp\Big(-\frac{\varepsilon^2 n}{2}\Big)\,. \tag{8.1}$$

On parle de *grandes déviations par rapport à la moyenne* car l'événement

$$\Big\{\Big|\frac{1}{n}\sum_i^n X_i - \mathbb{E}(X_1)\Big| \ge t\Big\} = \Big\{\Big|\sum_i^n X_i - \mathbb{E}\Big(\sum_i^n X_i\Big)\Big| \ge nt\Big\}$$

exprime une déviation de l'ordre $O(n)$. Ces déviations sont souvent très rares ; la quantité d'intérêt est la vitesse avec laquelle la probabilité de ces déviations tend vers 0 lorsque n diverge,

$$\limsup_{n\to\infty} \frac{1}{n} \ln P\Big(\frac{1}{n}\Big|\sum_{i=1}^{n} X_i - \mathbb{E}(X_1)\Big| \notin (-\varepsilon, \varepsilon)\Big)\,. \tag{8.2}$$

Lorsque cette quantité est strictement négative (ici elle est inférieure à $-\frac{\varepsilon^2}{2}$), la *convergence vers* 0 est *exponentiellement rapide*.

Grâce au théorème 8.1 on peut aussi estimer la probabilité de *déviations modérées* en posant $c = \varepsilon n^\alpha$ avec $1/2 < \alpha < 1$. Sous les hypothèses du théorème, de telles déviations sont aussi rares lorsque n diverge puisque $\alpha > 1/2$. On observe dans le comportement de l'inégalité (8.1) un changement qualitatif pour $\alpha = 1/2$: le membre de droite ne tend plus vers 0 lorsque $n \to \infty$, mais

$$P\Big(\Big|\sum_{i=1}^{n} X_i\Big| \geq t\sqrt{n}\Big) \leq 2\exp\Big(-\frac{t^2}{2}\Big)\,. \tag{8.3}$$

Dans le chapitre 12 on verra que ce changement qualitatif a effectivement lieu.

Exemple 8.4 Suite des exemples 2.4 et 6.4. Dans le cas du modèle d'Ising en champ moyen, la v.a. M_n, qui donne l'aimantation par spin, est la moyenne arithmétique des v.a. X_i ; ces v.a. ne sont pas indépendantes. Cependant, on a une situation analogue à celle de (8.1). Soit $E \subset (-1, 1)$ un sous-ensemble fermé et A_n l'événement

$$A_n := \{M_n \in E\}\,.$$

A partir des inégalités (6.5), en sommant sur les $x_k \in E$, on obtient

$$\frac{C_1}{\sqrt{n}Z_n}\exp(n\max_{x_k\in E} g_{h,\beta}(x_k)) \leq P(A_n) \leq \frac{(n+1)C_2}{Z_n}\exp(n\max_{x_k\in E} g_{h,\beta}(x_k))\,.$$

En utilisant (6.8) et (6.9), par un calcul similaire à celui de l'énergie libre de l'exemple 6.4, il existe une constante K telle que

$$\Big|\frac{1}{n}\ln P(A_n) - \Big(\max_{x\in[a,b]} g_{h,\beta}(x) - \max_{x'\in[-1,1]} g_{h,\beta}(x')\Big)\Big| \leq K\,\frac{\ln n}{n}\,.$$

Lorsque n tend vers l'infini,

$$\lim_{n\to\infty}\frac{1}{n}\ln P(A_n) = \max_{x\in E}\big(g_{h,\beta}(x) + \beta f(h,\beta)\big) \equiv \max_{x\in E}\kappa_{h,\beta}(x)\,. \tag{8.4}$$

La quantité $\kappa_{h,\beta}(x)$ est non-positive et

$$\kappa_{h,\beta}(x) = 0 \iff x \text{ est un maximum global de } g_{h,\beta}\,.$$

Si le sous-ensemble fermé E ne contient pas un maximum global de $g_{h,\beta}$, alors

$$\lim_{n\to\infty}\frac{1}{n}\ln P(A_n) = \max_{x\in E}\kappa_{h,\beta}(x) < 0\,. \tag{8.5}$$

Le résultat (8.5) est important. Il signifie que la probabilité d'observer une valeur de l'aimantation par spin, dans un sous-ensemble fermé E qui ne contient pas un maximum global de $g_{h,\beta}$, est exponentiellement petite lorsque n est grand. Ces grandes déviations par rapport aux maxima globaux de $g_{h,\beta}$ sont rares, et deviennent non observables dans la limite thermodynamique. Les valeurs observées de l'aimantation par spin à l'équilibre, dans la limite thermodynamique, sont donc celles des maxima globaux de $g_{h,\beta}$. Pour le calcul de ces maxima globaux de $g_{h,\beta}$, se référer à l'exemple 6.4 ; voir aussi figure 10.1.
Suite à l'exemple 10.4 section 10.3. □

8.3 Exercices

Exercice 8.1 a) Soit $q_1 \geq 0, \ldots, q_p \geq 0$ tels que $q_1 + \cdots + q_p = 1$. Vérifier l'identité

$$\sum_{\substack{n_1,\ldots,n_p \geq 0 : \\ n_1+\cdots+n_p=n}} \binom{n}{n_1\,, n_2\,, \cdots\,, n_p} q_1^{n_1} \cdots q_p^{n_p} = 1\,.$$

b) Si $n_1 \geq 0, \ldots, n_p \geq 0$ et $n_1 + \cdots + n_p = n$, montrer l'inégalité

$$\binom{n}{n_1\,, n_2\,, \cdots\,, n_p} \leq \exp(nH(X))$$

où $H(X)$ est l'entropie d'une v.a. prenant les valeurs $a_1, \ldots, a_p$ avec probabilité $P(X = a_i) = n_i/n$ (voir (6.3)).
Indication : partir de l'identité établie au point a).

Exercice 8.2 Soit Φ une fonction continue et convexe définie sur $\mathbb{R}$. Montrer l'inégalité de Jensen

$$\Phi(\mathbb{E}(X)) \leq \mathbb{E}(\Phi(X))\,.$$

Indication : utiliser la caractérisation géométrique d'une fonction convexe.

Exercice 8.3 On considère des v.a. indépendantes $X_1, X_2, \ldots$ qui ont la même loi et à valeur dans un ensemble fini $\mathbb{A}$. Le *temps de récurrence moyen de l'état* $a \in \mathbb{A}$ est par définition

$$\theta_a := P(X_2 = a | X_1 = a) + \sum_{k=2}^{\infty} kP(X_2 \neq a, \ldots, X_k \neq a, X_{k+1} = a | X_1 = a)\,.$$

Calculer θ_a en fonction de $p := P(X_i = a)$.

Exercice 8.4 On considère des v.a. aléatoires indépendantes de Bernoulli de paramètre $0 < p < 1$. On sait que le temps de récurrence moyen d'un « succès » $(X_i = 1)$ est de 500. Par exemple, ces v.a. modélisent une éruption volcanique d'un volcan donné pour lequel on considère qu'il y a au plus une éruption par an. Si elle a lieu pendant l'année i, alors $X_i = 1$.
a) Laquelle de ces affirmations semble correcte :

1) On observe une éruption tous les 500 ans.

2) La probabilité qu'une éruption se produise chaque année est 1/500.

b) Calculer la probabilité qu'aucune éruption n'ait lieu pendant 350 ans.
c) Calculer le nombre d'années consécutives maximales pour qu'aucune éruption ne se produise avec probabilité 0,95.
d) Soit

$$T := \min\{j \geq 1 \colon X_1 + \cdots + X_j \geq 1\}\,.$$

Calculer $P(T > k)$, $P(T > k + n \big| T > n)$.

Exercice 8.5 Soit $X_1, \ldots, X_n$ n v.a. réelles. Vérifier que la matrice de covariance $(\mathrm{Cov}(X_i, X_j))_{i,j}$ est non négative. $\mathrm{Cov}(X_1, X_2)$ est aussi appelée fonction de corrélation en mécanique statistique.

Exercice 8.6 On considère un gaz libre de n particules classiques (distinguables) qui se trouvent dans un cube $V \subset \mathbb{R}^3$ de côté L. On ne s'intéresse qu'à la position des particules dans V. Ce système est décrit par l'espace de probabilité $(\Omega, \mathcal{F}, P)$ tel que

$$\Omega := \{\omega = (\omega_1, \ldots, \omega_n) : \omega_j \in V \ , \ \forall\, j = 1, \ldots, n\} \subset \mathbb{R}^{3n} ;$$

ω_j est la position de la particule de nom j, et P est la mesure de probabilité uniforme sur Ω. Pour chaque $1 \leq k \leq n$ on définit la v.a. X_k, $X_k(\omega) := \omega_k$ qui donne la position de la particule indexée par k.
1) Calculer la loi de X_k. Montrer que les v.a. X_k sont indépendantes.
2) Soit $Q \subset V$ un sous-ensemble de volume $|Q| > 0$. N_Q est la v.a. qui donne le nombre de particules se trouvant dans Q,

$$N_Q := \big|\{\, j : X_j \in Q \,\}\big| .$$

Déterminer la loi de N_Q et calculer son espérance. Déterminer la loi conjointe de N_{Q_1} et N_{Q_2} si $Q_1 \cap Q_2 = \emptyset$. Est-ce que les v.a. N_{Q_1} et N_{Q_2} sont indépendantes ?

Exercice 8.7 On étudie les v.a. N_Q de l'exercice 8.6 dans la limite thermodynamique : on prend les limites $L \to \infty$ et $n \to \infty$ de sorte que $n/L^3 = \lambda > 0$.
1) Déterminer la loi et l'espérance de N_Q dans cette limite.
2) Montrer que dans cette limite les v.a. N_{Q_1} et N_{Q_2} sont indépendantes si $Q_1 \cap Q_2 = \emptyset$.
3) Dans la limite thermodynamique estimer la probabilité de l'événement

$$P((N_Q - \mathbb{E}(N_Q)) \geq a|Q|) \, , \ a > 0 \, .$$

Comparer le résultat de l'estimation pour $a = \lambda/2$, respectivement $a = u/\sqrt{|Q|}$ lorsque $|Q|$ est grand.
Indication : montrer que $\mathbb{E}(\exp(tN_Q)) < \infty$ et utiliser l'inégalité de Markov. Choisir ensuite t de façon optimale.

Exercice 8.8 Inégalité de Paley (1907-1933) et Zygmund (1900-1992). Si $X \geq 0$ est une v.a. telle que $\mathbb{E}(X^2) < \infty$ et si $a < \mathbb{E}(X)$, alors

$$P(X > a) \geq \frac{(\mathbb{E}(X) - a)^2}{\mathbb{E}(X^2)} \, .$$

Indication : minorer et majorer $\mathbb{E}(X I_{\{X>a\}})$; pour majorer utiliser l'inégalité de Cauchy-Schwarz (Schwarz (1843-1921)).

Exercice 8.9 a) Soit $A_1, \ldots, A_n$ n événements. Montrer l'inégalité

$$P(A_1 \cup \cdots \cup A_n) \geq P(A_1) + \cdots + P(A_n) - \sum_{i<j} P(A_i \cap A_j) \, .$$

b) Soit $X_1, \ldots, X_n$ des v.a. indépendantes et de même loi. Montrer que

$$\lim_{t\to\infty} \frac{P(\max_{i=1}^n X_i > t)}{nP(X_1 > t)} = 1\,.$$

Exercice 8.10 Soit $X_1, \ldots, X_n$ des v.a. indépendantes et de même loi.
a) Calculer la loi de

$$Y_n := \max_{i=1}^n X_i\,.$$

b) Montrer que $\lim_{n\to\infty} P(Y_n \leq n) = 1$ si $\mathbb{E}(X_i)$ existe.
c) On définit λ_p par l'équation

$$P(Y_n \leq \lambda_p) = p \quad (0 < p < 1)\,.$$

Montrer que pour n grand

$$P(X_1 > \lambda_p) \simeq \frac{\ln 1/p}{n}\,.$$

d) Déterminer le comportement de λ_p, lorsque n diverge, dans les cas suivants :

1) X_1 a une loi exponentielle de paramètre 1.

2) X_1 a une loi gaussienne $N(0,1)$.

3) X_1 a une loi de Pareto de paramètre $\alpha > 0$, $f_X(t) = \dfrac{\alpha}{(1+t)^{1+\alpha}} I_{\mathbb{R}^+}(t)$.

Remarque : si $p = 1/2$, $\lambda_{1/2}$ est la médiane.

Chapitre 9

Chaîne de Markov

Un des buts de la théorie des probabilités est de construire des modèles. Dans ce chapitre on étudie très brièvement une classe de modèles appelés chaînes de Markov (à nombre fini d'états). La théorie des chaînes de Markov a de multiples applications.

9.1 Chaîne de Markov à temps discret

Un *processus stochastique à temps discret* est une collection finie ou dénombrable de v.a. X_t, définies sur un même espace de probabilité, et qui sont indexées en général par le paramètre $t \in \{0, 1, \ldots, n\}$ ou $t \in \{0\} \cup \mathbb{N}$. Les v.a. X_t sont à valeur dans $\mathbb{A}$, *l'espace des états du processus* ; dans ce chapitre $\mathbb{A}$ est toujours un ensemble fini. Le paramètre t est interprété habituellement comme une variable temporelle, bien que n'importe quelle autre interprétation soit possible. Par exemple, le processus décrit l'évolution aléatoire d'un système qui peut se trouver dans $|\mathbb{A}|$ états possibles représentés par les éléments de $\mathbb{A}$. La v.a. X_k indique l'état du système au temps $t = k$.

Exemple 9.1 On considère le système formé de deux urnes U_1 et U_2 dont la composition est la suivante. Il y a $2N$ boules dont N sont rouges et N noires ; chaque urne contient la moitié des boules. A chaque unité de temps on procède à un tirage : on tire au hasard une boule de U_1 et au hasard une boule de U_2, puis on remet la boule tirée de U_1 dans U_2 ainsi que la boule tirée de U_2 dans U_1. L'état du système est déterminé par le nombre de boules noires dans l'urne U_1 ; les états possibles sont indexés par $0, 1, \ldots, N$, i.e. $\mathbb{A} = \{0, 1, \ldots, N\}$; $X_k = \#\{\text{boules noires de } U_1 \text{ au temps } t = k\}$. □

Lorsque $t \in \{0, 1, \ldots, n\}$ le processus est entièrement déterminé une fois que l'on a donné la loi conjointe des v.a. $X_0, X_1, \ldots, X_n$. Lorsque $t \in \{0\} \cup \mathbb{N}$, il faut donner les lois conjointes

$$P(X_0 = \omega_0, \ldots, X_m = \omega_m)$$

pour toutes les valeurs de m. Une *trajectoire*, ou *histoire*, jusqu'au temps n est déterminée par la suite des états $(a_0, a_1, \ldots, a_n)$ du processus jusqu'au temps n. L'espace fondamental qui décrit les trajectoires jusqu'au temps $t = n$ est

$$\Omega_n := \{\omega = (\omega_0, \omega_1, \ldots, \omega_n) : \omega_i \in \mathbb{A}\,, \forall\, i = 0, \ldots, n\} = \mathbb{A}^{n+1}\,.$$

La loi conjointe $P(X_0 = \omega_0, \ldots, X_n = \omega_n)$ est définie sur Ω_n ; désormais on considère que la v.a. X_k est définie sur Ω_n (représentation canonique des X_k)

$$X_k : \Omega_n \to \mathbb{A} \quad \omega \mapsto X_k(\omega) := \omega_k \,.$$

L'événement « l'état du système en $t = 3$ est a, et en $t = 5$ est b » s'exprime par

$$\{X_3 = a, X_5 = b\} \,.$$

Pour alléger l'écriture, à la place de $P(X_i = \omega_i, X_j = \omega_j, X_k = \omega_k)$, on écrit $P(\omega_i, \omega_j, \omega_k)$. De la proposition 4.1 on obtient

$$P(\omega_0, \ldots, \omega_n) = P(\omega_0)\, P(\omega_1|\omega_0)\, P(\omega_2|\omega_0, \omega_1) \cdots P(\omega_n|\omega_0, \omega_1, \ldots, \omega_{n-1}) \,.$$

$P(\omega_{j+1}|\omega_0, \ldots, \omega_j)$ donne la probabilité que le processus se trouve dans l'état ω_{j+1} au temps $t = j + 1$, sachant que sa trajectoire jusqu'au temps $t = j$ est donnée par $(\omega_0, \ldots, \omega_j)$.

Définition 9.1 *Un processus stochastique vérifie la* propriété de Markov *(1856-1922) si*

$$P(\omega_{j+1}|\omega_0, \omega_1, \ldots, \omega_j) = P(\omega_{j+1}|\omega_j) \quad \forall\, j \,.$$

Un tel processus stochastique est appelé chaîne de Markov.

La propriété de Markov exprime une perte de mémoire : la probabilité que le système se trouve dans l'état ω_{j+1} au temps $t = j + 1$, sachant le passé, ne dépend que de l'état ω_j du système au temps $t = j$. En supposant que les règles de l'évolution du système sont indépendantes du temps, il suffit de donner la loi de la v.a. X_0 et une matrice stochastique $\mathbf{M}$,

$$\mathbf{M}_{k\ell} := P(X_{j+1} = \ell \,|X_j = k) \,,$$

pour spécifier la chaîne de Markov. En effet

$$\begin{aligned} P(\omega_0, \ldots, \omega_n) &= P(\omega_0) P(\omega_1|\omega_0) P(\omega_2|\omega_1) \cdots P(\omega_n|\omega_{n-1}) \\ &= P(\omega_0) \mathbf{M}_{\omega_0\omega_1} \cdots \mathbf{M}_{\omega_{n-1}\omega_n} . \end{aligned}$$

A partir de cette formule, en sommant sur $\omega_1 \in \mathbb{A}, \ldots, \omega_{n-1} \in \mathbb{A}$, on obtient la probabilité que le processus est dans l'état ω_n au temps $t = n$, sachant que l'état initial au temps $t = 0$ est ω_0,

$$P(\omega_n|\omega_0) = \sum_{\omega_1, \ldots, \omega_{n-1}} \mathbf{M}_{\omega_0\omega_1} \cdots \mathbf{M}_{\omega_{n-1}\omega_n} = \mathbf{M}^n_{\omega_0\omega_n} \quad \text{(produit matriciel)}$$

où $\mathbf{M}^n_{\omega_0\omega_n}$ est l'élément d'indices $\omega_0\omega_n$ de la matrice $\mathbf{M}^n$ (matrice produit de n facteurs $\mathbf{M}$).

Certaines propriétés du processus peuvent être utilement décrites à l'aide d'un graphe. Soit $\mathcal{G}$ le graphe dont les sommets sont indexés par les éléments

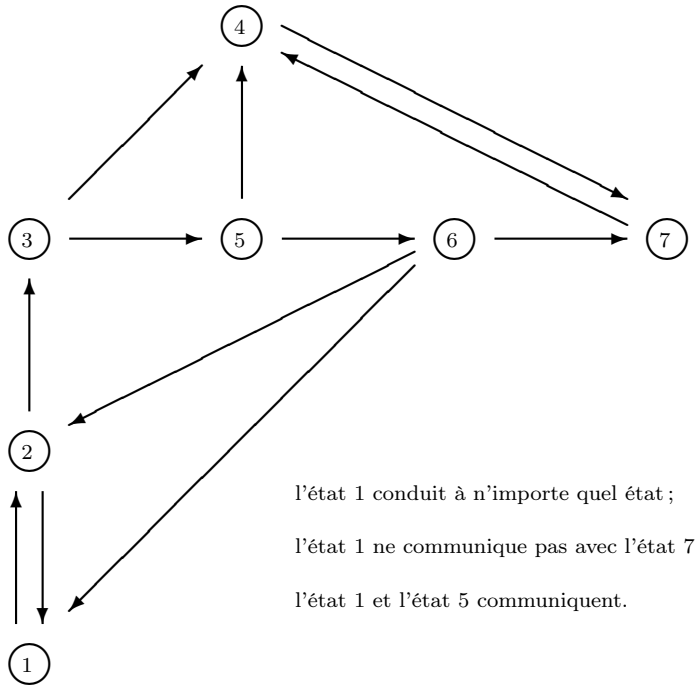


Figure 9.1 – Le graphe $\mathcal{G}$ associé à une matrice stochastique $\mathbf{M}$.

de $\mathbb{A}$. Deux sommets i et j sont reliés par un lien orienté $\langle i, j\rangle$ si et seulement si $\mathbf{M}_{ij} > 0$. L'état i *conduit à* j, $i \to j$, si et seulement si

$$\exists\, m \quad \text{tel que} \quad P(X_m = j | X_0 = i) > 0\,.$$

L'état i *communique avec* j, $i \leftrightarrow j$, si et seulement si $i \to j$ et $j \to i$.

Exemple 9.2 On lance 100 fois une pièce de monnaie équilibrée. Chaque résultat de l'expérience est également probable. Une *séquence* de Piles (Faces) est une succession maximale de Piles (Faces). Dans la suite de symboles

$$\underbrace{F}\,\underbrace{PPP}\,\underbrace{FF}\,\underbrace{P}\,\underbrace{F}\,\underbrace{PP}\,\underbrace{FF}$$

il y a 7 séquences ; la plus longue est de longueur 3 et la dernière est FF.

On définit l'événement $E_{100}(k)$ « dans une suite de 100 lancers la séquence la plus longue est de longueur $\geq k$ ». Pour calculer la probabilité de cet événement on introduit un automate aléatoire. On considère le cas explicite $k = 6$. L'automate est un système possédant 6 états notés $1, \ldots, 6$. On pose

$$\begin{aligned} L_m &:= \text{longueur de la séquence la plus longue après } m \text{ lancers} ; \\ \ell_m &:= \text{longueur de la dernière séquence après } m \text{ lancers.} \end{aligned}$$

Dans l'exemple ci-dessus $m = 12$, $L_{12} = 3$ et $\ell_{12} = 2$. Les règles de l'évolution de l'automate sont données ainsi : l'automate enregistre L_m et ℓ_m ; au temps m l'automate se trouve dans l'état k, $k \leq 5$, si et seulement si $L_m < 6$ et $\ell_m = k$; il se trouve dans l'état 6 si et seulement si $L_m \geq 6$. On choisit comme état initial l'état 1 qui correspond à l'état de l'automate au temps $t = 1$. Il est facile de vérifier que l'automate a la propriété de Markov et de calculer sa matrice stochastique $\mathbf{M}$, $\mathbf{M}_{k\ell} = P(X_{j+1} = \ell \,|\, X_j = k)$,

$$\mathbf{M} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{9.1}$$

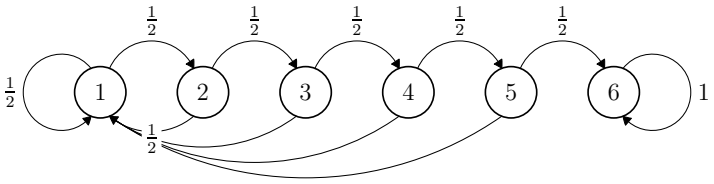


Figure 9.2 – Le graphe associé à la matrice stochastique (9.1).

Cet automate a la particularité qu'une fois dans l'état 6 il y reste pour toujours. L'état 6 est un *état absorbant*. On choisit l'état initial $P(X_1 = 1) = 1$ au temps $t = 1$ par commodité. On obtient

$$\begin{aligned} P(E_{100}(6)) &= P(X_{100} = 6) \\ &= P(X_{100} = 6 \,|\, X_1 = 1) P(X_1 = 1) = \mathbf{M}^{99}_{1,6}\,. \end{aligned}$$

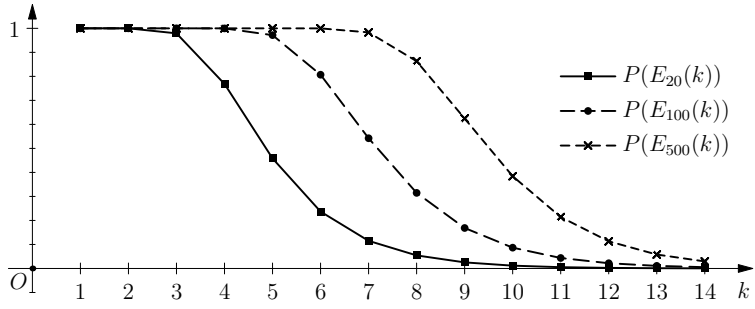


Figure 9.3 – $k \mapsto P(E_n(k))$ pour $n = 20$, $n = 100$ et $n = 500$.

Pour 100 lancers, au centième près, on calcule

k	:	4	5	6	7	8	9	10	11
$P(E_{100}(k))$	:	1,00	0,97	0,81	0,54	0,31	0,17	0,09	0,04

Remarque 9.1 On peut montrer des résultats très précis concernant les longues séquences. Par exemple P. Erdös (1913-1996) et A. Rényi (1921-1970) ont établi le résultat suivant. Soit deux constantes C_1 et C_2 telles que $0 < C_1 < 1 < C_2 < \infty$; alors avec probabilité un, il existe $N_0(\omega, C_1, C_2)$ tel que

$$\lfloor C_1 \log_2 N \rfloor \leq Z_N(\omega) \leq \lfloor C_2 \log_2 N \rfloor \quad \text{si } N \geq N_0(\omega, C_1, C_2)\,,$$

où $Z_N(\omega)$ est la longueur de la plus longue séquence de Piles pour N lancers ; $\lfloor x \rfloor$ est la partie entière de x et $\log_2 x$ est le logarithme en base 2 de x. □

9.2 Chaîne de Markov ergodique I

Une classe importante de chaînes de Markov est celle constituée des *chaînes de Markov ergodiques* : ce sont les chaînes de Markov dont la matrice stochastique $\mathbf{M}$ est *ergodique*, i.e. telle que

$$\exists\, k\,, \quad \mathbf{M}_{ij}^k > 0 \quad \forall\, i, j\,.$$

Si $\mathbf{M}$ est ergodique, chaque état communique avec n'importe quel autre état. Il est facile de donner des exemples de chaînes de Markov qui ne sont pas ergodiques. Il suffit de donner une matrice stochastique telle que deux états ne communiquent pas. Le théorème suivant est le théorème principal pour les chaînes de Markov ergodiques.

Théorème 9.1 *Si* $\mathbf{M}$ *est ergodique, alors il existe une unique mesure de probabilité* π *sur* $\mathbb{A}$ *telle que pour tout* $k \in \mathbb{A}$

$$\pi(k) = \lim_{n\to\infty} \mathbf{M}_{ik}^n > 0 \quad \forall\, i$$

et

$$\sum_{k\in\mathbb{A}} \pi(k)\mathbf{M}_{kj} = \pi(j)\,.$$

Le terme *ergodique* a été introduit en mécanique statistique par L. Boltzmann. Dans la situation particulière considérée ici cela signifie que le système « oublie » sa condition initiale (première affirmation du théorème 9.1) :

$$\forall\, i \in \mathbb{A}\colon \quad P(X_n = k | X_0 = i) \equiv \mathbf{M}_{ik}^n \to \pi(k) \quad \text{si } n \to \infty\,.$$

Si la loi de l'état initial est la mesure de probabilité μ_0, $P(X_0 = j) = \mu_0(j)$, alors on obtient par la formule des probabilités totales la loi de X_n,

$$\begin{aligned} P(X_n = k) &= \sum_{j\in\mathbb{A}} P(X_n = k | X_0 = j) P(X_0 = j) \\ &= \sum_{j\in\mathbb{A}} \mu_0(j) \mathbf{M}^n_{jk} \equiv \mu_n(k)\,. \end{aligned}$$

Lorsque $\mathbf{M}$ est ergodique, dans la limite $n \to \infty$,

$$\lim_{n\to\infty} P(X_n = k) = \lim_{n\to\infty} \sum_{j\in\mathbb{A}} \mu_0(j)\mathbf{M}^n_{jk} = \sum_{j\in\mathbb{A}} \mu_0(j)\pi(k) = \pi(k)\,.$$

En particulier si $\mu_0 = \pi$, les propriétés aléatoires de la chaîne sont les mêmes quel que soit n, dans le sens que pour tout n $P(X_n = j) = \pi(j)$, car

$$\sum_{k\in\mathbb{A}} \pi(k)\mathbf{M}_{kj} = \pi(j)\,.$$

Cette dernière identité s'écrit plus simplement $\pi\mathbf{M} = \pi$ si la mesure de probabilité π est considérée comme un vecteur ligne. La chaîne se trouve dans un régime stationnaire. La mesure de probabilité π sur $\mathbb{A}$ est la *mesure de probabilité stationnaire* ou *mesure invariante* de la chaîne. Dans cette situation l'état de la chaîne change au cours du temps, mais la probabilité de se trouver dans l'état k reste $\pi(k)$ à n'importe quel instant ; la loi de X_n est π pour tout $t = n$. Un point important est la vitesse avec laquelle on atteint le régime stationnaire. La preuve qui suit montre que ce régime est atteint exponentiellement rapidement (voir (9.2)).

Preuve du théorème 9.1 Soit $\mathcal{M}$ l'ensemble de toutes les mesures de probabilité sur $\mathbb{A}$. Pour comparer deux éléments de $\mathcal{M}$ on introduit une distance, la *distance en variation totale*

$$d(\mu,\mu') := \frac{1}{2}\sum_{i\in\mathbb{A}} |\mu(i) - \mu'(i)|\,.$$

Clairement $d(\mu,\mu') \le 1$. On exprime cette distance sous une autre forme. Soit

$$\mathbb{A}^+ := \{i \in \mathbb{A} :\ \mu(i) - \mu'(i) > 0\} \ \text{ et } \ \mathbb{A}^- := \{i \in \mathbb{A} :\ \mu(i) - \mu'(i) < 0\}\,.$$

On peut écrire

$$0 = \sum_{i\in\mathbb{A}} \big(\mu(i) - \mu'(i)\big) = \sum_{i\in\mathbb{A}^+} \big(\mu(i) - \mu'(i)\big) - \sum_{i\in\mathbb{A}^-} \big(\mu'(i) - \mu(i)\big)\,.$$

Par conséquent la distance $d(\mu,\mu')$ s'écrit aussi

$$\begin{aligned} d(\mu,\mu') &= \frac{1}{2}\sum_{i\in\mathbb{A}^+} \big(\mu(i) - \mu'(i)\big) + \frac{1}{2}\sum_{i\in\mathbb{A}^-} \big(\mu'(i) - \mu(i)\big) \\ &= \sum_{i\in\mathbb{A}^+} \big(\mu(i) - \mu'(i)\big)\,. \end{aligned}$$

Soit μ et $\mu' \in \mathcal{M}$; comme $\mathbf{M}$ est une matrice stochastique, $\mu\mathbf{M} \in \mathcal{M}$ et $\mu'\mathbf{M} \in \mathcal{M}$ puisque

$$\sum_{j\in\mathbb{A}}(\mu\mathbf{M})(j) = \sum_{j\in\mathbb{A}}\sum_{i\in\mathbb{A}}\mu(i)\mathbf{M}_{ij} = \sum_{i\in\mathbb{A}}\mu(i)\sum_{j\in\mathbb{A}}\mathbf{M}_{ij} = \sum_{i\in\mathbb{A}}\mu(i) = 1\,.$$

La distance en variation totale vérifie la propriété importante

$$d(\mu\mathbf{M}, \mu'\mathbf{M}) \leq d(\mu, \mu')\,.$$

En effet, si l'on pose comme ci-dessus

$$\widetilde{\mathbb{A}_+} := \big\{j \in \mathbb{A} : \sum_i \mu_i\mathbf{M}_{ij} > \sum_i \mu'_i\mathbf{M}_{ij}\big\}\,,$$

alors

$$\begin{aligned} d(\mu\mathbf{M}, \mu'\mathbf{M}) &= \sum_{j\in\widetilde{\mathbb{A}_+}}\Big(\sum_{i\in\mathbb{A}}\big(\mu(i)-\mu'(i)\big)\mathbf{M}_{ij}\Big) \\ &\leq \sum_{j\in\widetilde{\mathbb{A}_+}}\sum_{i\in\mathbb{A}^+}\big(\mu(i)-\mu'(i)\big)\mathbf{M}_{ij} \\ &= \sum_{i\in\mathbb{A}^+}\big(\mu(i)-\mu'(i)\big)\sum_{j\in\widetilde{\mathbb{A}_+}}\mathbf{M}_{ij} \leq d(\mu,\mu')\,. \end{aligned}$$

On remarque que $\mathbb{A} \neq \widetilde{\mathbb{A}_+}$, car sinon

$$1 = \sum_j\sum_i \mu(i)\mathbf{M}_{ij} > \sum_j\sum_i \mu'(i)\mathbf{M}_{ij} = 1\,.$$

Si $\mathbf{M}_{ij} > 0$ pour tout i, j, il existe $\delta > 0$ tel que

$$\max_i \sum_{j\in\widetilde{\mathbb{A}_+}}\mathbf{M}_{ij} \leq 1-\delta\,.$$

Dans ce cas $d(\mu\mathbf{M}, \mu'\mathbf{M}) \leq (1-\delta)d(\mu,\mu')$.

Soit $\mu_0 \in \mathcal{M}$, $\mu_n := \mu_0\mathbf{M}^n$ et $\varepsilon > 0$. Comme $\mathbf{M}$ est ergodique, il existe $k \in \mathbb{N}$ et $\alpha > 0$ tels que

$$\mathbf{M}^k_{ij} \geq \alpha > 0 \;\; \forall\, i, j \quad \text{et} \quad (1-\alpha)^m < \varepsilon$$

(si m est suffisamment grand). Par itération, pour tout $n \geq 1$,

$$\begin{aligned} d(\mu_{km}, \mu_{km+n}) &= d(\mu_0\mathbf{M}^{km}, \mu_n\mathbf{M}^{km}) \\ &\leq (1-\alpha)\, d(\mu_0\mathbf{M}^{k(m-1)}, \mu_n\mathbf{M}^{k(m-1)}) \\ &\cdots \\ &\leq (1-\alpha)^m d(\mu_0,\mu_n) \leq \varepsilon\,. \end{aligned}$$

Ces estimations montrent que pour tout j, la suite $\mu_n(j)$, $n \geq 1$, est une suite de Cauchy ; par conséquent elle converge. On pose $\pi(j) := \lim_n \mu_n(j)$.

$$\pi\mathbf{M} = \lim_n \mu_n\mathbf{M} = \lim_n \mu_{n+1} = \pi\,.$$

Cette mesure de probabilité π est unique, car si $\pi'\mathbf{M} = \pi'$,

$$d(\pi, \pi') = d(\pi\mathbf{M}^k, \pi'\mathbf{M}^k) \leq (1-\alpha)d(\pi, \pi')\,.$$

Ceci implique $d(\pi, \pi') = 0$, i.e. $\pi = \pi'$. Comme $\pi\mathbf{M}^k = \pi$, $\pi(j) > 0$ pour tout $j \in \mathbb{A}$.

Les estimations ci-dessus donnent aussi une estimation de la vitesse de convergence de μ_n vers π. Il existe une constante $C < \infty$ telle que

$$d(\mu_n, \pi) \leq C \exp\Big(- \ln\Big(\frac{1}{1-\alpha}\Big)^{1/k} n\Big)\,. \tag{9.2}$$

□

Les chaînes de Markov ergodiques sont étudiées aussi dans la section 10.6 en rapport avec la loi des grands nombres. Le chapitre 11 est consacré aux marches aléatoires qui constituent une autre famille d'exemples de chaînes de Markov.

9.3 Exercices

Exercice 9.1 On considère n boules numérotées par $i = 1, 2, \ldots, n$; on aligne ces boules de gauche à droite par ordre croissant de i. On fait une permutation des boules, chaque permutation ayant la même probabilité.
a) Calculer la probabilité de l'événement E_k « la boule k reste à sa place initiale ».
b) Exprimer, à l'aide des événements E_k, l'événement $E(n)$ « aucune boule ne reste à sa place initiale ».
Calculer la probabilité de l'événement $E(n)$.
Indication : on peut calculer la probabilité de l'événement complémentaire en utilisant la proposition 2.2.
c) Déterminer la limite de cette probabilité lorsque $n \to \infty$.
d) Calculer la probabilité de l'événement $F_k(n)$ « k et seulement k boules sont non permutées ».
e) Estimer cette probabilité lorsque n est grand.

Exercice 9.2 a) Soit $2N$ boules dont N sont rouges et N sont noires. On considère deux urnes, U_1 et U_2, contenant chacune N de ces boules. On fait le tirage suivant : on tire au hasard une boule de U_1 et une boule de U_2 ; puis on remet la boule tirée de U_1 dans U_2 ainsi que la boule tirée de U_2 dans U_1. Le nombre de boules noires dans U_1 après le $m^{\text{ième}}$ tirage est noté X_m. Montrer

que le processus stochastique ainsi défini est une chaîne de Markov et calculer sa matrice stochastique

$$P(X_{m+1} = j \,|\, X_m = i)\,.$$

b) On suppose que U_1 et U_2 sont vides. On remplit l'urne U_1 en choisissant au hasard N boules parmi les $2N$ boules. Les autres boules sont mises dans U_2. Calculer la probabilité $\pi(k)$ que U_1 contienne k boules noires et $N-k$ boules rouges.
c) Montrer que la mesure de probabilité π définie au point b) est la mesure invariante de la chaîne.

Exercice 9.3 Soit $\mathbb{A} = \{1, 2, 3, 4, 5, 6\}$.
a) Dessiner le graphe de la matrice stochastique $\mathbf{M} = (\mathbf{M}_{ij})$, $1 \le i, j \le 6$, définie ci-dessous.
b) Quels sont les sous-ensembles des états qui communiquent entre eux ?
c) Un sous-ensemble C de $\mathbb{A}$ est *fermé* si

$$(i \to j \quad \text{et} \quad i \in C) \implies j \in C\,.$$

Trouver les sous-ensembles fermés.
d) Est-ce que $\mathbf{M}$ est ergodique ?

$$\mathbf{M} := \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Exercice 9.4 Dans ce qui suit toutes les matrices sont des matrices stochastiques de type $r \times r$. Un vecteur $\mathbf{x} \in \mathbb{R}^r$ est *non négatif* si et seulement si $x_i \ge 0$ pour tout i.
Montrer que les trois conditions suivantes sont équivalentes.
a) $\mathbf{M}$ est stochastique.
b) Pour tout $\mathbf{x}$ non négatif le vecteur $\mathbf{Mx}$ est non négatif, et le vecteur $\mathbf{1}$ (dont toutes les composantes sont égales à 1) est un vecteur propre à droite de valeur propre 1, $\mathbf{M1} = \mathbf{1}$.
c) Si $\mu = (\mu_1, \cdots, \mu_r)$ est une mesure de probabilité, alors $\mu' := \mu\mathbf{M}$ est une mesure de probabilité.

Exercice 9.5 On considère une chaîne de Markov avec espace des états $\mathbb{A} = \{1, 2, \ldots, r\}$. L'application X_n à valeur dans $\mathbb{A}$, $n \ge 0$, donne l'état de la chaîne au temps n. L'évolution de la chaîne est déterminée par la matrice stochastique $\mathbf{M}$. Soit μ une mesure de probabilité sur $\mathbb{A}$, $P(X_0 = i) = \mu(i)$.
a) Si $a_j \in \mathbb{A}$, $j = i, \ldots, i+k$, exprimer à l'aide de $\mathbf{M}$

$$P(a_i, a_{i+1}, \ldots, a_{i+k}) = P(X_i(\omega) = a_i, \ldots, X_{i+k}(\omega) = a_{i+k})\,.$$

b) On définit deux algèbres de Boole : $\mathcal{F}_{\leq n}$ qui est l'algèbre des événements qui sont des unions finies d'ensembles du type

$$[a_0, \ldots, a_n] := \{X_0 = a_0, \ldots, X_n = a_n\}\,, \quad a_i \in \mathbb{A}\,,$$

et $\mathcal{F}_{\geq n}$ qui est l'algèbre des événements qui sont des unions finies d'ensembles $[b_n, \ldots, b_m]$, $m \geq n$ et $b_i \in \mathbb{A}$.

Si $A \in \mathcal{F}_{\leq n}$ et $B \in \mathcal{F}_{\geq n}$, montrer que

$$P(A \cap B | X_n = i) = P(A|X_n = i)P(B|X_n = i)\,.$$

(A et B sont indépendants conditionnellement à $\{X_n = i\}$).

Exercice 9.6 On lance n fois de façon indépendante une pièce de monnaie équilibrée. Le résultat de l'expérience est décrit complètement en indiquant dans l'ordre la longueur maximale des séquences de Faces et de Piles (voir exemple 9.2) ; par convention on commence toujours par une séquence de Faces, celle-ci étant de longueur zéro si le premier lancer donne Pile. Soit l'événement E « on a obtenu exactement r séquences distinctes de Piles ».

Quelle est la probabilité conditionnelle de l'événement E, sachant qu'on a obtenu lors de l'expérience k Piles (on suppose que $k \geq r$) ?
Indication : si les longueurs des séquences de Faces sont fixées, le nombre de séquences de Piles compatibles avec l'événement $E \cap \{k \text{ Piles}\}$ est égal au nombre des rangements de k boules dans r boîtes, au moins une boule par boîte, sans enregistrer quelle boule va dans quelle boîte.

Exercice 9.7 Soit A un événement concernant une expérience aléatoire, $P(A) = p$, $0 < p < 1$. On répète cette expérience aléatoire de façon indépendante jusqu'à ce que l'événement A se réalise.
Calculer la probabilité qu'on doive répéter l'expérience n fois jusqu'à la première réalisation de A.
Calculer la probabilité qu'on doive répéter l'expérience n fois jusqu'à ce qu'on ait observé pour la première fois r réalisations de A.

Exercice 9.8 Soit la chaîne de Markov définie par la matrice stochastique

$$\mathbf{M} = \begin{pmatrix} 0 & 3/10 & 1/10 & 3/5 \\ 1/10 & 1/10 & 7/10 & 1/10 \\ 1/10 & 7/10 & 1/10 & 1/10 \\ 9/10 & 1/10 & 0 & 0 \end{pmatrix}.$$

a) Dessiner le graphe associé à cette matrice.
b) Calculer la mesure de probabilité stationnaire de la chaîne.

Exercice 9.9 On considère le jeu suivant. Le joueur A choisit une suite de trois lettres de l'alphabet $\{P, F\}$, par exemple FPP. Le joueur B choisit alors une autre suite de trois lettres, par exemple FFP. On lance une pièce de monnaie équilibrée autant de fois que nécessaire jusqu'à ce qu'on observe pour

la première fois une des suites choisies par les joueurs. Le joueur est gagnant si c'est sa suite qui apparaît.
a) Avec le choix ci-dessus, montrer que B gagne avec probabilité $2/3$.
b) Pour le choix ci-dessus, le jeu peut être décrit avec un automate à six états. Donner le graphe qui spécifie cet automate ainsi que la matrice stochastique associée.
Indication : il y a un état initial et deux états terminaux qui correspondent aux événements « A gagne » et « B gagne » (ces deux états sont absorbants).
c) Montrer que PFF gagne sur FFP, FFP gagne sur FPP, FPP gagne sur PPF et PPF gagne sur PFF.

Exercice 9.10 On considère une urne contenant $a \geq 1$ boules blanches et $b \geq 1$ boules noires. A chaque unité de temps on tire une boule ; si la boule est noire le processus s'arrête et si la boule est blanche on la remet dans l'urne et on ajoute une autre boule blanche, puis on refait un tirage.
a) Sachant qu'on a fait $i-1$ tirages, calculer la probabilité que le processus ne s'arrête pas au $i^{\text{ième}}$ tirage.
b) Soit N le nombre de tirages nécessaires pour que le processus s'arrête. Calculer $P(N > k)$ et $P(N = k)$. Montrer que si $b > 1$, $\lim_{k\to\infty} kP(N > k) = 0$; montrer que si $b = 1$, alors ce n'est pas le cas.
c) Montrer que l'espérance de la v.a. N existe si et seulement si le nombre de boules noires $b > 1$.
d) Calculer $\mathbb{E}(N)$ lorsque $b > 1$.
Indication : on peut exprimer $P(N > k)$ comme une différence de deux termes.

Chapitre 10

La loi des grands nombres

Dans les chapitres 10, 11 et 12 on étudie le comportement asymptotique de la somme de n v.a. réelles indépendantes $X_1, \ldots, X_n$ lorsque n devient grand. Dans la plupart des cas les v.a. ont la même loi, i.e. elles sont *identiquement distribuées*. On dit que les v.a. $X_1, \ldots, X_n$ sont *i.i.d.* si elles sont indépendantes, identiquement distribuées. On étudie le comportement de la somme $\sum_i X_i$ sur différentes *échelles* de la manière suivante : on sélectionne (de manière appropriée) une suite monotone $a(n)$, $n \geq 1$, de nombres positifs et on étudie

$$\frac{1}{a(n)} \sum_{i=1}^{n} X_i \quad \text{lorsque } n \to \infty.$$

Les $a(n)$ jouent le rôle de grandeurs de référence pour exprimer la valeur de la somme des v.a. :

$$P\Big(c \leq \frac{1}{a(n)} \sum_{i=1}^{n} X_i \leq d\Big) = P\Big(a(n)\, c \leq \sum_{i=1}^{n} X_i \leq a(n)\, b\Big)\,.$$

Les cas étudiés sont

1) chapitre 10 : $a(n) = n$;
2) chapitre 11 : $a(n) = 1$ pour tout n, i.e. on étudie $\sum_i X_i$;
3) chapitre 12 : $a(n) = O(\sqrt{n})$.

10.1 Théorème de la loi des grands nombres

C'est un des théorèmes les plus importants. Le théorème de la loi des grands nombres (LGN) donne une relation fondamentale entre

probabilité d'un événement et fréquence relative d'un événement,

et plus généralement entre l'espérance et la moyenne empirique. Il permet de comprendre pourquoi un phénomène aléatoire présente des aspects *réguliers* lorsqu'il est examiné sur une échelle appropriée. Une des caractéristiques d'une expérience aléatoire, telle qu'elle est définie dans ce livre, est que l'on peut

répéter cette expérience. On définit la *fréquence relative d'un événement* $A \in \mathcal{F}$ par le rapport

$$\frac{\#\text{ de réalisations de } A \text{ lors de } n \text{ répétitions indépendantes de l'expérience}}{n}\,.$$

La fréquence relative de A est une *variable aléatoire* ; la probabilité $P(A)$ est un nombre réel qui est calculé à partir de l'espace de probabilité $(\Omega, \mathcal{F}, P)$ décrivant l'expérience aléatoire. La LGN permet de comparer $P(A)$ avec la valeur de la fréquence relative de A, i.e. les résultats théoriques (du modèle de l'expérience) avec les résultats empiriques (de l'expérience).

Théorème 10.1 (Loi des grands nombres) *Soit $X_1, X_2, \ldots$ une suite de v.a. i.i.d. possédant une espérance. Alors pour tout $\varepsilon > 0$*

$$\lim_{n\to\infty} P\Big(\Big|\frac{X_1 + \cdots + X_n}{n} - \mathbb{E}(X_1)\Big| \geq \varepsilon\Big) = 0\,.$$

La *moyenne empirique* est la v.a.

$$\frac{1}{n}\sum_{j=1}^{n} X_j\,,$$

et la *moyenne théorique* est le nombre réel $\mathbb{E}(X)$. Dans le cas de v.a. bornées et i.i.d. l'inégalité de Hoeffding donne un résultat beaucoup plus fort, car on obtient aussi une estimation de la vitesse de convergence qui est exponentiellement rapide. Le théorème 10.1 est optimal si les v.a. sont i.i.d. et si l'on ne suppose que l'existence de l'espérance.

Jakob Bernoulli (1654-1705) a montré ce théorème pour la première fois pour les v.a. $X_i = I_A$ où A est un événement donné. Dans ce cas $\mathbb{E}(X_1) = P(A)$ et

$$\frac{1}{n}\sum_{j=1}^{n} X_j = \text{fréquence relative de } A \equiv \mathrm{freq}_n(A)\,.$$

Pour ce cas, on peut choisir dans l'inégalité de Hoeffding a_i et b_i tels que $b_i - a_i = 1$ puisque $X_i = I_A$. Cette inégalité donne la relation suivante entre fréquence d'un événement A et sa probabilité,

$$\forall\,\varepsilon > 0\colon \quad P\Big(\Big|\frac{\mathrm{freq}_n(A) - P(A)}{P(A)}\Big| \geq \varepsilon\Big) \leq 2\exp\big(-2(\varepsilon P(A))^2 n\big)\,. \qquad (10.1)$$

L'inégalité (10.1) donne, en fonction du nombre n de répétitions de l'expérience, une estimation explicite de la probabilité de l'événement «l'erreur relative entre la fréquence $\mathrm{freq}_n(A)$ de A et $P(A)$ est plus grande que ε». Cette probabilité tend vers 0 exponentiellement vite avec le nombre n de répétitions.

Exemple 10.1 On considère l'expérience de l'exemple 4.6. On a une urne qui est composée de n jetons dont k sont noirs, les autres blancs. On fait un tirage sans remise de ces n jetons, mais pendant le tirage on *ne prend pas connaissance* de la couleur des jetons qui sont tirés. Soit A l'événement « le dernier jeton tiré est noir ». Si l'on ne prend pas connaissance de la couleur des jetons qui sont tirés, $P(A) = k/n$ bien qu'il ne reste à ce moment qu'un seul jeton dans l'urne. La valeur de $P(A)$ exprime *notre manque d'information* au sujet de la couleur des jetons. Comment concilier ceci avec le fait que les tirages ne dépendent pas du fait que l'on connaît on non la couleur des jetons ?

La LGN donne la réponse. L'*expérience* qu'on fait *consiste à tirer les jetons sans prendre connaissance de la couleur de ceux-ci*. Si l'on répète n fois *cette même expérience*, lorsque n devient grand, la fréquence de l'événement A est $\approx k/n$ avec une probabilité proche de 1. □

Exemple 10.2 On considère l'expérience de l'exemple 4.5. On modifie le pari de la manière suivante : on *répète l'expérience* et chaque fois que l'événement E est réalisé, i.e. on a tiré 1 jeton noir et 9 blancs, on a la possibilité de gagner 10 francs suisses si l'on désigne l'urne qui a été utilisée pour le tirage des 10 jetons, et on perd 10 francs suisses si l'on donne une mauvaise réponse. Dans ce cas la stratégie à adopter est claire si l'on peut jouer autant de fois qu'on veut : on désigne toujours U_1. En effet, $P(C_1|E) = 0,556$ et la LGN affirme qu'avec probabilité proche de 1 la fréquence relative

$$\frac{\#\text{ de réalisations de } E \cap C_1 \text{ lors des } n \text{ expériences}}{\#\text{ de réalisations de } E \text{ lors des } n \text{ expériences}} \approx 0{,}556$$

si n est suffisamment grand. □

Preuve du théorème 10.1 Lorsque les v.a. sont bornées, le théorème suit de l'inégalité de Hoeffding. Si les v.a. possèdent seulement une variance, on pose

$$Z_n := \frac{1}{n}\sum_{j=1}^{n}\big(X_j - \mathbb{E}(X_j)\big)\,.$$

La loi des grands nombres est alors une conséquence de l'inégalité de Chebyshev. En effet, $\lim\limits_{n\to\infty} \mathrm{Var} Z_n = 0$ puisque les v.a. sont indépendantes,

$$\begin{aligned}\mathrm{Var} Z_n &= \frac{1}{n^2}\sum_{i,j}\mathbb{E}\big((X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j)\big)\\ &= \frac{1}{n^2}\sum_{i}\mathbb{E}\big((X_i - \mathbb{E}(X_i))^2\big) = \frac{1}{n}\mathrm{Var} X_1\,. \qquad (10.2)\end{aligned}$$

Par conséquent, pour tout $\varepsilon > 0$

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_1)\Big| \geq \varepsilon\Big) = P(|Z_n| \geq \varepsilon) \leq \frac{1}{n\varepsilon^2}\mathrm{Var} X_1 \overset{n\to\infty}{\longrightarrow} 0\,.$$

Noter que l'estimation de la vitesse de convergence vers 0 est considérablement moins bonne que dans le cas où l'on peut utiliser l'inégalité de Hoeffding.

Pour traiter le cas général, on utilise une idée simple, exprimée dans le lemme 10.1, qui est de se ramener au cas des v.a. bornées, pour lequel l'inégalité de Chebyshev (ou de Hoeffding) est valable.

Lemme 10.1 *Soit n v.a. $X_1, \dots, X_n$ et $0 < a < \infty$. On introduit les v.a. tronquées $X'_i := X_i I_{\{|X_i| \le a\}}$ qui sont égales à X_i si $|X_i| \le a$ et nulles si $|X_i| > a$. On pose*

$$S_n := X_1 + \cdots + X_n \quad et \quad S'_n := X'_1 + \cdots + X'_n \,.$$

Alors, pour tout c et pour tout t

$$P\big(|S_n - c| > t\big) \le P\big(|S'_n - c| > t\big) + P(S_n \ne S'_n) \,.$$

Preuve L'événement $\{|S_n - c| > t\}$ se décompose en l'union des événements $\{|S_n - c| > t\} \cap \{S'_n = S_n\}$ et $\{|S_n - c| > t\} \cap \{S'_n \ne S_n\}$. □

Soit $\delta > 0$, et $\mu = \mathbb{E}(X_1)$, $c = n\mu$, $t = \varepsilon n$, $a = n\delta$. On décompose les v.a. X_i :

$$X'_i := X_i I_{\{|X_i| \le a\}} \,, \quad X_i = X'_i + X_i I_{\{|X_i| > a\}} \,.$$

La v.a. X'_1 est bornée et on pose $\mu' := \mathbb{E}(X'_1)$. A partir des inégalités

$$\mathrm{Var} Y \le \mathbb{E}(Y^2) \quad \text{et} \quad |X'_1| \le n\delta$$

on obtient

$$\mathrm{Var}\,(X'_1) \le \mathbb{E}(X'_1)^2 \le n\delta\, \mathbb{E}\big(|X'_1|\big) \le n\delta\, \mathbb{E}(|X_1|) \,. \tag{10.3}$$

Avec les notations du lemme 10.1

$$P(S_n \ne S'_n) \le \sum_{k=1}^{n} P(X'_k \ne X_k) = nP(|X_1| > n\delta) \,.$$

Ce lemme permet d'écrire

$$P\Big(\Big|\frac{S_n}{n} - \mu\Big| > \varepsilon\Big) \le P\Big(\Big|\frac{S'_n}{n} - \mu\Big| > \varepsilon\Big) + nP(|X_1| > n\delta) \,. \tag{10.4}$$

Comme $\lim_{a\to\infty} \mathbb{E}(X'_i) = \mathbb{E}(X_i)$ (théorème 7.2), on peut choisir n suffisamment grand de sorte que $|\mu - \mu'| \le \varepsilon/2$. En tenant compte de (10.3) et de l'indépendance des v.a., on obtient par l'inégalité de Chebyshev

$$P\Big(\Big|\frac{S'_n}{n} - \mu\Big| > \varepsilon\Big) \le P\Big(\Big|\frac{S'_n}{n} - \mu'\Big| \ge \frac{\varepsilon}{2}\Big) \le \frac{4n^2\delta\, \mathbb{E}(|X_1|)}{n^2\varepsilon^2} \,.$$

De (10.4) on conclut que

$$P\Big(\Big|\frac{S_n}{n} - \mu\Big| > \varepsilon\Big) \le \frac{4n^2\delta\, \mathbb{E}(|X_1|)}{n^2\varepsilon^2} + nP(|X_1| > n\delta) \,.$$

En prenant d'abord la limite $n \to \infty$, on obtient par le lemme 8.1

$$\forall \delta > 0 : \quad \lim_{n\to\infty} P\Big(\Big|\frac{S_n}{n} - \mu\Big| > \varepsilon\Big) \le \frac{4\delta\,\mathbb{E}(|X_1|)}{\varepsilon^2}\,.$$

Il suffit de prendre maintenant la limite $\delta \to 0$. □

Remarque 10.1 Si l'espérance n'existe pas, le comportement de la moyenne empirique peut être très différent de celui décrit dans le théorème 10.1. Par exemple, si $X_1, X_2, \ldots$ sont i.i.d. et X_i est une v.a. de Cauchy de paramètre 1, alors pour tout n la moyenne empirique est une v.a. de Cauchy de paramètre 1. En effet, $X_1 + \cdots + X_n$ a une loi de Cauchy de paramètre n (proposition 6.4) et la fonction de répartition de la moyenne empirique est

$$F(t) = P\Big(\frac{X_1 + \cdots + X_n}{n} \le t\Big) = P(X_1 + \cdots + X_n \le nt)\,.$$

En dérivant par rapport à t on obtient la densité d'une v.a. de Cauchy de paramètre 1. □

10.2 Processus stochastique faiblement corrélé

La LGN a une validité plus grande que celle indiquée dans le théorème 10.1. La condition i.i.d. des v.a. peut être remplacée par d'autres conditions.

On suppose que les X_i ne sont pas indépendantes (ni identiquement distribuées) et on définit comme dans la preuve du théorème 10.1

$$Z_n := \frac{1}{n}\sum_{j=1}^{n} \big(X_j - \mathbb{E}(X_j)\big)\,.$$

Si $\lim\limits_{n\to\infty} \mathrm{Var} Z_n = 0$, alors par l'inégalité de Chebyshev le théorème 10.1 est vrai. C'est le cas si les X_i sont non corrélées et identiquement distribuées (voir (10.2)).

On peut encore remplacer cette dernière condition par une condition plus faible. Soit $X_1, X_2, \ldots$ un processus stochastique. Le processus stochastique est *faiblement corrélé* si

1) il existe $C < \infty$ tel que $|\mathrm{Cov}(X_i, X_j)| \le C$ pour tout i, j ;
2) il existe $r(n) \ge 0$, $\lim\limits_{n\to\infty} r(n) = 0$ et

$$|\mathrm{Cov}(X_i, X_j)| \le r(|i-j|) \quad \forall\, i, j\,.$$

Théorème 10.2 *Si le processus stochastique est faiblement corrélé, alors*

$$\forall\, \varepsilon > 0\colon \quad \lim_{n\to\infty} P\Big(\Big|\frac{1}{n}\sum_{j=1}^{n}(X_j - \mathbb{E}(X_j))\Big| \ge \varepsilon\Big) = 0\,.$$

Preuve On pose

$$Z_n := \frac{1}{n}\sum_{j=1}^{n}(X_j - \mathbb{E}(X_j))\,,$$

et on vérifie que $\mathrm{Var}Z_n \to 0$ si $n \to \infty$. Soit $p \in \mathbb{N}$;

$$\begin{aligned}\mathrm{Var}Z_n &= \frac{1}{n^2}\sum_{i,j}\mathrm{Cov}(X_i, X_j)\\ &= \frac{1}{n^2}\sum_{i,j\,:\,|i-j|\le p}\mathrm{Cov}(X_i, X_j) + \frac{1}{n^2}\sum_{i,j\,:\,|i-j|>p}\mathrm{Cov}(X_i, X_j)\\ &\le \frac{(2p+1)C}{n} + r(p) \to r(p) \quad \text{si } n \to \infty.\end{aligned}$$

On a utilisé le fait qu'il y a au plus $2p+1$ indices i tels que $|i-j| \le p$ pour j fixé. Mais p est arbitraire ; par conséquent $\lim_n \mathrm{Var}Z_n = 0$. □

Si le processus stochastique $X_1, X_2, \ldots$ est faiblement corrélé et si

$$\lim_{n\to\infty}\mathbb{E}(X_n) = \mu\,,$$

alors l'affirmation du théorème 10.2 peut s'énoncer :

$$\forall\,\varepsilon > 0\colon \qquad \lim_{n\to\infty} P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}X_i - \mu\Big| \ge \varepsilon\Big) = 0\,. \tag{10.5}$$

En effet, si les espérances des X_n convergent vers μ,

$$\forall\,\delta\ \exists\, N_\delta \text{ tel que} \quad n \ge N_\delta \implies |\mu - \mathbb{E}(X_n)| \le \delta\,;$$

cela implique

$$\lim_{n\to\infty}\Big|\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}(X_k) - \mu\Big| \le \delta \quad \forall\,\delta\,.$$

Le résultat (10.5) découle du théorème 10.2.

Exemple 10.3 Voir section 10.6 pour des exemples de processus stochastiques faiblement corrélés. □

Exemple 10.4 Suite et fin des exemples 2.4, 6.4 et 8.4. On considère le cas $h = 0$ qui est le plus intéressant. Les v.a. X_i (voir (2.4)) sont identiquement distribuées, mais non indépendantes, et $\mathbb{E}(X_i) = 0$. Si $\beta \le 1$, alors la fonction $g_{0,\beta}$ a un seul maximum global en $x = 0$. La loi des grands nombres est valide pour les moyennes empiriques M_n de l'aimantation (voir (8.5)) et

$$\forall\,\varepsilon > 0\colon \qquad \lim_{n\to\infty}\frac{1}{n}\ln P\big(|M_n| \ge \varepsilon\big) < 0\,.$$

Cela signifie que les déviations par rapport à la moyenne sont rares (comme pour (8.1)). La LGN est vérifiée, et on a un bon contrôle de la vitesse de convergence.

Si $\beta > 1$, la fonction $g_{0,\beta}$ a deux maxima globaux en $\pm m^*(\beta) \neq 0$ (voir figure 6.8). Un phénomène nouveau apparaît, une *transition de phase du premier ordre* : en l'absence de champ magnétique le système possède une aimantation (par spin) non nulle qui est égale à $\pm m^*(\beta)$. L'aimantation est $m^*(\beta) > 0$ si l'on approche $h = 0$ à partir de $h > 0$, en prenant la limite $h \to 0$ (voir (6.7)). L'aimantation par spin passe de valeurs strictement positives à des valeurs strictement négatives lorsque h passe de valeurs positives à des valeurs négatives. L'aimantation dans ce modèle est un *paramètre d'ordre*.

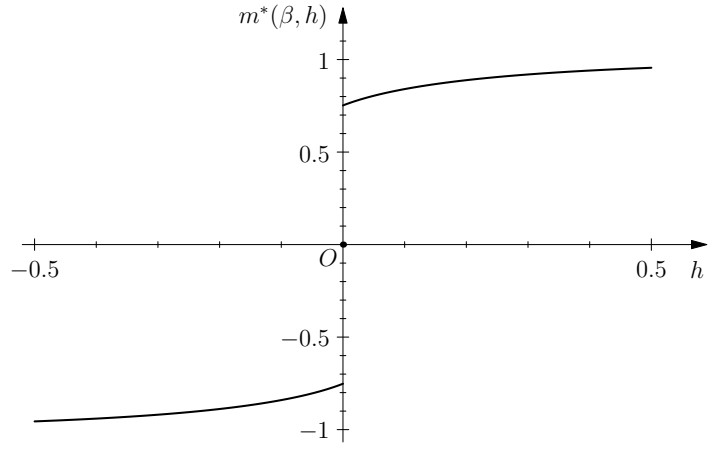


Figure 10.1 Aimantation du modèle de Curie-Weiss pour $\beta = 1{,}3$.

Cette discontinuité de l'aimantation est caractéristique d'une transition de phase du premier ordre. *La loi des grands nombres n'est plus vérifiée* : l'aimantation par spin du système est concentrée en $\pm m^*(\beta)$ et non en 0. Ce sont ces valeurs qui sont observées lorsque n est grand (voir exemple 8.4), alors que l'espérance de l'aimantation par spin est nulle. En effet, si $\varepsilon > 0$ est assez petit, l'intervalle $[-\varepsilon, \varepsilon]$ ne contient pas les maxima globaux $\pm m^*(\beta)$, et par (8.4)

$$\lim_{n\to\infty} \frac{1}{n} \ln P(M_n \in [-\varepsilon, \varepsilon]) = \max_{x\in[-\varepsilon,\varepsilon]} \kappa_{0,\beta}(x) < 0\,.$$

Par conséquent

$$\lim_{n\to\infty} P(|M_n| \geq \varepsilon) = 1 - \lim_{n\to\infty} P(|M_n| \leq \varepsilon) = 1\,.$$

Comme l'aimantation moyenne est concentrée autour de $\pm m^*(\beta)$ et que son espérance est nulle, lorsque n tend vers l'infini l'aimantation moyenne prend la valeur $m^*(\beta)$ avec probabilité $1/2$ et $-m^*(\beta)$ avec probabilité $1/2$. □

10.3 Loi forte des grands nombres

Le contrôle de la vitesse de convergence dans la LGN est important. Pour l'exploiter on fait appel au lemme de Borel (1871-1956) et Cantelli (1875-1966).

Lemme 10.2 (Borel-Cantelli) *Soit A_n une suite d'événements tels que*

$$\sum_{n\geq 1} P(A_n) < \infty\,.$$

Avec probabilité un, seulement un nombre fini de ces événements sont réalisés simultanément.

Preuve Si B est l'événement « un nombre infini des A_n sont réalisés », alors

$$\omega \in B \implies \omega \in \bigcup_{k\geq m} A_k \quad \forall\, m\,.$$

$$P(B) \leq P\Big(\bigcup_{k\geq m} A_k\Big) \leq \sum_{k\geq m} P(A_k) \to 0 \quad \text{si } m \to \infty.$$

□

Soit $X_1, X_2 \ldots$ un processus stochastique tel que $\mathbb{E}(X_i) = \mathbb{E}(X_1)$ existe pour tout i. On pose $S_n = \sum_{i=1}^{n} X_i$, et on suppose que pour tout ε il existe $\delta(\varepsilon, n)$ tel que

$$P\Big(\Big|\frac{S_n}{n} - \mathbb{E}(X_1)\Big| > \varepsilon\Big) \leq \delta(\varepsilon, n) \quad \text{avec} \quad \sum_{n\geq 1} \delta(\varepsilon, n) < \infty\,. \tag{10.6}$$

C'est le cas, par exemple, si les v.a. sont i.i.d. et bornées. Soit

$$A_m(\varepsilon) := \Big\{\Big|\frac{S_m(\omega)}{m} - \mathbb{E}(X_1)\Big| > \varepsilon\Big\}\,.$$

Sous l'hypothèse (10.6), pour tout $\varepsilon > 0$

$$\sum_{n\geq 1} P(A_n(\varepsilon)) < \infty\,.$$

L'événement $E(\varepsilon)$ « un nombre fini des $A_m(\varepsilon)$ sont réalisés » a une probabilité égale à un. Cet événement s'écrit

$$E(\varepsilon) = \Big\{\omega\colon \exists\, N(\varepsilon,\omega) \text{ tel que } m \geq N(\varepsilon,\omega) \text{ implique } \Big|\frac{S_m(\omega)}{m} - \mathbb{E}(X_1)\Big| \leq \varepsilon\Big\}\,.$$

Si l'on pose $\varepsilon = k^{-1}$, $k \geq 1$, alors l'événement

$$E := \bigcap_k E(k^{-1})$$

a encore une probabilité égale à un. Par conséquent, si $\omega \in E$,

$$\lim_m \frac{S_m(\omega)}{m} = \mathbb{E}(X_1)\,.$$

On a donc convergence ponctuelle de $S_m(\omega)/m$ vers la constante $\mathbb{E}(X_1)$ avec probabilité un. Ceci constitue la version forte de la LGN. Dans le cas de v.a. i.i.d. on a le théorème 10.3.

Théorème 10.3 (Loi forte des grands nombres) *Soit $X_1, X_2, \ldots$ une suite de v.a. i.i.d. possédant une espérance. Alors avec probabilité un*

$$\lim_{m\to\infty} \frac{S_m(\omega)}{m} = \mathbb{E}(X_1)\,.$$

10.4 Fonction de répartition empirique

Soit X une v.a. aléatoire et F_X sa fonction de répartition. Comment obtenir des informations sur F_X ?

Si $X_1, \ldots, X_n$ sont i.i.d. et $X_i \overset{\mathcal{L}}{=} X$, on définit la *fonction de répartition empirique* (constante par morceaux)

$$\widehat{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x\}}\,.$$

La fonction $\widehat{F}_n(x)$ est une v.a. puisque qu'elle donne la fréquence relative de l'événement $\{X \le x\}$. Pour une réalisation $x_1, \ldots, x_n$ des v.a. i.i.d. X_i,

$$x \mapsto \widehat{F}_n(x) = \frac{1}{n}\#\{i\colon x_i \le x\}$$

est une fonction de répartition : elle est monotone croissante, continue à droite,

$$\lim_{x\to-\infty} \widehat{F}_n(x) = 0 \quad \text{et} \quad \lim_{x\to\infty} \widehat{F}_n(x) = 1\,.$$

Elle définit donc une loi (discrète) sur $\mathbb{R}$. Si φ est une fonction réelle définie sur $\mathbb{R}$, l'espérance de φ par rapport à cette loi est simplement la moyenne empirique

$$\frac{1}{n}\sum_{i=1}^{n} \varphi(x_i)\,.$$

De l'équation de Hoeffding on obtient : pour tout $\varepsilon > 0$,

$$\sup_{x\in\mathbb{R}} P\big(\big|\widehat{F}_n(x) - F_X(x)\big| \ge \varepsilon\big) \le 2\exp\big(-2\varepsilon^2 n\big) \overset{n\to\infty}{\longrightarrow} 0\,.$$

On peut démontrer un résultat plus fort, le théorème de Glivenko (1897-1940) et Cantelli, en utilisant la loi forte des grands nombres.

Théorème 10.4 (Glivenko-Cantelli) *Soit $X_1, X_2, \ldots$ des v.a. i.i.d. telles que $X_i \stackrel{\mathcal{L}}{=} X$ où X est une v.a. donnée. Soit $\widehat{F}_n$ la fonction de répartition empirique de $X_1, \ldots, X_n$. Alors, avec probabilité un,*

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F_X(x)\right| = 0\,.$$

Preuve On donne la preuve lorsque la fonction de répartition F_X de X est continue. Soit $\varepsilon > 0$; il existe

$$-\infty = x_0 < x_1 < \cdots < x_{k+1} = \infty$$

tels que

$$F_X(x_{i+1}) - F_X(x_i) \leq \frac{\varepsilon}{2} \quad \forall\, i \leq k\,,$$

car la fonction de répartition F_X est monotone et continue. On compare les fonctions F_X et $\widehat{F}_n$ en utilisant le fait qu'elles sont monotones. Pour tout $x \in (x_i, x_{i+1}]$,

$$\widehat{F}_n(x_i) - F_X(x_{i+1}) \leq \widehat{F}_n(x) - F_X(x) \leq \widehat{F}_n(x_{i+1}) - F_X(x_i)$$

et

$$F_X(x_i) - \widehat{F}_n(x_{i+1}) \leq F_X(x) - \widehat{F}_n(x) \leq F_X(x_{i+1}) - \widehat{F}_n(x_i)\,.$$

On en déduit que

$$|\widehat{F}_n(x) - F_X(x)| \leq \max\left\{\widehat{F}_n(x_{i+1}) - F_X(x_i), F_X(x_{i+1}) - \widehat{F}_n(x_i)\right\}.$$

Par conséquent,

$$\begin{aligned}\Delta_n :&= \max_{i=0,\ldots,k} \sup_{x\in(x_i,x_{i+1}]} |\widehat{F}_n(x) - F_X(x)| \\ &= \max_{i=0,\ldots,k} \max\left\{\widehat{F}_n(x_{i+1}) - F_X(x_i), F_X(x_{i+1}) - \widehat{F}_n(x_i)\right\}.\end{aligned}$$

Par la loi forte des grands nombres, avec probabilité un,

$$\lim_{n\to\infty} \widehat{F}_n(x_i) = F_X(x_i) \quad \forall\, i \leq k\,.$$

Par conséquent, avec probabilité un, l'événement

$$E(\varepsilon) := \{\limsup_{n\to\infty} \Delta_n \leq \varepsilon\}$$

est réalisé. Il en est de même pour

$$E := \bigcap_{k\geq 1} E(k^{-1})\,,$$

ce qui prouve le théorème. □

10.5 Principe de la méthode de Monte-Carlo

On présente ici l'idée de base de la méthode de Monte-Carlo, non la méthode d'un point de vue pratique, en considérant le calcul de l'intégrale d'une fonction continue $h : [0,1]^p \to \mathbb{R}$,

$$I := \int_{[0,1]^p} h(t_1, \ldots, t_p)\, dt_1 \cdots dt_p\,.$$

La méthode de Monte-Carlo est une méthode probabiliste qui permet de faire des économies de temps de calcul, mais qui peut parfois conduire à des erreurs à cause de son caractère aléatoire. On introduit les v.a. i.i.d. $X_1, \ldots, X_p$, chacune étant uniformément distribuée sur $[0,1]$. Si

$$Y := h(X_1, \ldots, X_p)\,,$$

$$\mathbb{E}(Y) = \int_{[0,1]^p} h(t_1, \ldots, t_p)\, dt_1 \cdots dt_p\,.$$

Soit $Y_1, Y_2, \ldots, Y_n$ des v.a. i.i.d., $Y_i \sim \mu_Y$. Pour tout $\varepsilon > 0$

$$\lim_{n\to\infty} P\Big(\Big|\frac{1}{n}\sum_{j=1}^{n} Y_j - I\Big| \geq \varepsilon\Big) = 0\,.$$

De façon concrète, on construit un échantillon $\omega = (\omega_1, \ldots, \omega_n)$ de longueur n ; chaque $\omega_i \in [0,1]^p$ est obtenu en utilisant p fois un GNA pour générer les p coordonnées de ω_i, puis on calcule la moyenne empirique

$$\frac{1}{n}\sum_{j=1}^{n} Y_j(\omega_j)\,.$$

La v.a. Y est bornée, $|Y| \leq M$. On utilise l'inégalité de Hoeffding pour estimer la probabilité de l'événement : la moyenne empirique des Y_i donne un résultat qui diffère de I d'au moins $\varepsilon > 0$;

$$P\Big(\Big|\frac{1}{n}\sum_{j=1}^{n} Y_j - I\Big| \geq \varepsilon\Big) \leq 2\exp\Big(-\frac{n\varepsilon^2}{2M^2}\Big)\,.$$

Un point important est que l'estimée ci-dessus est *indépendante de la dimension* p de l'hypercube sur lequel on calcule l'intégrale de h. Lorsque p est grand, l'avantage d'utiliser la méthode de Monte-Carlo devient évident si l'on accepte un certain risque. Le risque est quantifié par la probabilité d'obtenir un résultat dont la précision est moins bonne que ε. Le risque est inférieur à δ si n vérifie

$$2\exp\Big(-\frac{n\varepsilon^2}{2M^2}\Big) \leq \delta\,.$$

10.6 Chaîne de Markov ergodique II

On considère une chaîne de Markov avec r états représentés par les éléments de $\mathbb{A} = \{1, \ldots, r\}$. On suppose que les règles de l'évolution du processus sont indépendantes du temps et sont donc déterminées par une matrice stochastique $\mathbf{M}$,

$$\mathbf{M}_{k\ell} := P(X_{j+1} = \ell \,|X_j = k)\,.$$

Les v.a. qui donnent l'état du processus au temps $0, 1, \ldots$ sont $X_0, X_1, \ldots$. On définit deux autres v.a.

$$N_n(i) := \mathrm{card}\{k : 0 \le k \le n\,,\ X_k = i\}\,;$$
$$N_n(ij) := \mathrm{card}\{k : 0 \le k \le n-1\,,\ X_k = i \text{ et } X_{k+1} = j\}\,.$$

La v.a. $N_n(i)$ donne le nombre de fois que le processus a passé dans l'état i jusqu'au temps n. On peut l'interpréter comme le temps de *séjour du processus stochastique dans l'état i* jusqu'au temps n. La v.a. $N_n(ij)$ indique le nombre de fois que le processus a passé successivement par les états i et j jusqu'au temps n.

On suppose que la matrice stochastique $\mathbf{M}$ est ergodique. La chaîne de Markov possède une mesure invariante π, $\pi\mathbf{M} = \pi$. Dans le régime stationnaire du processus stochastique, la probabilité que celui-ci passe successivement par les états i et j est donnée par $\pi(i)\mathbf{M}_{ij}$.

Théorème 10.5 *Si* $\mathbf{M}$ *est ergodique, alors pour tout* $\varepsilon > 0$

$$\lim_{n\to\infty} P\Big(\Big|\frac{N_n(i)}{n} - \pi(i)\Big| \ge \varepsilon\Big) = 0 \quad \forall\, i\,,$$

$$\lim_{n\to\infty} P\Big(\Big|\frac{N_n(ij)}{n} - \pi(i)\mathbf{M}_{ij}\Big| \ge \varepsilon\Big) = 0 \quad \forall\, i, j\,.$$

La première affirmation donne une interprétation de la mesure stationnaire π : $\pi(i)$ représente le *temps de séjour moyen dans l'état i lorsque n tend vers l'infini.* Dans ce théorème on ne précise pas l'état initial de la chaîne. Les affirmations sont indépendantes de l'état initial (ergodicité).

Preuve La démonstration est semblable dans les deux cas ; on considère le deuxième. C'est une conséquence de la loi des grands nombres pour des v.a. non indépendantes (non identiquement distribuées). Soit deux états i et j. On définit

$$Y_k := \begin{cases} 1 & \text{si } X_k = i \text{ et } X_{k+1} = j \\ 0 & \text{sinon.} \end{cases}$$

On fixe un état initial de la chaîne en donnant la loi de X_0, $\mu(i) := P(X_0 = i)$. On montre que le processus définit par les v.a. Y_k est faiblement corrélé.

Comme $\mathbf{M}$ est ergodique (théorème 9.1)

$$\lim_{n\to\infty} \mathbf{M}^n_{kj} = \pi(j) \quad \forall\, k\,. \tag{10.7}$$

Si l'état initial de la chaîne est donné par μ,

$$P(X_n = i) = \sum_{\ell=1}^{r} \mu(\ell)\mathbf{M}^n_{\ell\, i}\,.$$

Par la propriété de Markov et (10.7)

$$\mathbb{E}(Y_k) = P(X_k = i, X_{k+1} = j) = \sum_{\ell=1}^{r} \mu(\ell)\mathbf{M}^k_{\ell\, i}\mathbf{M}_{ij} \overset{k\to\infty}{\longrightarrow} \pi(i)\mathbf{M}_{ij}\,,$$

et

$$\mathbb{E}(Y_k Y_{k+m}) = \sum_{\ell=1}^{r} \mu(\ell)\mathbf{M}^k_{\ell\, i}\mathbf{M}_{ij}\mathbf{M}^{m-1}_{ji}\mathbf{M}_{ij}\,.$$

Calcul de $\mathrm{Cov}(Y_k, Y_{k+m})$.

$$\begin{aligned}\mathbb{E}(Y_k Y_{k+m}) - \mathbb{E}(Y_k)\mathbb{E}(Y_{k+m}) &= \sum_{\ell=1}^{r} \mu(\ell)\mathbf{M}^k_{\ell\, i}\mathbf{M}_{ij}\mathbf{M}^{m-1}_{ji}\mathbf{M}_{ij} - \mathbb{E}(Y_k)\mathbb{E}(Y_{k+m}) \\ &= \mathbb{E}(Y_k)\big(\mathbf{M}^{m-1}_{ji}\mathbf{M}_{ij} - \mathbb{E}(Y_{k+m})\big)\,.\end{aligned}$$

La suite $\mathbb{E}(Y_n)$ est une suite de Cauchy car elle converge. De ce fait et de (10.7) on obtient, uniformément en $k \geq 1$,

$$\begin{aligned}|\mathbb{E}(Y_{m+k}) - \mathbf{M}^{m-1}_{ji}\mathbf{M}_{ij}| &\leq |\mathbb{E}(Y_{m+k}) - \mathbb{E}(Y_m)| + |\mathbb{E}(Y_m) - \mathbf{M}^{m-1}_{ji}\mathbf{M}_{ij}| \\ &\to 0 \quad \text{si } m \to \infty.\end{aligned}$$

Par conséquent

$$r(m) := \sup_k \big|\mathbb{E}(Y_k Y_{k+m}) - \mathbb{E}(Y_k)\mathbb{E}(Y_{k+m}))\big| \overset{m\to\infty}{\longrightarrow} 0\,.$$

Comme

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}(Y_k) \overset{n\to\infty}{\longrightarrow} \pi(i)\mathbf{M}_{ij}$$

on peut appliquer le théorème 10.2 sous la forme (10.5),

$$P\Big(\Big|\frac{N_n(ij)}{n} - \pi(i)\mathbf{M}_{ij}\Big| \geq \varepsilon\Big) \to 0 \quad 1 \leq i, j \leq r\,.$$

□

10.7 Exercices

Exercice 10.1 On considère une suite i.i.d. de v.a. de Bernoulli de paramètre $0 < p < 1$. On pose $S_n = X_1 + \cdots + X_n$. Montrer que

$$P\Big(\Big|\frac{S_n}{n} - p\Big| \geq \varepsilon\Big) \leq \frac{1}{4\varepsilon^4 n^2}\,.$$

Montrer la loi forte des grands nombres sous ces hypothèses.

Exercice 10.2 Démontrer la première partie du théorème 10.5.

Exercice 10.3 On considère l'espace de probabilité d'un GNA. On pose

$$Y(\omega) := \begin{cases} 0 & \text{si } 0 \le \omega < \frac{1}{2} \\ 1 & \text{si } \frac{1}{2} \le \omega < 1. \end{cases}$$

Soit θ la transformation de $[0,1)$ sur $[0,1)$,

$$\omega \mapsto \theta(\omega) := 2\omega \mod 1\,.$$

On définit pour tout entier $n \ge 0$, $X_n := Y \circ \theta^n$. $X_n(\omega)$ donne le $(n+1)^{\text{ième}}$ chiffre du développement de ω en base 2. Noter que la transformation θ définit un automate *déterministe* (système dynamique) dont l'espace des états est $[0,1)$: si ω est donné, $\theta(\omega)$ est univoquement déterminé.
a) Montrer que pour tout $k \in \mathbb{N}$, les v.a. $X_0, \ldots, X_k$ sont i.i.d.. Donner la loi des X_k.
b) On appelle *motif* m_r *de longueur* r une suite $a_0 a_2 \cdots a_{r-1}$, $a_i \in \{0,1\}$, $i = 0, \ldots, r-1$. Soit m_r un motif fixé (par exemple si $r = 10$, $m_{10} = 0100011110$) ; on définit la v.a. $N_n(\omega; m_r)$ de la manière suivante : $N_n(\omega; m_r)$ est égal au nombre de fois que r chiffres consécutifs coincident avec le motif m_r dans $X_0(\omega) \cdots X_{n-1}(\omega)$. Montrer qu'avec probabilité un

$$\lim_{n\to\infty} \frac{N_n(\omega; m_r)}{n} = \frac{1}{2^r}\,.$$

Indication : décomposer $N_n(\omega; m_r)$ en r sommes constituées chacune par des v.a. i.i.d..
c) On suppose qu'on connaît ω avec une précision finie qui permet de déterminer les dix premiers chiffres $a_0 a_1 \cdots a_9$ du développement en base 2 de ω. Calculer pour $k = 5$ et $k = 10$ (en fonction des b_i)

$$P(X_k = b_0, \ldots X_{k+9} = b_9 | X_0 = a_0, \ldots X_9 = a_9)\,.$$

Le résultat exprime la dépendance sensible du système dynamique par rapport aux conditions initiales.

Exercice 10.4 On considère une densité de probabilité bornée, $f(x) \le M$, qui est nulle en dehors de l'intervalle $[a,b]$. On utilise deux GNA indépendants U_1 et U_2 et on pose

$$X := a + (b-a)U_1 \quad \text{et} \quad Y = MU_2\,.$$

Si $Y \le f(X)$ on garde la valeur de X, sinon on tire un autre point (X,Y) jusqu'à ce que $Y \le f(X)$.
a) Montrer que la densité de la loi de la v.a. Z, qui est définie par ce procédé, est f.
Indication : calculer la fonction de répartition de Z.

b) On considère une fonction continue $g : [0,1] \to [0,1]$. En s'inspirant du premier point, donner une variante de la méthode de Monte-Carlo pour calculer l'intégrale de la fonction g sur $[0,1]$.

Exercice 10.5 Suite de l'exercice 9.10. Soit N la v.a. définie dans l'exercice 9.10 ; on considère une suite de v.a. i.i.d. N_j avec $N_j \stackrel{\mathcal{L}}{=} N$. On suppose que $b = 1$ (une seule boule noire). Montrer que pour tout $\varepsilon > 0$,

$$\lim_{n\to\infty} P\Big(\Big|\frac{N_1 + \cdots + N_n}{n \ln n} - a\Big| \geq \varepsilon\Big) = 0\,.$$

Indication : utiliser la méthode de la preuve du théorème 10.1 en introduisant des v.a. tronquées $N'_j = N_j I_{N_j \leq M}$ et en choisissant $M = \lfloor an \ln n \rfloor$. Montrer que le deuxième terme dans le lemme 10.1 tend vers 0.

Chapitre 11

Marche aléatoire

Dans ce chapitre on étudie l'exemple 2.1 de la marche aléatoire sur $\mathbb{Z}$. A la fin du chapitre on examine aussi le cas où la marche a lieu sur $\mathbb{Z}^2$ et $\mathbb{Z}^3$. Soit X_k, $k = 1, 2, \ldots$, des v.a. i.i.d. de Bernoulli, $P(X_k = \pm 1) = 1/2$. Le marcheur part toujours de 0 ; sa position initiale est $S_0 := 0$. Il fait un pas à chaque unité de temps ; au temps $k-1$ il fait un pas (d'une unité) à droite si $X_k = 1$ et un pas (d'une unité) à gauche si $X_k = -1$. La position du marcheur après n pas (au temps n) est

$$S_n := \sum_{j=1}^{n} X_j \,.$$

Une marche ω jusqu'au temps n est spécifiée par les positions du marcheur $\omega(k)$ en $k = 0, 1, \ldots, n$; l'ensemble de toutes les marches jusqu'au temps n est noté Ω_n. La mesure de probabilité sur Ω_n est la mesure uniforme, $P_n(\{\omega\}) = 2^{-n}$, puisqu'il y a 2^n marches différentes et que chaque marche est également probable. Les v.a. S_k et X_k sont définies par

$$S_k(\omega) := \omega(k) \quad \text{et} \quad X_k(\omega) := \omega(k) - \omega(k-1) \,.$$

Il est commode de représenter une marche par une ligne brisée, appelée aussi *chemin*, comme sur la figure 11.1, et de considérer qu'un chemin est le graphe d'une fonction $t \mapsto \omega(t)$, avec $t \in \mathbb{R}$. Par la suite on omet en général l'indice n dans Ω_n ou P_n.

Cette expérience aléatoire peut être interprétée de différentes manières. Par exemple, il s'agit d'un jeu équitable ; on reçoit 1 franc par partie gagnée et on donne 1 franc par partie perdue. Ici S_n représente le gain (ou perte) du joueur après n parties.

Les v.a. X_k sont i.i.d. ; la loi des grands nombres est vérifiée, mais elle ne donne essentiellement *aucune* information sur cette expérience. En effet, dans la loi des grands nombres la v.a. S_n est exprimée sur l'échelle $a(n) = n$, i.e. $S_n = x\,n$. Or on sait déjà, par l'inégalité de Hoeffding, que $|S_n| \ll n$. De l'inégalité (8.3), avec $t = \sqrt{2a \ln n}$, on obtient

$$P\big(|S_n| \geq \sqrt{2an \ln n}\big) \leq \frac{2}{n^a} \,.$$

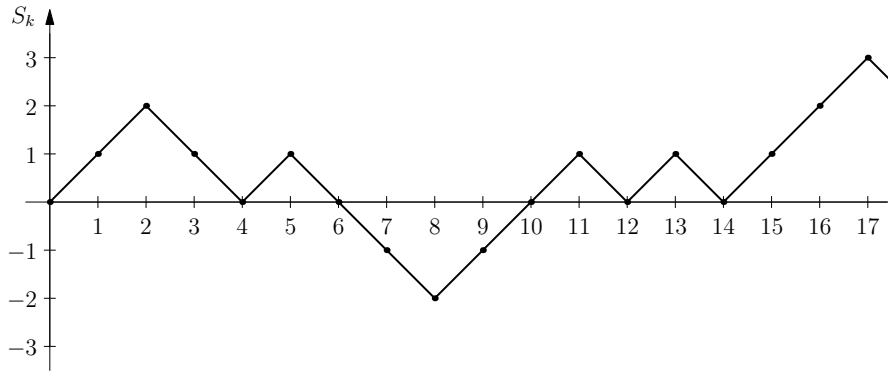


Figure 11.1 – Chemin d'une marche aléatoire.

Si $a > 1$, par le lemme de Borel-Cantelli, on sait qu'avec probabilité un, au plus un nombre fini des événements $\{|S_n| \geq \sqrt{2an \ln n}\}$, $n \geq 1$, sont réalisés. Par conséquent, si $a > 1$, avec probabilité un

$$\limsup_{n\to\infty} \frac{|S_n|}{\sqrt{2an \ln n}} \leq 1 \,.$$

Il existe un résultat beaucoup plus fort et remarquable dû à Khinchine (1894-1959),

Théorème 11.1 (Loi du logarithme itéré) *Si les X_i, $i \geq 1$, sont i.i.d., $\mathbb{E}(X_i) = 0$ et $\mathrm{Var} X_i = 1$, alors avec probabilité un*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n \ln\ln n}} = 1 \quad \textit{et} \quad \liminf_{n\to\infty} \frac{S_n}{\sqrt{2n \ln\ln n}} = -1 \,.$$

En résumé, l'échelle $a(n) = n$ est totalement inappropriée pour étudier le comportement de S_n en détail. Dans ce chapitre on étudie la marche aléatoire sur l'échelle $a(n) \equiv 1$ et dans la section 12.1 sur l'échelle $a(n) = O(\sqrt{n})$.

11.1 Marche aléatoire sur $\mathbb{Z}$, retour à l'origine

Pour calculer $P(S_n = r)$, il suffit de compter le nombre $N_{n,r}$ de marches différentes, partant de 0 au temps $t = 0$ et arrivant à r au temps $t = n$. Soit p le nombre de pas à droite et q le nombre de pas à gauche. On a $p + q = n$ et $p - q = r$, ce qui donne $2p = n + r$.

$$P(S_n = r) = N_{n,r}\, 2^{-n} = \binom{n}{\frac{n+r}{2}} 2^{-n} \,, \quad \binom{x}{y} = 0 \text{ si non défini} \,.$$

Un événement important est le retour à l'origine qui peut avoir lieu seulement aux temps pairs ; on définit $u_0 := 1$ et $u_{2n} := P(S_{2n} = 0)$, $n \geq 1$. Par la formule de Stirling (voir exemple 3.8)

$$u_{2n} := P(S_{2n} = 0) = \binom{2n}{n} 2^{-2n} \simeq \frac{1}{\sqrt{\pi n}}. \tag{11.1}$$

Un autre événement important est le *premier retour à l'origine au temps* $2n$ dont la probabilité est par définition $f_0 := 0$, f_{2n} si $n \geq 1$,

$$f_{2n} := P(S_1 \neq 0, \ldots, S_{2n-1} \neq 0, S_{2n} = 0).$$

En utilisant la propriété de Markov de la marche aléatoire, les probabilités u_{2n} (retour à l'origine au temps $2n$) et f_{2k} (premier retour à l'origine au temps $2k$) sont reliées ainsi :

$$u_{2n} = f_2 u_{2n-2} + f_4 u_{2n-4} + \cdots f_{2n} u_0 \quad \text{si } n \geq 1. \tag{11.2}$$

Le terme $f_2 u_{2n-2} = P(S_2 = 0, S_{2n} = 0)$, $f_4 u_{2n-4}$ est la probabilité de l'événement $\{S_2 \neq 0, S_4 = 0, S_{2n} = 0\}$ etc.

Lemme 11.1 (Principe du miroir) *Soit $s \geq 1$ et $r \geq 1$. Il existe une correspondance $1-1$ entre les chemins qui partent de $A := (1, s)$, passent par $(m, 0)$ pour un $1 < m < n$ et arrivent en $B := (n, r)$, et les chemins qui partent de $A' := (1, -s)$ et arrivent en B.*

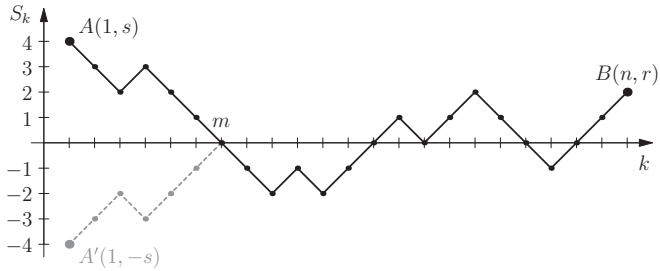


Figure 11.2 – Illustration de la construction du principe du miroir.

Preuve Soit un chemin, qui est le graphe de la fonction $t \mapsto \omega(t)$, telle que $\omega(1) = s$, $\omega(n) = r$ et m est le premier instant tel que $\omega(m) = 0$. Ce chemin part de A, arrive en B et passe par $(m, 0)$. On construit un chemin, qui part de A' et arrive en B, et qui est le graphe de la fonction ω' obtenue par la transformation

$$\omega'(t) := \begin{cases} -\omega(t) & \text{si } t \leq m\,; \\ \omega(t) & \text{sinon.} \end{cases}$$

Si l'on applique cette transformation à ω', on retrouve la fonction ω. Ceci permet de définir une bijection entre les deux ensembles de chemins. □

Le lemme suivant joue un rôle central dans l'analyse de la marche aléatoire.

Lemme 11.2 *Pour la marche aléatoire sur* $\mathbb{Z}$,

1) $P(S_1 \neq 0, \ldots, S_{2n} \neq 0) = P(S_2 \neq 0, S_4 \neq 0 \ldots, S_{2n} \neq 0) = u_{2n}$.

2) $P(S_1 \geq 0, \ldots, S_{2n} \geq 0) = u_{2n}$.

Preuve Si l'événement $\{S_k \neq 0\,,\ 1 \leq k \leq 2n\}$ est réalisé, alors soit l'événement $\{S_1 > 0, \ldots, S_{2n} > 0\}$ est réalisé, soit $\{S_1 < 0, \ldots, S_{2n} < 0\}$ est réalisé. Par symétrie ces événements ont la même probabilité. Pour montrer 1) il suffit de montrer

$$P(S_1 > 0, \ldots, S_{2n} > 0) = \frac{u_{2n}}{2}\,. \tag{11.3}$$

Si $\{S_1 > 0, \ldots, S_{2n} > 0\}$ a lieu, alors $S_1 = 1$ et

$$P(S_1 > 0, \ldots, S_{2n} > 0) = \sum_{r \geq 1} P(S_1 = 1, S_2 > 0, \ldots, S_{2n-1} > 0, S_{2n} = 2r)\,.$$

Par le lemme 11.1, le nombre de chemins, qui partent de $A = (1,1)$ et arrivent en $B = (2n, 2r)$ et tels que $\omega(t) > 0$ pour tout $t \leq 2n$, est égal au nombre de chemins, qui partent de A et arrivent en B, moins le nombre de chemins, qui partent de $A' = (1,-1)$ et arrivent en B ; ce nombre est $N_{2n-1,2r-1} - N_{2n-1,2r+1}$. Par conséquent

$$\begin{aligned} P(S_1 > 0, \ldots, S_{2n-1} > 0, S_{2n} = 2r) &= \frac{1}{2^{2n}} \left(N_{2n-1,2r-1} - N_{2n-1,2r+1}\right) \\ &= \frac{1}{2}\, P(S_{2n-1} = 2r-1) - \frac{1}{2}\, P(S_{2n-1} = 2r+1)\,. \end{aligned}$$

En sommant sur $r \geq 1$ on obtient (la somme est finie)

$$P(S_1 > 0, \ldots, S_{2n} > 0) = \frac{1}{2}\, P(S_{2n-1} = 1)\,. \tag{11.4}$$

D'autre part, si $S_{2n} = 0$, alors $S_{2n-1} = 1$ ou $S_{2n-1} = -1$; u_{2n} s'écrit donc

$$\begin{aligned} u_{2n} &= P(S_{2n-1} = 1 \,\text{et}\, X_{2n} = -1) + P(S_{2n-1} = -1 \,\text{et}\, X_{2n} = 1) \\ &= \frac{1}{2}\, P(S_{2n-1} = 1) + \frac{1}{2}\, P(S_{2n-1} = -1) = P(S_{2n-1} = 1)\,. \end{aligned} \tag{11.5}$$

En effet, par symétrie $P(S_{2n-1} = -1) = P(S_{2n-1} = 1)$, et les v.a. S_{2n-1} et X_{2n} sont indépendantes. Les résultats (11.4) et (11.5) prouvent le point 1).

Si une marche vérifie $S_1 > 0, \ldots S_{2n} > 0$, alors $S_1 = 1$ et $S_k \geq 1$ pour tout $k = 2, \ldots, 2n$. Par conséquent l'équation (11.3) s'écrit

$$\begin{aligned} \frac{u_{2n}}{2} &= P(S_1 > 0, \ldots, S_{2n} > 0) \\ &= P(S_2 \geq 1, \ldots, S_{2n} \geq 1 | S_1 = 1) P(S_1 = 1) \\ &= P(S_1 \geq 0, \ldots, S_{2n-1} \geq 0) P(S_1 = 1) \\ &= P(S_1 \geq 0, \ldots, S_{2n} \geq 0) P(S_1 = 1) = \frac{1}{2} P(S_1 \geq 0, \ldots, S_{2n} \geq 0)\,. \end{aligned}$$

□

A partir de ces résultats on calcule aisément la probabilité du premier retour à l'origine au temps $2n$.

$$\begin{aligned} f_{2n} &= P\Big(S_2 \neq 0, \dots, S_{2n-2} \neq 0,\, S_{2n} = 0\Big) \\ &= P\Big(S_2 \neq 0, \dots, S_{2n-2} \neq 0\Big) - P\Big(S_2 \neq 0, \dots, S_{2n-2} \neq 0,\, S_{2n} \neq 0\Big) \\ &= u_{2n-2} - u_{2n}\,. \end{aligned}$$

En utilisant l'expression de u_{2n-2} on obtient

$$\begin{aligned} f_{2n} &= \frac{(2n-2)!}{(n-1)!\,(n-1)!}\, 2^{-2(n-1)} - u_{2n} \qquad (11.6) \\ &= \overbrace{\frac{(2n)!}{n!\,n!}\, 2^{-2n}}^{u_{2n}} \frac{4n^2}{(2n-1)2n} - u_{2n} = \frac{u_{2n}}{2n-1}\,. \end{aligned}$$

Les événements « premier retour à l'origine au temps $2n$ », $n \geq 1$, sont disjoints,

$$f_2 + f_4 + \cdots \leq 1\,.$$

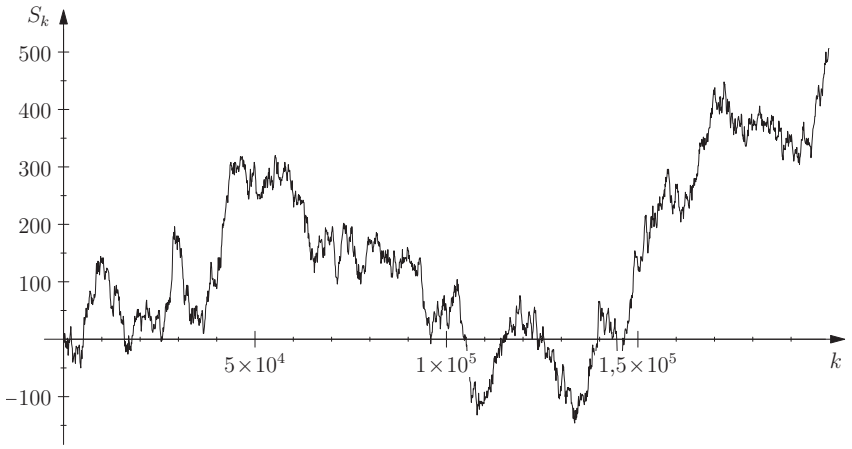


Figure 11.3 – Marche aléatoire sur $\mathbb{Z}$.

Pour la marche aléatoire sur $\mathbb{Z}$

$$\lim_{n\to\infty} \big(f_2 + f_4 + \cdots + f_{2n}\big) = \lim_{n\to\infty} (u_0 - u_{2n}) = u_0 = 1\,. \qquad (11.7)$$

Pour une marche arbitrairement longue on retourne à l'origine avec probabilité un. On définit la v.a. T, *temps du premier retour à l'origine*, par

$$T(\omega) := \min\{k \geq 1 :\, S_k(\omega) = 0\}\,.$$

La loi de T est donnée par

$$P(T = 2n+1) = 0 \quad \text{et} \quad P(T = 2n) = f_{2n}\,, \quad n \geq 0\,.$$

La v.a. T n'a pas d'espérance,

$$\mathbb{E}(T) = \sum_{k\geq 1} 2kP(T = 2k) = \infty\,.$$

On n'a pas de LGN pour T ; on ne peut pas définir un temps de récurrence moyen. La v.a. qui donne le temps du deuxième retour à l'origine s'écrit T_1+T_2 avec T_1 et T_2 indépendantes et $T_i \stackrel{\mathcal{L}}{=} T$, $i = 1, 2$.

Soit $A_r := \{\text{les } r \text{ premiers pas sont faits à droite}\}$; la probabilité conditionnelle de faire le pas suivant à droite, sachant A_r, est $1/2$ puisque les pas sont indépendants. La loi des grands nombres indique que la fréquence relative du nombre de pas à droite tend vers $1/2$ lorsque n diverge. Cependant *on ne peut pas conclure* que lors d'une *marche particulière* les pas à droite sont « compensés » la plupart du temps par les pas à gauche. Le fait que $u_{2n} \approx (\sqrt{\pi n})^{-1}$ et que $\mathbb{E}(T) = \infty$ montrent que ce n'est pas du tout ce qui se passe (sinon on aurait un temps de récurrence moyen et $u_{2n} \not\to 0$ lorsque $n \to \infty$). Les propriétés d'une marche aléatoire sont surprenantes. Ceci est particulièrement manifeste dans la section suivante 11.2.

11.2 Marche aléatoire sur $\mathbb{Z}$, loi de l'arc-sinus

On considère des marches jusqu'au temps $2n$ (marches de $2n$ pas) qui partent de l'origine. On calcule les probabilités des événements suivants

$$\{\text{le dernier passage à l'origine a lieu au temps } 2k\}\,,$$
$$\{\text{le temps passé à droite de l'origine vaut } 2k\}\,.$$

Théorème 11.2 (Loi discrète de l'arc-sinus pour la dernière visite) *Soit Z_{2n} la v.a indiquant le dernier temps de passage à l'origine pour une marche effectuée dans l'intervalle de temps $[0, 2n]$,*

$$Z_{2n}(\omega) := \max\{k :\ S_k(\omega) = 0\,,\ 1 \leq k \leq 2n\}\,.$$

La loi de Z_{2n} est $P(Z_{2n} = 2k + 1) = 0$ et

$$P(Z_{2n} = 2k) = P(Z_{2n} = 2n - 2k) = u_{2k}\, u_{2n-2k}\,.$$

Preuve On écrit

$$\begin{aligned} &P(S_{2k} = 0, S_{2k+2} \neq 0, \ldots, S_{2n} \neq 0) \\ &= P(S_{2k+2} \neq 0, \ldots, S_{2n} \neq 0 | S_{2k} = 0) P(S_{2k} = 0)\,. \end{aligned}$$

En utilisant le lemme 11.2

$$P(S_{2k+2} \neq 0, \ldots, S_{2n} \neq 0 | S_{2k} = 0) = u_{2n-2k}\,,$$

car à partir de $2k$ on effectue encore $2n-2k$ pas sans passer par l'origine. □

La loi de Z_{2n} est symétrique, $P(Z_{2n}=2k)=P(Z_{2n}=2n-2k)$. La probabilité est maximale pour des valeurs petites ou grandes de k. Soit

$$f(x) := \frac{1}{\pi\sqrt{x(1-x)}} \quad 0<x<1\,.$$

On vérifie (voir (5.3))

$$\int_0^1 f(y)\,dy = \pi^{-1}B\Big(\frac{1}{2},\frac{1}{2}\Big) = \pi^{-1}\Big(\Gamma\Big(\frac{1}{2}\Big)\Big)^2 = 1$$

et

$$\int_0^x f(y)\,dy = \frac{2}{\pi}\,\arcsin\sqrt{x}\,.$$

C'est une *loi béta* $B(1/2,1/2)$; on la nomme aussi *loi de l'arc-sinus* à cause de sa fonction de répartition. Dans la limite $n\to\infty$, la fonction de répartition de la v.a. $Z_{2n}/2n$ converge vers la fonction de répartition d'une v.a. de loi de l'arc-sinus. En effet, $u_{2n}\simeq(\pi\,n)^{-1/2}$ et

$$P(Z_{2n}=2k) = P\Big(\frac{Z_{2n}}{2n}=\frac{k}{n}\Big) \simeq \frac{1}{n}\,f(x_k) \quad \text{avec} \quad x_k := \frac{k}{n}\,.$$

La fonction de répartition F_{2n} de $Z_{2n}/2n$ vérifie $(0<x<1)$

$$\begin{aligned} F_{2n}(x) = P\Big(\frac{Z_{2n}}{2n}\le x\Big) &= \sum_{k:\,k\le xn} P(Z_{2n}=2k) \\ &\simeq \sum_{k:\,x_k\le x} \frac{1}{n}\,f(x_k) \qquad \Big(x_{k+1}-x_k=\frac{1}{n}\Big) \\ &\to \int_0^x f(y)\,dy = \frac{2}{\pi}\arcsin\sqrt{x} \equiv A(x)\ . \end{aligned}$$

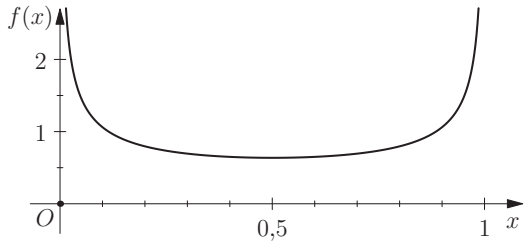


Figure 11.4 – Densité de la loi de l'arc-sinus.

Si $x\simeq 0{,}025 = 1/40$, $A(X)\simeq 0{,}1$. Pour une marche particulière effectuée dans l'intervalle de temps $[0,\,2n]$ et choisie au « hasard » (la probabilité sur

l'ensemble des marches est uniforme), la probabilité que le dernier passage à l'origine ait lieu avant le temps $2n/40$ est $\approx 10^{-1}$. Par exemple, si un marcheur fait un pas toutes les secondes pendant 365 jours, l'événement correspond à passer à l'origine *pour la dernière fois* avant le 10 janvier à 3h. du matin. Ce résultat montre très clairement que pour une marche aléatoire symétrique les « pas à droite » *ne sont pas* compensés la plupart du temps par les « pas à gauche ». Car si c'était le cas, cette probabilité tendrait vers zéro lorsque $n \to \infty$, et la dernière visite à l'origine aurait lieu à un temps $2k$ proche de $2n$.

La loi des grands nombres indique que la fréquence relative de l'événement « le dernier passage à l'origine a lieu avant le temps $2n/40$ » est $\simeq 10^{-1}$ si le nombre N d'observations indépendantes est grand. En moyenne (moyenne empirique), une marche sur 10 passe pour la dernière fois à l'origine avant le temps $2n/40$ si la moyenne empirique est calculée avec N grand.

Théorème 11.3 (Loi discrète de l'arc-sinus pour le temps de séjour) *La probabilité que dans l'intervalle de temps* $[0, 2n]$ *le marcheur passe* $2k$ *unités de temps du côté droit de l'origine et* $2n-2k$ *du côté gauche vaut* $u_{2k}\, u_{2n-2k}$. *Par convention : si* $S_1 = 1$, *alors tant que* $S_k \geq 0$ *on considère qu'on est à droite de l'origine ; si* $S_1 = -1$, *alors tant que* $S_k \leq 0$ *on considère qu'on est à gauche de l'origine.*

Preuve Toutes les marches ont $2n$ pas. Soit $b_{2k,2n}$ la probabilité de l'événement « le temps passé à droite de l'origine vaut $2k$ ». On doit montrer $b_{2k,2n} = u_{2k}\, u_{2n-2k}$. Si $2k = 2n$ (voir lemme 11.2),

$$P(S_1 \geq 0, \ldots, S_{2n} \geq 0) = u_{2n} \quad \Longrightarrow \quad b_{2n,2n} = u_{2n} = u_{2n}\, u_0\,.$$

Par symétrie $b_{0,2n} = u_0\, u_{2n}$. Il reste à considérer les cas $1 \leq k \leq n-1$ où l'on a toujours un premier retour à l'origine au temps $2r < 2n$. Il y a deux possibilités.
a) Les $2r$ premiers pas sont à droite de l'origine, et donc $r \leq k \leq n-1$ et la partie de la marche après le point $(2r, 0)$ contient exactement $2k - 2r$ pas à droite de l'origine. Le nombre de telles marches vaut

$$2^{2r-1} f_{2r} \cdot 2^{2n-2r} b_{2k-2r,2n-2r} = \frac{1}{2}\, 2^{2r} f_{2r} \cdot 2^{2n-2r} b_{2k-2r,2n-2r}\,.$$

b) Les $2r$ premiers pas sont à gauche de l'origine, et donc $r \leq n-k$ et la partie de la marche après le point $(2r, 0)$ contient exactement $2k$ pas à droite de l'origine. Le nombre de telles marches vaut

$$2^{2r-1} f_{2r} \cdot 2^{2n-2r} b_{2k,2n-2r} = \frac{1}{2}\, 2^{2r} f_{2r} \cdot 2^{2n-2r} b_{2k,2n-2r}\,.$$

Par conséquent pour $1 \leq k \leq n-1$,

$$b_{2k,2n} = \frac{1}{2} \sum_{r=1}^{k} f_{2r}\, b_{2k-2r,2n-2r} + \frac{1}{2} \sum_{r=1}^{n-k} f_{2r}\, b_{2k,2n-2r}\,. \tag{11.8}$$

On procède par induction. L'identité $b_{2k,2n} = u_{2k}\, u_{2n-2k}$ est vraie si $n = 1$; supposons qu'elle soit vraie pour $m \leq n-1$. Dans ce cas (11.8) devient (voir (11.2))

$$b_{2k,2n} = \frac{1}{2} u_{2n-2k} \underbrace{\sum_{r=1}^{k} f_{2r}\, u_{2k-2r}}_{u_{2k}} + \frac{1}{2} u_{2k} \underbrace{\sum_{r=1}^{n-k} f_{2r}\, u_{2n-2k-2r}}_{u_{2n-2k}} .$$

□

Corollaire 11.1 *Si $0 < x < 1$, la probabilité qu'au plus xn unités de temps sont passées du côté droit de l'origine, et au moins $(1-x)n$ du côté gauche, tend vers $\frac{2}{\pi} \arcsin \sqrt{x}$ lorsque $n \to \infty$.*

Pour n grand, pendant au moins 97,5 % du temps de la marche, le marcheur est du même côté de l'origine avec probabilité $0{,}1 + 0{,}1 = 0{,}2$. En effet, avec probabilité 0,1 le marcheur passe au plus $1/40 = 0{,}025$ du temps à gauche de l'origine et avec la même probabilité au plus $1/40$ du temps à droite de l'origine.

11.3 Comportement récurrent/transitoire

Les marches aléatoires peuvent être définies sur $\mathbb{Z}^d$, $d \geq 2$, exactement de la même façon, ou sur tout autre graphe (fini ou infini). Dans le cas de $\mathbb{Z}^2$, le marcheur voyage sur les sites de $\mathbb{Z}^2$ et à chaque unité de temps il fait un pas dans une des quatre directions du réseau. La marche est *symétrique* si chaque pas est équiprobable et effectué de manière indépendante. On suppose que le marcheur part de l'origine $O = (0,0)$ comme dans la section 11.1. Sa position au temps n (ou après n pas) est notée S_n, et T désigne comme avant le temps du premier retour à l'origine,

$$T = \min\{k \geq 1 \colon S_k = O\} .$$

L'événement {le marcheur retourne à l'origine} peut s'écrire de deux manières différentes,

$$\{\, \exists n \geq 1 \text{ tel que } S_n = O\} \quad \text{ou} \quad \{T < \infty\} .$$

La probabilité de retourner à l'origine est notée p (sur $\mathbb{Z}$, $p = 1$),

$$p := P(T < \infty) = \sum_{k \geq 1} P(T = k) .$$

La probabilité de revenir au moins deux fois à l'origine est

$$\sum_{k\geq 1} P(T=k)P(\exists\, m\geq 1\colon S_{k+m}=O|T=k) =$$
$$\sum_{k\geq 1} P(T=k)P(\exists\, m\geq 1\colon S_{m}=O|S_0=O) =$$
$$\sum_{k\geq 1} P(T=k)p = p^2\,.$$

De la même manière on montre que la probabilité de visiter au moins k fois l'origine est égale à p^k. Par conséquent,

$$p_k := \text{probabilité de visiter } O \text{ exactement } k \text{ fois} = p^k - p^{k+1} = p^k(1-p)\,.$$

De ce résultat on obtient la propriété importante

$$P(\#\text{ retours en } O \text{ est fini}) = \sum_{k\geq 0} p^k(1-p) = \begin{cases} 0 & \text{si } p=1 \\ 1 & \text{si } p<1. \end{cases}$$

Une telle propriété est appelée *loi zéro-un* : soit elle est vraie avec probabilité un, soit sa négation est vraie avec probabilité un. On peut aussi calculer l'espérance du nombre N de retours en O. Si $p<1$,

$$\mathbb{E}(N) = \sum_{k\geq 1} k\,p_k = \sum_{k\geq 1} kp^k(1-p) = p(1-p)\frac{d}{dp}\frac{1}{1-p} = \frac{p}{1-p}\,.$$

Cette formule est encore valable si $p=1$, car dans ce cas $\mathbb{E}(N)=\infty$. Cette relation fournit un critère pour savoir si $p<1$: il faut et il suffit que $\mathbb{E}(N)<\infty$, i.e.

$$\mathbb{E}(N) = \mathbb{E}\Big(\sum_{n\geq 1} I_{\{S_n=O\}}\Big) = \sum_{n\geq 1} P(S_n=O) < \infty\,.$$

Théorème 11.4 (Pólya (1887-1985)) *Soit une marche aléatoire sur $\mathbb{Z}^d$ partant de l'origine.*

1. *Si $d=1$ et $d=2$, il y a une infinité de retours à l'origine avec probabilité un.*
2. *Si $d\geq 3$, il y a une probabilité non nulle de ne jamais retourner à l'origine ; avec probabilité un le marcheur s'échappe à l'infini.*

Dans le premier cas du théorème, lorsque $p=1$, on dit que la *marche est récurrente* ; dans le deuxième cas la *marche est transitoire*. La probabilité de retour à l'origine pour une marche arbitrairement longue est $p\approx 0{,}34$ si $d=3$ et $p\approx 0{,}1$ si $d=6$. Pour $d=1$ et $d=2$ la marche est récurrente.

Preuve On montre pour $d = 2$ que $\mathbb{E}(N) = \infty$ en calculant $P(S_{2n} = O)$.

$$\begin{aligned} P(S_{2n} = O) &= \frac{1}{4^{2n}} \sum_{k=0}^{n} \underbrace{\binom{2n}{2k}}_{\substack{\text{choix des pas} \\ \text{horizontaux}}} \underbrace{\binom{2k}{k}}_{\substack{\text{choix des pas} \\ \text{vers la gauche}}} \underbrace{\binom{2(n-k)}{n-k}}_{\substack{\text{choix des pas} \\ \text{vers le haut}}} \\ &= \frac{1}{4^{2n}} \frac{(2n)!}{(n!)^2} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \frac{n!}{k!(n-k)!} \\ &= \frac{1}{4^{2n}} \binom{2n}{n} \sum_{k=0}^{n} \binom{n}{k} \binom{n}{n-k} = \frac{1}{2^{2n}} \binom{2n}{n} \frac{1}{2^{2n}} \binom{2n}{n} \\ &\simeq \frac{1}{\pi n} \,. \end{aligned}$$

Par conséquent $\mathbb{E}(N) = \infty$ et donc $p = 1$. On a utilisé l'identité

$$\sum_{k=0}^{n} \binom{n}{k} \binom{n}{n-k} = \binom{2n}{n} \,.$$

(Compter le nombre de sous-ensembles à n éléments d'un ensemble de $2n$ boules dont n sont noires et n blanches). On revient à l'origine avec probabilité un, et donc par la loi zéro-un du retour à l'origine on y revient une infinité de fois.

Si $d \geq 3$, on estime $P(S_{2n} = O)$. Il suffit de considérer le cas $d = 3$.

$$\begin{aligned} P(S_{2n} = O) &= \frac{1}{6^{2n}} \sum_{\substack{k,j,\ell: \\ k+j+\ell=n}} \underbrace{\binom{2n}{2k \quad 2j \quad 2\ell}}_{\substack{\text{choix des pas dans les} \\ \text{directions des trois axes}}} \binom{2k}{k} \binom{2j}{j} \binom{2\ell}{\ell} \\ &= \frac{1}{2^{2n}} \binom{2n}{n} \sum_{\substack{k,j,\ell: \\ k+j+\ell=n}} \Big[\frac{1}{3^n} \binom{n}{k \quad j \quad \ell} \Big]^2 . \end{aligned}$$

On montre (voir ci-dessous) qu'il existe une constante M telle que

$$\Big[\frac{1}{3^n} \binom{n}{k \quad j \quad \ell} \Big] \leq \frac{M}{n} \,. \tag{11.9}$$

Par conséquent (voir (3.2))

$$\begin{aligned} P(S_{2n} = O) &\leq \frac{M}{n} \frac{1}{2^{2n}} \binom{2n}{n} \sum_{\substack{k,j,\ell: \\ k+j+\ell=n}} \frac{1}{3^k} \frac{1}{3^j} \frac{1}{3^\ell} \binom{n}{k \quad j \quad \ell} \\ &= \frac{M}{n} \frac{1}{2^{2n}} \binom{2n}{n} \big(\frac{1}{3} + \frac{1}{3} + \frac{1}{3} \big)^n \\ &\leq \frac{\text{const.}}{n^{3/2}} \,. \end{aligned}$$

Ceci montre que $\mathbb{E}(N) < \infty$. Dans ce cas, si $x \in \mathbb{Z}^3$ et si $S_k = x$, alors la probabilité de revenir en ce site une infinité de fois est nulle. Pour toute boule

de rayon R centrée en O la probabilité de revenir une infinité de fois dans cette boule est nulle, quel que soit le rayon R. Cela signifie que le marcheur s'échappe à l'infini avec probabilité un.

Vérification de (11.9). Si $n = 3m$, alors par (3.2)

$$\frac{1}{3^n}\binom{n}{k\ \ j\ \ \ell} \le \frac{1}{3^{3m}}\binom{3m}{m\ \ m\ \ m} \le \sqrt{\frac{3}{\mathrm{e}}}\,\frac{1}{m}\,.$$

La première inégalité s'obtient en écrivant le quotient

$$\frac{k!\,j!\,\ell!}{m!\,m!\,m!} = \frac{(m+a)!\,(m+b)!\,(m+c)!}{m!\,m!\,m!}\,.$$

Ce quotient est plus grand que 1, car $a+b+c=0$. Si $n = 3m-i$, $i = 1, 2$, on se ramène au cas précédent en utilisant

$$\frac{1}{3^n}\binom{n}{k\ \ j\ \ \ell} \le \frac{3^i}{3^{3m}}\binom{n+i}{k+i\ \ j\ \ \ell}\,.$$

□

11.4 Exercices

Exercice 11.1 On considère une élection avec deux candidats P et Q. Le candidat P obtient p votes et le candidat Q obtient q votes, $p > q$. On procède au dépouillement en tirant successivement au hasard les bulletins de vote de l'urne. Tirage sans remplacement ! Quelle est la probabilité que durant le dépouillement le candidat P a toujours strictement plus de voix que le candidat Q ?
Indication : représenter ce dépouillement par une marche aléatoire.

Exercice 11.2 Pour la marche symétrique sur $\mathbb{Z}$, montrer que pour tout k

$$\lim_{n\to\infty} \sqrt{\pi n} P(S_{2n} = 2k) = 1\,.$$

Exercice 11.3 On considère la marche symétrique sur $\mathbb{Z}$ partant de l'origine. Soit r un entier non négatif et $k \le r$.
a) Montrer que

$$P(\max_{\ell=1}^{n} S_\ell \ge r, S_n = k) = P(S_n = 2r - k)\,.$$

Indication : utiliser le principe du miroir.
b) Calculer la probabilité

$$P(\max_{\ell=1}^{n} S_\ell = r, S_n = k)\,.$$

c) Calculer la probabilité

$$P(\max_{\ell=1}^{n} S_\ell = r)\,.$$

Exercice 11.4 On considère une marche aléatoire asymétrique sur $\mathbb{Z}$,

$$P(X=1)=p \quad \text{et} \quad P(X=-1)=q\,,$$

$p+q=1$ et $p\neq q$. Est-ce que la marche est récurrente ?

Exercice 11.5 Dans $\mathbb{R}^2$ on considère les vecteurs $\mathbf{e}_1=(1,0)$ et $\mathbf{e_2}=(0,1)$, ainsi que les vecteurs $\mathbf{m}_1:=\mathbf{e}_1+\mathbf{e}_2$ et $\mathbf{m}_2:=\mathbf{e}_1-\mathbf{e}_2$. On définit

$$\begin{aligned}\mathbb{L}_1 &:= \{\mathbf{x}=k\mathbf{e}_1 \colon k\in\mathbb{Z}\}\\ \mathbb{L}_2 &:= \{\mathbf{x}=k\mathbf{e}_2 \colon k\in\mathbb{Z}\}\\ \mathbb{L} &:= \{\mathbf{x}=k\mathbf{m}_1+\ell\mathbf{m}_2 \colon k\in\mathbb{Z} \text{ et } \ell\in\mathbb{Z}\}\,.\end{aligned}$$

Le graphe $\mathbb{L}$ est isomorphe à $\mathbb{Z}^2$ et on considère sur $\mathbb{L}$ la marche aléatoire symétrique. Montrer que c'est équivalent à considérer deux marches aléatoires symétriques et indépendantes sur $\mathbb{L}_1$ et $\mathbb{L}_2$.

Chapitre 12

Théorème de la limite centrale

Comme mentionné au début du chapitre 10 la quantité S_n, somme de v.a. i.i.d., peut être étudiée sur trois échelles ; elle se comporte de manière très différente sur chacune des trois échelles.

Dans le chapitre 11 on a considéré S_n sur l'échelle $a(n) \equiv 1$. En empruntant une terminologie propre à la physique, on peut dire que dans le chapitre 11 on étudie S_n sur l'*échelle microscopique*. Le comportement de S_n est sans aucune régularité et souvent inattendu et surprenant. C'est le régime où les effets aléatoires sont maximaux ; ce régime persiste pour tout n. Par opposition, dans le chapitre 10 l'échelle de référence est $a(n) = n$. C'est une *échelle macroscopique* : l'unité caractéristique pour exprimer S_n est proportionnelle à n, le nombre de v.a.. Sur cette échelle il y a de la régularité (LGN). Le caractère aléatoire disparait lorsque n diverge. Si $X_1, X_2, \dots$ sont i.i.d. et possèdent une espérance, l'événement $(t > 0)$

$$\Big\{\frac{1}{n}\sum_i^n X_i \geq \mathbb{E}(X_1) + t\Big\} = \Big\{\sum_i^n X_i \geq \mathbb{E}\Big(\sum_i^n X_i\Big) + nt\Big\}$$

exprime une grande déviation par rapport à la moyenne. L'inégalité de Hoeffding montre que ces grandes déviations sont *rares* lorsque $n \to \infty$. Cette inégalité suggère un comportement différent lorsqu'on étudie les *petites déviations par rapport à la moyenne*, i.e. des événements du type $(t > 0)$

$$\Big\{\frac{1}{n}\sum_i^n X_i \geq \mathbb{E}(X_1) + \frac{t}{\sqrt{n}}\Big\} = \Big\{\sum_i^n X_i \geq \mathbb{E}\Big(\sum_i^n X_i\Big) + \sqrt{n}t\Big\}\,.$$

(Voir (8.3) section 8.2). Cette échelle intermédiaire, qu'on peut appeler *échelle mésoscopique*, correspond à $a(n) = \sqrt{n}$. Sur cette échelle la v.a. pertinente est

$$Y_n := \frac{\sum_{i=1}^n \big[X_i - \mathbb{E}(X_i)\big]}{\sqrt{n}}\,.$$

Il y a un fait remarquable et fondamental : l'émergence d'un phénomène aléatoire « *universel* » (au sens des classes d'universalité des physiciens). Si n diverge, le comportement des v.a. Y_n est *le même* asymptotiquement dès que $\mathbb{E}(X_1)$ et $\mathbb{E}(X_1^2)$ sont fixés ; de plus les petites déviations sont *typiques*.

12.1 La loi binomiale et la loi normale

La loi binomiale $B_i(n;p)$ est une des lois fondamentales de la théorie des probabilités. On a vu dans la section 6.6 que dans la limite des événements rares, $n\,p \to \lambda$ et $n \to \infty$, cette loi converge vers la loi de Poisson π_λ. Ici on étudie une autre limite de cette loi dans le cadre de la marche aléatoire sur $\mathbb{Z}$. On montre par un calcul explicite que sur l'échelle $a(n) = \sqrt{n}$ la loi de $S_n/\sqrt{n}$ tend vers une loi gaussienne. Puis on montre la relation entre la marche aléatoire et le mouvement Brownien.

Calcul de $P(S_{2n} = 2r)$ On fait j pas à droite et l pas à gauche, $j + l = 2n$ et $j - l = 2r$, ce qui donne $j = n + r$.

$$\text{Prob}\,(S_{2n} = 2r) = \frac{1}{2^{2n}} \binom{2n}{n+r} = \frac{1}{2^{2n}} \frac{(2n)!}{(n+r)!(n-r)!}\,.$$

En utilisant la formule de Stirling, $n! \simeq \sqrt{2\pi n}\, n^n\, \mathrm{e}^{-n}$,

$$\begin{aligned}
\text{Prob}\,(S_{2n} = 2r) &\simeq \frac{\sqrt{2\pi 2n}}{\sqrt{2\pi(n+r)}\sqrt{2\pi(n-r)}} \\
&\quad \cdot \frac{(2n)^{2n}\,\mathrm{e}^{-2n}\,2^{-2n}}{(n+r)^{n+r}\,(n-r)^{n-r}\,\mathrm{e}^{-n-r}\,\mathrm{e}^{-n+r}} \\
&= \frac{1}{\sqrt{\pi n}} \frac{1}{\sqrt{1-\frac{r^2}{n^2}}} \Big(1+\frac{r}{n}\Big)^{-n-r} \Big(1-\frac{r}{n}\Big)^{-n+r} \\
&= \frac{1}{\sqrt{\pi n}} \frac{1}{\sqrt{1-\frac{r^2}{n^2}}} \frac{1}{\big(1-\frac{r^2}{n^2}\big)^n} \Big(\frac{1-\frac{r}{n}}{1+\frac{r}{n}}\Big)^r\,.
\end{aligned}$$

On écrit

$$\frac{1-\frac{r}{n}}{1+\frac{r}{n}} = 1 - \frac{2r}{n} + O(r^2 n^{-2})\,.$$

On obtient un résultat non trivial lorsque $n \to \infty$ si $r \simeq \sqrt{n}$. On définit x_r par l'équation

$$r \equiv x_r \frac{\sqrt{n}}{\sqrt{2}}\,.$$

Avec cette notation,

$$\Big(1-\frac{r^2}{n^2}\Big)^n = \Big(1-\frac{x_r^2}{2n}\Big)^n \to \mathrm{e}^{-\frac{x_r^2}{2}}$$

et

$$\Big(1-\frac{2r}{n}\Big)^r = \Big(1-\sqrt{\frac{2}{n}}\,x_r\Big)^{\sqrt{\frac{n}{2}}\,\cdot\,x_r} \to \mathrm{e}^{-x_r^2}\,.$$

Par conséquent

$$\text{Prob}\,(S_{2n} = 2r) = \text{Prob}\,\Big(\frac{S_{2n}}{\sqrt{2n}} = x_r\Big) \simeq \frac{1}{\sqrt{\pi n}} \frac{\mathrm{e}^{-x_r^2}}{\mathrm{e}^{-\frac{x_r^2}{2}}} = \frac{1}{\sqrt{\pi n}}\,\mathrm{e}^{-\frac{x_r^2}{2}}\,.$$

Comme $x_{r+1} - x_r = \frac{\sqrt{2}}{\sqrt{n}}$, on en déduit

$$\begin{aligned}\mathrm{Prob}\left(a \le \frac{S_{2n}}{\sqrt{2n}} \le b\right) &= \sum_{a \le x_r \le b} \mathrm{Prob}\left(\frac{S_{2n}}{\sqrt{2n}} = x_r\right) \\ &\simeq \sum_{a \le x_r \le b} \frac{1}{\sqrt{\pi n}}\, \mathrm{e}^{-\frac{x_r^2}{2}} \\ &= \sum_{a \le x_r \le b} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x_r^2}{2}} (x_{r+1} - x_r) \\ &\to \int_a^b \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}\, dt\,.\end{aligned}$$

On examine le comportement des marches aléatoires, lorsque n tend vers l'infini, d'une manière un peu différente, en prenant une limite du continu de la manière suivante. On fixe un temps macroscopique t ; on fait un pas à chaque unité de temps τ (temps microscopique) ; ces deux échelles de temps sont liées par la relation $t = n\tau$. La longueur d'un pas est h et on pose

$$Y_k := hX_k \quad \text{et} \quad \tilde{S}_n := \sum_{i=1}^{n} Y_i = h \sum_{i=1}^{n} X_i\,.$$

La v.a. $\tilde{S}_n$ donne la position du marcheur après n pas. La variance de $\tilde{S}_n$,

$$\mathrm{Var}\tilde{S}_n = n\mathrm{Var}(Y_1) = nh^2\,,$$

donne l'espérance du carré de la distance du marcheur à l'origine. Soit D une constante fixée ; on choisit h en fonction de n de sorte que

$$h^2 n = \frac{h^2}{\tau} n\tau = \frac{h^2}{\tau} t = Dt \quad \text{i.e.} \quad h = \sqrt{\tau D} = \sqrt{\frac{tD}{n}}\,.$$

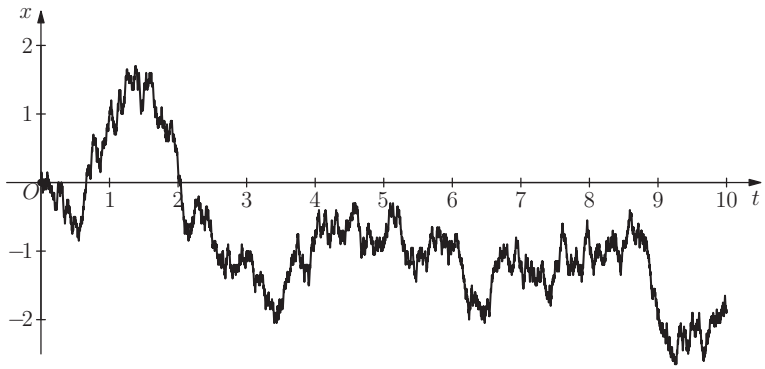


Figure 12.1 – Vers la limite du continu : $D = 1$, $h = \sqrt{\tau}$ et $\tau = 0{,}0025$.

On étudie sous ces conditions la *limite du continu* $\tau \to 0$. On note que la « vitesse » h/τ du marcheur diverge dans cette limite : le mouvement est de plus en plus rapide et erratique. Par le calcul précédent on obtient, dans la limite $n \to \infty$, la fonction de répartition de la loi de la position du marcheur au temps t :

$$\begin{aligned} P\big(\tilde{S}_n \le x\big) &= P\Big(\sqrt{\frac{Dt}{n}} S_n \le x\Big) \\ &= P\Big(\frac{S_n}{\sqrt{n}} \le \frac{x}{\sqrt{Dt}}\Big) \\ &\to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x}{\sqrt{Dt}}} \mathrm{e}^{-\frac{s^2}{2}}\, ds \\ &= \frac{1}{\sqrt{2\pi Dt}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{u^2}{2Dt}}\, du\,. \end{aligned}$$

Si B_t désigne la position du marcheur au temps t, dans cette limite, B_t est une v.a. gaussienne $N(0, Dt)$.

Plus généralement, en utilisant la propriété de Markov de la marche aléatoire, on peut calculer comme précédemment la probabilité que le marcheur passe au temps (macroscopique) t dans l'intervalle (a, b) et arrive au temps $t + s$ ($s = m\tau$) dans l'intervalle (c, d). Pour simplifier un peu l'écriture on choisit $D = 1$.

$$\begin{aligned} &\sum_{a<x<b} \sum_{c<y<d} P\big(\tilde{S}_n = x\big) P\big(\tilde{S}_{n+m} = y | \tilde{S}_n = x\big) \\ &= \sum_{a<x<b} \sum_{c<y<d} P\big(\tilde{S}_n = x\big) P\big(\tilde{S}_m = y - x\big) \\ &= \sum_{a<x<b} \sum_{c<y<d} P\Big(\frac{S_n}{\sqrt{n}} = \frac{x}{\sqrt{t}}\Big) P\Big(\frac{S_m}{\sqrt{m}} = \frac{y-x}{\sqrt{s}}\Big) \\ &\to \int_a^b dx \int_c^d dy \frac{1}{\sqrt{2\pi t}} \mathrm{e}^{-\frac{x^2}{2t}} \frac{1}{\sqrt{2\pi s}} \mathrm{e}^{-\frac{(y-x)^2}{2s}} \qquad \text{(lorsque } \tau \to 0) \\ &= P\big(B_t \in (a,b), B_{t+s} \in (c,d)\big)\,. \end{aligned}$$

Ce calcul montre que les v.a. $B_{t+s} - B_t$ et B_s ont la même loi $N(0, s)$ et que les v.a. B_t et $B_{t+s} - B_t$ sont indépendantes. La généralisation au cas

$$P(B_{t_1} \in I_1, B_{t_1+t_2} \in I_2, \ldots, B_{t_1+\cdots+t_k} \in I_k)$$

où les I_j sont des intervalles est immédiate. Les v.a. B_t qui sont indéxées par le paramètre continu $t \in \mathbb{R}^+$ définissent un *processus stochastique continu* appelé *mouvement Brownien*. Pour tout $t_1, \ldots, t_k$ et $k \in \mathbb{N}$ arbitraire les v.a.

B_{t_1}, $B_{t_2} - B_{t_1}$, $\ldots$, $B_{t_k} - B_{t_{k-1}}$ sont indépendantes ;
$B_{t_1} \sim N(0, t_1)$ et $B_{t_j} - B_{t_{j-1}} \sim N(0, t_j - t_{j-1})$, $2 \le j \le n$.

12.2 Théorème de De Moivre-Laplace

Dans cette section on démontre le théorème de De Moivre (1667-1754) et Laplace (1749-1827), par un calcul explicite semblable à celui de la section 12.1, mais plus précis. Ce théorème est un cas particulier du théorème de la limite centrale de la section suivante. La démontration consiste principalement à établir le lemme 12.1 qui a son propre intérêt.

Théorème 12.1 (De Moivre-Laplace) *Soit X_k, $k \geq 1$, une suite de v.a. i.i.d. de Bernoulli de paramètre $0 < p < 1$. Alors pour tout $-\infty < a, b < \infty$*

$$\lim_{n\to\infty} P\Big(a < \frac{S_n - pn}{\sqrt{np(1-p)}} \leq b\Big) = \frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-\frac{x^2}{2}}\, dx\,.$$

Preuve L'analyse repose sur le lemme 12.1.

Lemme 12.1 *Soit $0 < A < \infty$, $q = 1 - p$ et*

$$v_{nk} := \frac{k - np}{\sqrt{npq}}\,.$$

Si $|v_{nk}| \leq A$, alors il existe des constantes d_1 et d_2 indépendantes de n telles que

$$\underbrace{\mathrm{e}^{-\frac{d_1}{\sqrt{n}}}\Big(1 - \frac{d_2}{\sqrt{n}}\Big)}_{\equiv D_-(n)} \leq \left| \frac{\binom{n}{k}\, p^k\, q^{n-k}}{\frac{1}{\sqrt{2\pi npq}}\mathrm{e}^{-\frac{v_{nk}^2}{2}}} \right| \leq \underbrace{\mathrm{e}^{\frac{d_1}{\sqrt{n}}}\Big(1 + \frac{d_2}{\sqrt{n}}\Big)}_{\equiv D_+(n)}\,.$$

On donne la démonstration du théorème puis celle du lemme. Soit k une valeur de $S_n = \sum_{j=1}^n X_j$; pour alléger l'écriture on pose $v_{nk} \equiv v_k$.

$$S_n = k \iff \frac{S_n - pn}{\sqrt{npq}} = v_k\,.$$

$$P\Big(a < \frac{S_n - pn}{\sqrt{npq}} \leq b\Big) = \sum_{k:\ a < v_k \leq b} \binom{n}{k}\, p^k\, q^{n-k}\,.$$

On suppose que

$$v_{j-1} \leq a < v_j < \cdots < v_l \leq b < v_{\ell+1}\,.$$

Les quantités v_k vérifient l'hypothèse du lemme 12.1 car a et b sont fixés. Comme

$$v_{k+1} - v_k = \frac{1}{\sqrt{npq}}\,,$$

on obtient

$$D_-(n)\Big(\sum_{k=j}^{\ell} \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{v_k^2}{2}} \underbrace{(v_{k+1} - v_k)}_{\frac{1}{\sqrt{npq}}}\Big) \leq \sum_{k=j}^{\ell} \binom{n}{k}\, p^k\, q^{n-k}$$

et

$$\sum_{k=j}^{\ell} \binom{n}{k} p^k q^{n-k} \le D_+(n)\Big(\sum_{k=j}^{\ell} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{v_k^2}{2}} (v_{k+1} - v_k) \Big) .$$

Le théorème est démontré en prenant la limite $n \to \infty$. □

Preuve du lemme Par calcul en utilisant la formule de Stirling ; n est fixé et on pose $v_{nk} \equiv v_k$. On utilise le lemme suivant.

Lemme 12.2 *Si* $|x| \le 2/3$, *alors* $| \ln(1+x) - (x - \frac{x^2}{2}) | \le |x|^3$.

Preuve $\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots \equiv x - \frac{x^2}{2} + \theta(x)$, avec

$$|\theta(x)| \le \sum_{n\ge 3} \frac{|x|^n}{n} \le \frac{1}{3} \sum_{n\ge 3} |x|^n = \frac{1}{3} \frac{|x|^3}{1-|x|} \le |x|^3 \quad \text{si } |x| \le 2/3 .$$

□

Par définition

$$\begin{aligned} v_k \equiv v_{nk} = \frac{k-np}{\sqrt{npq}} &\iff k = np + v_k \sqrt{npq} \\ &\iff n-k = nq - v_k \sqrt{npq} . \end{aligned}$$

La formule de Stirling,

$$c_1(n) \sqrt{2\pi n}\, n^n \, \mathrm{e}^{-n} < n! < c_2(n) \sqrt{2\pi n}\, n^n \, \mathrm{e}^{-n}$$

avec $c_1(n)$, $c_2(n) \to 1$ si $n \to \infty$, permet d'écrire

$$\frac{c_1(n)}{c_2(k) c_2(n-k)} \varphi(n,k) \sqrt{\frac{n}{2\pi(n-k)k}} \le \binom{n}{k} p^k q^{n-k}$$

et

$$\binom{n}{k} p^k q^{n-k} \le \frac{c_2(n)}{c_1(k) c_1(n-k)} \varphi(n,k) \sqrt{\frac{n}{2\pi(n-k)k}} .$$

Dans ces expressions

$$\begin{aligned} \varphi(n,k) &= \Big(\frac{np}{k}\Big)^k \Big(\frac{nq}{n-k}\Big)^{n-k} \\ &= \Big(\frac{k - \sqrt{npq} v_k}{k}\Big)^k \Big(\frac{n-k+\sqrt{npq} v_k}{n-k}\Big)^{n-k} \\ &= \Big(1 - \frac{\sqrt{npq}}{k} v_k\Big)^k \Big(1 + \frac{\sqrt{npq}}{n-k} v_k\Big)^{n-k} . \end{aligned}$$

Par conséquent

$$\begin{aligned}\ln\varphi(n,k) &= k\left[-\frac{\sqrt{nqp}}{k}v_k - \frac{npq}{2k^2}v_k^2 + \theta\Big(-\frac{\sqrt{npq}}{k}v_k\Big)\right]\\ &\quad + (n-k)\left[\frac{\sqrt{npq}}{n-k}v_k - \frac{npq}{2(n-k)^2}v_k^2 + \theta\Big(\frac{\sqrt{npq}}{n-k}v_k\Big)\right]\\ &= -\frac{n^2pq}{2k(n-k)}v_k^2 + k\theta\Big(-\frac{\sqrt{npq}}{k}v_k\Big) + (n-k)\theta\Big(\frac{\sqrt{npq}}{n-k}v_k\Big)\,.\end{aligned}$$

Enfin

$$\begin{aligned}k(n-k) &= \Big(np + v_k\sqrt{npq}\Big)\Big(nq - v_k\sqrt{npq}\Big)\\ &= n^2pq - v_k^2npq + nqv_k\sqrt{npq} - v_knp\sqrt{npq}\\ &= n^2pq\Big(1 - \frac{v_k^2}{n} + \frac{qv_k}{\sqrt{npq}} - \frac{pv_k}{\sqrt{npq}}\Big)\\ &\equiv n^2pq(1-\delta(n,k))\,.\end{aligned}$$

En mettant ces résultats ensemble, il existe, uniformément en $|v_k| \le A$, des constantes d_1 et d_2 telles que

$$D_-(n)\frac{1}{\sqrt{2\pi npq}}\mathrm{e}^{-\frac{v_{nk}^2}{2}} \le \binom{n}{k}\, p^k\, q^{n-k} \le D_+(n)\frac{1}{\sqrt{2\pi npq}}\mathrm{e}^{-\frac{v_{nk}^2}{2}}\,.$$

□

Pour des v.a. de Bernoulli, la somme S_n obéit à une loi binomiale $\mathcal{B}_i(n,p)$. Lorsque n est grand, le théorème 12.1 (ou le théorème 12.2 dans des cas plus généraux) permet de calculer approximativement

$$P\Big(a \le \frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{Var}S_n}} \le b\Big) \approx \frac{1}{\sqrt{2\pi}}\int_a^b \mathrm{e}^{-\frac{t^2}{2}}\,dt\,.$$

Il faut faire attention au fait que a et b sont des constantes. Si la probabilité de l'événement est petite, il faut que n soit grand si l'on veut utiliser cette approximation et avoir une erreur relative petite.

Remarque 12.1 La loi $B_i(n;p)$ est discrète et donc aussi celle de T_n,

$$T_n = \frac{S_n - np}{\sqrt{np(1-p)}}\,.$$

Si B est l'ensemble des valeurs de la v.a. T_n,

$$1 = P(T_n \in B) \ne \frac{1}{\sqrt{2\pi}}\int_B \mathrm{e}^{-\frac{s^2}{2}}\,ds = 0\,.$$

Pour une approximation ponctuelle de la loi $\mathcal{B}_i(n,p)$ voir le lemme 12.1. □

Exemple 12.1 On réalise 100 lancers d'une pièce de monnaie équilibrée. La v.a. qui compte le nombre de Piles est

$$S_{100} = \sum_{k=1}^{100} X_k \quad \text{avec} \quad \mathbb{E}(S_{100}) = 50 \text{ et } \mathrm{Var} S_{100} = 25\,.$$

D'après le théorème 12.1

$$\begin{aligned} P\big(50 - 5x \leq S_{100} \leq 50 + 5x\big) &= P\Big(\Big|\frac{S_{100} - 50}{5}\Big| \leq x\Big) \\ &\simeq \frac{1}{\sqrt{2\pi}} \int_{-x}^{+x} \mathrm{e}^{-\frac{t^2}{2}}\, dt\,. \end{aligned}$$

Dans cet exemple, une déviation standard $SD = \sqrt{\mathrm{Var} S_{100}} = 5$. La probabilité d'observer une valeur de $S_{100} \in [50 - SD,\, 50 + SD]$ est $\approx 0{,}68$ et la probabilité d'observer une valeur de $S_{100} \in [50 - 3SD,\, 50 + 3SD]$ est $\approx 0{,}997$. Si l'on obtient le résultat $S_{100} \geq 67$, la probabilité de ce résultat est très petite puisque $67 \notin [35,\, 65] = [50 - 3SD,\, 50 + 3SD]$. Si l'on a fait l'expérience qu'une seule fois, il est raisonnable d'avoir des doutes sur le fait que la pièce est équilibrée, même si ce résultat est possible, car la probabilité $P(S_{100} > 65) \approx 0{,}0015$. *Mais* si l'on a fait l'expérience plus de 10^6 fois et qu'on rapporte uniquement un des résultats peu probables, il n'y a pas de raison de douter que la pièce est équilibrée. Par la loi des grands nombres on s'attend à ce qu'on trouve des résultats avec $S_{100} = 67$, mais avec une fréquence relative très petite. On a ici une illustration du *principe de la loterie :* la probabilité que « je gagne » est très différente de la probabilité qu'« il y a un gagnant ». Cette dernière probabilité est grande si le nombre N de participants est grand ; si N augmente cette probabilité augmente. Lorsque N est très grand il est surprenant de ne pas avoir un gagnant. □

12.3 Théorème de la limite centrale

Le théorème 12.2 est appelé en anglais « *Central Limit Theorem* ». La terminologie est due à Pólya *à cause du rôle central de ce théorème* en théorie des probabilités et en statistique. La terminologie usuelle en français « théorème de la limite centrale » (TLC) indique que ce théorème décrit le comportement du centre de la distribution par opposition à la queue de la distribution [1].

On considère des v.a. réelles et indépendantes $X_1, \ldots, X_n$. Chaque v.a. X_j a une espérance μ_j et une variance σ_j^2 telle que $0 < \sigma_j^2 < \infty$. On se ramène au cas $\mu_j = 0$ en introduisant de nouvelles v.a. centrées,

$$Y_j = X_j - \mu_j\,.$$

1. Les noms de mathématiciens célèbres sont attachés à ce théorème, Laplace, Poisson, Chebyshev, Liapunov (1859-1918), Markov, Lindeberg (1876-1932), Bernstein, Lévy (1886-1971), Cramér (1893-1985), Feller (1906-1970), Kolmogorov.

La première partie du théorème 12.2 décrit le comportement d'une somme de v.a. i.i.d. sur l'échelle $a(n) = \sqrt{n}$ lorsque $n \to \infty$. Le caractère essentiel du théorème 12.2 est celui d'un *théorème d'approximation* ($2^{\text{ième}}$ partie). On s'intéresse à l'effet cumulé d'un grand nombre de v.a. $\widehat{Y}_j$, $1 \leq j \leq n$, (causes aléatoires) indépendantes et « petites » dans le sens suivant : $\mathbb{E}(\widehat{Y}_j) = 0$ et la variance de chaque v.a. $\widehat{Y}_j$ vaut n^{-1}.

Il y a un cas pour lequel l'affirmation du théorème est exacte pour n fini. Soit X_j, $j = 1, \ldots, n$, des v.a. i.i.d. et $X_j \sim N(m, \sigma^2)$. La v.a.

$$\widehat{Y}_j = \frac{X_j - m}{\sigma\sqrt{n}} \sim N(0, 1/n)\,.$$

La variance de $\sum_j \widehat{Y}_j$ est la somme des $\mathrm{Var}(\widehat{Y_j})$ et donc égale à 1 et

$$\sum_j \widehat{Y}_j \sim N(0, 1)\,.$$

Par conséquent

$$P\Big(\sum_{j=1}^{n} \widehat{Y}_j \leq t\Big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \mathrm{e}^{-\frac{s^2}{2}}\, ds = \Phi(t)\,.$$

Ce cas particulier est souvent utilisé en statistique ; la justification de ce choix s'appuie sur le théorème 12.2 qui est aussi vrai pour des v.a. non identiquement distribuées (voir section 12.4). Le théorème de la limite centrale concerne le comportement asymptotique de la fonction de répartition de

$$\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}} \quad \text{lorsque} \quad n \to \infty\,.$$

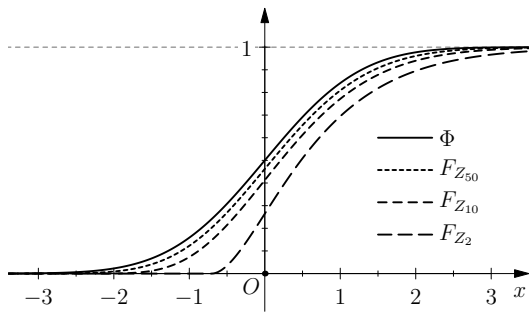


Figure 12.2 – Comparaison des fonctions de répartition pour les v.a. $Z_n = (\sqrt{n})^{-1}(S_n - n)$ où S_n est la somme de n v.a. i.i.d. de loi exponentielle de paramètre $\lambda = 1$. Comparer avec la fonction de répartition de la loi exponentielle de la figure 6.6.

Théorème 12.2 (Théorème central de la théorie des probabilités)
1) *Soit $X_1, X_2, \dots$ des v.a. i.i.d. telles que l'espérance et la variance existent ;* $\mathrm{Var}(X_1) = \sigma^2$ *et* $0 < \sigma^2 < \infty$. *Alors*

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \Big| P\Big(\frac{1}{\sqrt{n\sigma^2}} \sum_{k=1}^{n} \big(X_k - \mathbb{E}(X_k)\big) \le x \Big) - \Phi(x) \Big| = 0 \,.$$

2) *Si de plus* $\mathbb{E}(|X_1 - \mathbb{E}(X_1)|^3) < \infty$, *alors*

$$\sup_{x\in\mathbb{R}} \Big| P\Big(\frac{1}{\sqrt{n\sigma^2}} \sum_{k=1}^{n} \big(X_k - \mathbb{E}(X_k)\big) \le x \Big) - \Phi(x) \Big| \le \frac{\mathbb{E}(|X_1 - \mathbb{E}(X_1)|^3)}{\sigma^3 \sqrt{n}} \,.$$

L'estimée du point 2) est due à Berry et Esseen (1918-2001). Une preuve complète (avec une estimée un peu plus faible) est donnée en section 12.4.

L'hypothèse $0 < \mathrm{Var}(X_1) = \sigma^2 < \infty$ est importante. Soit $X_1, X_2, \dots$ des v.a. i.i.d., telles que $\mathbb{E}(X_1) = 1$ et $\mathrm{Var}(X_1) = 1$. Le théorème 12.2 affirme que

$|S_n - n|$ est de l'ordre de $\sqrt{n}$ avec grande probabilité ;

$P(S_n \ge n) \approx 1/2$ lorsque n devient grand.

Ce résultat, combiné avec le théorème 11.1, signifie qu'asymptotiquement la v.a. S_n *fluctue autour de sa moyenne sur l'échelle* $a(n) = \sqrt{n}$:

a) La fonction de répartition de la loi de $(\sqrt{n})^{-1}(S_n - n)$ converge vers celle de la loi $N(0,1)$ de moyenne nulle ; l'écart de S_n à sa moyenne sur cette échelle est positif ou négatif avec égale probabilité, lorsque $n \to \infty$.

b) Avec probabilité un, cet écart sur cette échelle prend n'importe quelle valeur réelle lorsque n diverge,

$$\limsup_{n\to\infty} \frac{S_n(\omega) - n}{\sqrt{n}} = \infty \quad \text{et} \quad \liminf_{n\to\infty} \frac{S_n(\omega) - n}{\sqrt{n}} = -\infty \,. \tag{12.1}$$

Si $\mathrm{Var}(X_1)$ n'existe pas, on peut avoir un comportement complètement différent. La seule chose que l'on sait est que

$$S_n = n\mathbb{E}(X_1) + r_n \quad \text{et} \quad \lim_{n\to\infty} \frac{|r_n|}{n} = 0 \,.$$

On peut construire (exemple 12.2) des v.a. i.i.d. avec $\mathbb{E}(X_1) = 1$, telles que $|S_n - n|$ est au moins de l'ordre de $n/\ln n$, et $S_n - n$ a le même signe avec probabilité tendant vers un si $n \to \infty$:

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} P\Big(S_n < n - \frac{(1-\varepsilon)n}{\ln n} \Big) = 1 \,.$$

Dans cet exemple, asymptotiquement lorsque $n \to \infty$, la valeur de S_n *ne fluctue pas autour de* $n = \mathbb{E}(S_n)$, quelle que soit l'échelle $a(n)$ que l'on considère.

Exemple 12.2 Soit X une v.a. discrète qui prend les valeurs,

$$P(X = \mathrm{e}^k) := \frac{\mathrm{e}^{-k}}{k(k+1)} \,,\ k \ge 1 \quad \text{et} \quad P(X = 0) := 1 - \sum_{k\ge 1} \frac{\mathrm{e}^{-k}}{k(k+1)} \,.$$

L'espérance de X est égale à

$$\mathbb{E}(X) = \sum_{k\geq 1} \mathrm{e}^k \frac{\mathrm{e}^{-k}}{k(k+1)} = \sum_{k\geq 1} \Big(\frac{1}{k} - \frac{1}{k+1}\Big) = 1\,.$$

Soit $X_1, X_2, \ldots$ une suite de v.a. i.i.d., $X_i \overset{\mathcal{L}}{=} X$, et $S_n = X_1 + \cdots + X_n$. Pour étudier S_n on utilise la méthode du lemme 10.1 en introduisant des v.a. tronquées

$$X_i' := X_i\, I_{\left\{X_i \leq \frac{n}{\ln n}\right\}}\,.$$

D'une part

$$\begin{aligned} P(X_i = X_i',\ i = 1, \ldots, n) &= 1 - P(\exists\, j\colon X_j \neq X_j') \\ &\geq 1 - \sum_{j=1}^{n} P(X_j \neq X_j') \\ &= 1 - nP\Big(X_1 > \frac{n}{\ln n}\Big)\,; \end{aligned}$$

d'autre part, en utilisant la proposition I.3 et en posant

$$k^* := \max\Big\{k\colon \mathrm{e}^k \leq \frac{n}{\ln n}\Big\}\,,$$

$$\begin{aligned} P\Big(X_1 > \frac{n}{\ln n}\Big) &= \sum_{k>k^*} \frac{\mathrm{e}^{-k}}{k(k+1)} \leq \int_{k^*}^{\infty} \frac{\mathrm{e}^{-x}}{x^2}\, dx \\ &= \frac{\mathrm{e}^{-k^*}}{(k^*)^2} - \int_{k^*}^{\infty} \frac{2\mathrm{e}^{-x}}{x^3}\, dx \\ &\leq \frac{\mathrm{e}^{-k^*}}{(k^*)^2} \approx \frac{\ln n}{n}\Big(\ln \frac{n}{\ln n}\Big)^{-2}\,. \end{aligned}$$

Par conséquent

$$\lim_{n\to\infty} P(X_i = X_i',\ i = 1, \ldots, n) = 1\,. \tag{12.2}$$

On calcule

$$\mathbb{E}(X_1') = \sum_{k=1}^{k^*} \Big(\frac{1}{k} - \frac{1}{k+1}\Big) \approx 1 - \Big(\ln \frac{n}{\ln n}\Big)^{-1} \approx 1 - \frac{1}{\ln n}$$

et

$$\mathrm{Var}(X_1') \leq \mathbb{E}((X_1')^2) \leq \sum_{k=1}^{k^*} \frac{\mathrm{e}^k}{k^2} \approx \int_1^{k^*} \frac{\mathrm{e}^x}{x^2}\, dx\,.$$

L'intégrale est estimée en utilisant

$$\int_3^y \frac{\mathrm{e}^x}{x^2}\, dx \leq \frac{\mathrm{e}^y}{y^2} + 2\int_3^y \frac{\mathrm{e}^x}{x^3}\, dx \leq \frac{\mathrm{e}^y}{y^2} + 2\int_3^y \frac{\mathrm{e}^x}{3x^2}\, dx\,,$$

ce qui donne

$$\frac{1}{3}\int_3^y \frac{\mathrm{e}^x}{x^2}\,dx \le \frac{\mathrm{e}^y}{y^2}\,.$$

Par l'inégalité de Chebyshev et les résultats précédents,

$$P\Big(\sum_{i=1}^n \big|X_i' - \mathbb{E}(X_i')\big| \ge \frac{\varepsilon n}{\ln n}\Big) \le \frac{n\mathrm{Var}(X_1')}{\varepsilon^2 n^2}(\ln n)^2 \overset{n\to 0}{\longrightarrow} 0\,. \tag{12.3}$$

A partir de (12.2), (12.3) et du lemme 10.1,

$$P\Big(\big|S_n - n\mathbb{E}(X_1')\big| \ge \frac{\varepsilon n}{\ln n}\Big) \le P\Big(\big|S_n' - n\mathbb{E}(X_1')\big| \ge \frac{\varepsilon n}{\ln n}\Big) + P(S_n \ne S_n')$$

converge vers 0, lorsque $n \to \infty$. Par conséquent,

$$\forall\,\varepsilon > 0 \quad \lim_{n\to\infty} P\Big(-\frac{(1+\varepsilon)n}{\ln n} \le S_n - n \le -\frac{(1-\varepsilon)n}{\ln n}\Big) = 1\,.$$

La LGN est bien sûr vérifiée puisque $\lim_n P(|S_n - n| \ge \varepsilon n) = 0$. □

Exemple 12.3 Si λ est grand ($\lambda \in \mathbb{N}$ pour simplifier), la loi de Poisson π_λ est la loi d'une somme de v.a. i.i.d. de loi de Poisson π_1. Si $Z \sim \pi_\lambda$, la fonction de répartition G_λ de la v.a. $Z_\lambda := (Z-\lambda)/\sqrt{\lambda}$ vérifie

$$\sup_{x\in\mathbb{R}} \big|G_\lambda(x) - \Phi(x)\big| \le O\Big(\frac{1}{\sqrt{\lambda}}\Big)\,.$$

A partir de ce résultat on peut montrer que

$$\lim_{n\to\infty} \frac{\sqrt{2\pi n}\mathrm{e}^{-n}n^n}{n!} = 1\,.$$

Soit $X_1, \ldots, X_n$ des v.a. i.i.d. de loi de Poisson π_1, et $S_n := X_1 \cdots + X_n$. Par le TLC

$$P\Big(-1 \le \frac{S_n - n}{\sqrt{n}} \le 0\Big) \overset{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}}\int_{-1}^0 \mathrm{e}^{-\frac{t^2}{2}}\,dt\,.$$

D'autre part $S_n \sim \pi_n$, ce qui permet d'écrire ($\ell \in \mathbb{N}$)

$$\begin{aligned} P\Big(-1 \le \frac{S_n - n}{\sqrt{n}} \le 0\Big) &= \sum_{0\le\ell\le\sqrt{n}} \mathrm{e}^{-n}\frac{n^{n-\ell}}{(n-l)!} \\ &= \mathrm{e}^{-n}\frac{n^n}{n!}\sum_{0\le\ell\le\sqrt{n}} \frac{n(n-1)\cdots(n-\ell+1)}{n^\ell} \\ &= \mathrm{e}^{-n}\frac{n^n}{n!}\sum_{0\le\ell\le\sqrt{n}} \prod_{j=1}^{\ell-1}\Big(1-\frac{j}{n}\Big)\,. \end{aligned}$$

Pour $0 \le \ell \le \sqrt{n}$, on a les inégalités (voir solution exercice 3.10)

$$\mathrm{e}^{-\frac{\ell^2}{2n}}\mathrm{e}^{-\frac{1}{3\sqrt{n}}} \le \prod_{j=1}^{\ell-1}\Big(1-\frac{j}{n}\Big) \le \mathrm{e}^{-\frac{\ell^2}{2n}}\mathrm{e}^{\frac{1}{2\sqrt{n}}}\,.$$

Par conséquent,

$$\begin{aligned}\lim_{n\to\infty} P\Big(-1 \le \frac{S_n - n}{\sqrt{n}} \le 0\Big) &= \frac{1}{\sqrt{2\pi}} \int_{-1}^{0} \mathrm{e}^{-\frac{t^2}{2}}\, dt \\ &= \lim_{n\to\infty} \frac{\sqrt{n}\mathrm{e}^{-n} n^n}{n!}\, \frac{1}{\sqrt{n}} \sum_{0\le \ell \le \sqrt{n}} \prod_{j=1}^{\ell-1} \Big(1 - \frac{j}{n}\Big) \\ &= \lim_{n\to\infty} \frac{\sqrt{n}\mathrm{e}^{-n} n^n}{n!} \int_{-1}^{0} \mathrm{e}^{-\frac{t^2}{2}}\, dt\,.\end{aligned}$$

□

12.4 Preuve du théorème de la limite centrale

On démontre le théorème de la limite centrale avec une estimée (12.6) plus faible que celle de Berry-Esseen, ainsi que la première partie du théorème 12.2. Cette preuve est due à Lindenberg (1922). Il y a d'autres preuves de ce théorème qui sont plus courtes, mais qui utilisent des outils de l'analyse harmonique. La preuve ci-dessous a le mérite d'être directe et elle ne fait appel qu'aux notions introduites dans ce livre.

Soit $X_1, X_2, \ldots$ des v.a. indépendantes,

$$\mathbb{E}(X_j) \equiv \mu_j \quad \text{et} \quad 0 < \mathrm{Var} X_j \equiv \sigma_j^2 < \infty\,.$$

On ne suppose pas que les v.a. sont identiquement distribuées. On pose

$$s_n^2 := \sigma_1^2 + \cdots + \sigma_n^2\,.$$

On introduit les v.a.

$$Y_j := \frac{X_j - \mu_j}{s_n} \quad \text{et} \quad S_n := \sum_{j=1}^{n} Y_j$$

telles que

$$\mathrm{Var} Y_j = \frac{\sigma_j^2}{s_n^2} \quad \text{et} \quad \mathbb{E}(Y_j) = 0\,.$$

On introduit aussi n v.a. indépendantes

$$Z_j \sim N(0, \sigma_j^2/s_n^2)$$

qui sont aussi indépendantes des X_k et on pose

$$T_n = Z_1 + \cdots + Z_n\,.$$

Par définition, la loi de T_n est $N(0,1)$ et

$$\mathrm{Var}(Y_j) = \mathrm{Var}(Z_j) \quad \text{et} \quad \mathrm{Var}(S_n) = \mathrm{Var}(T_n) = 1\,.$$

On estime supérieurement la différence

$$P(S_n \le t) - P(T_n \le t) = \mathbb{E}\big(I_{(-\infty,t]}(S_n)\big) - \mathbb{E}\big(I_{(-\infty,t]}(T_n)\big)\,.$$

Soit $h : \mathbb{R} \to [0,1]$ une fonction monotone décroissante de classe C^3 ayant les propriétés suivantes :

$$h(s) = 1 \quad \text{si} \quad s \le 0 \quad \text{et} \quad h(s) = 0 \quad \text{si} \quad s \ge 1\,.$$

Soit $0 < \delta \le 1$ et $t \in \mathbb{R}$ fixés ; on définit $f_\delta : \mathbb{R} \to [0,1]$,

$$f_\delta(x) := h\big(\delta^{-1}(x-t)\big)\,.$$

Par définition de h, $I_{(-\infty,t]}(x) \le f_\delta(x)$ pour tout $x \in \mathbb{R}$ et par conséquent

$$\begin{aligned} P(S_n \le t) - P(T_n \le t) \le{}& \big(\mathbb{E}\big(f_\delta(S_n)\big) - \mathbb{E}\big(f_\delta(T_n)\big)\big) \\ &+ \big(\mathbb{E}\big(f_\delta(T_n)\big) - \mathbb{E}\big(I_{(-\infty,t]}(T_n)\big)\big)\,. \end{aligned}$$

Le $2^{\text{ième}}$ terme de droite est facilement contrôlé,

$$0 \le \mathbb{E}\big(f_\delta(T_n)\big) - \mathbb{E}\big(I_{(-\infty,t]}(T_n)\big) = \frac{1}{\sqrt{2\pi}} \int_t^{t+\delta} h\big(\delta^{-1}(x-t)\big) \mathrm{e}^{-\frac{x^2}{2}}\, dx \le \delta\,.$$

La différence $\mathbb{E}\big(f_\delta(S_n)\big) - \mathbb{E}\big(f_\delta(T_n)\big)$ est exprimée par une somme télescopique en introduisant les v.a.

$$U_k := \sum_{1 \le j < k} Z_j + \sum_{k < j \le n} Y_j\,.$$

Avec ces v.a.,

$$\mathbb{E}\big[f_\delta\big(S_n\big)\big] - \mathbb{E}\big[f_\delta\big(T_n\big)\big] = \sum_{k=1}^{n} \Big\{\mathbb{E}\big[f_\delta\big(U_k + Y_k\big)\big] - \mathbb{E}\big[f_\delta\big(U_k + Z_k\big)\big]\Big\}\,.$$

On estime chaque terme de la somme ci-dessus. Comme f_δ est de classe C^3, on peut écrire un développement de Taylor au deuxième ordre avec reste ; on développe $f_\delta(U_k(\omega) + Y_k(\omega))$ pour ω fixé.

$$\begin{aligned} f_\delta\big(U_k + Y_k\big) = {}& f_\delta\big(U_k\big) + Y_k\, f_\delta^{(1)}\big(U_k\big) + \big(Y_k^2/2\big)\, f_\delta^{(2)}\big(U_k\big) \\ &+ \big(Y_k^2/2\big)\Big[f_\delta^{(2)}\big(U_k^*\big) - f_\delta^{(2)}\big(U_k\big)\Big]\,, \end{aligned} \tag{12.4}$$

avec $U_k^* = U_k + W_k$ et $W_k \le |Y_k|$; on estime le reste en notant que

$$A := \sup_s |h^{(3)}(s)| \implies |f_\delta^{(3)}(x)| \le A\delta^{-3}$$

et

$$|\, f_\delta^{(2)}\big(U_k^*\big) - f_\delta^{(2)}\big(U_k\big)| = \Big| \int_{U_k}^{U_k + W_k} f_\delta^{(3)}(x)\, dx \Big| \le A\delta^{-3}|Y_k|\,. \tag{12.5}$$

On a une expression similaire pour $f_\delta(U_k + Z_k)$.

On peut maintenant estimer $\mathbb{E}\big[f_\delta(S_n)\big] - \mathbb{E}\big[f_\delta(T_n)\big]$ en tenant compte de (12.4), de l'expression similaire pour $f_\delta(U_k+Z_k)$, de l'estimation du reste (12.5) et des identités $\mathbb{E}(Y_k) = \mathbb{E}(Z_k) = 0$ et $\mathrm{Var}(Y_k) = \mathrm{Var}(Z_k)$. Les v.a. U_k, Y_k et Z_k sont indépendantes et pour $p = 1, 2$,

$$\mathbb{E}\big((Y_k)^p\, f_\delta^{(p)}(U_k)\big) = \mathbb{E}\big((Y_k)^p\big)\mathbb{E}\big(f_\delta^{(p)}(U_k)\big)$$

et

$$\mathbb{E}\big((Z_k)^p\, f_\delta^{(p)}(U_k)\big) = \mathbb{E}\big((Z_k)^p\big)\mathbb{E}\big(f_\delta^{(p)}(U_k)\big)\,.$$

Le résultat final est

$$\mathbb{E}\big(f_\delta(S_n)\big) - \mathbb{E}\big(f_\delta(T_n)\big) \le \frac{A\delta^{-3}}{2} \sum_{j=1}^{n} \big\{\mathbb{E}(|Y_j|^3) + \mathbb{E}(|Z_j|^3)\big\}\,.$$

Si $V \sim N(0, \tau^2)$, on obtient par intégration par parties

$$\mathbb{E}(|V|^3) = \frac{2}{\sqrt{2\pi\tau^2}} \int_0^\infty t^3\, \mathrm{e}^{-\frac{t^2}{2\tau^2}}\, dt = \frac{4}{\sqrt{2\pi}}\tau^3\,.$$

Par conséquent

$$\begin{aligned}\sum_{j=1}^{n} \mathbb{E}(|Z_j|^3) &= \frac{4}{\sqrt{2\pi}s_n^3} \sum_{j=1}^{n} \sigma_j^3 \\ &\le \frac{4}{\sqrt{2\pi}s_n^2} \sum_{j=1}^{n} \sigma_j^2 \Big(\max_k \frac{\sigma_k}{s_n}\Big) \\ &= \frac{4}{\sqrt{2\pi}} \max_k \frac{\sigma_k}{s_n}\,.\end{aligned}$$

On établit une borne inférieure pour $P(S_n \le t) - P(T_n \le t)$ de la même manière, en utilisant, à la place de f_δ, la fonction g_δ telle que pour tout $x \in \mathbb{R}$ $g_\delta(x) \le I_{(-\infty,t]}(x)$:

$$g_\delta(x) := h\big(\delta^{-1}(x - t + \delta)\big)\,.$$

En combinant les deux bornes et choisissant $\delta = s_n^{-1/4}$, on obtient

$$\sup_t \big|P(S_n \le t) - \Phi(t)\big| \le \tag{12.6}$$
$$s_n^{-1/4}\Big(\frac{A}{2}\sum_{j=1}^{n}\frac{\mathbb{E}(|X_j - \mu_j|^3)}{s_n^2} + A\sqrt{\frac{2}{\pi}}\max_j \sigma_j + 1\Big)\,.$$

Lorsque les v.a. sont i.i.d. $s_n^2 = n\sigma^2$; l'inégalité (12.6) devient

$$\sup_t \big|P(S_n \le t) - \Phi(t)\big| \le C\, n^{-\frac{1}{8}}\,.$$

Ceci démontre la deuxième partie du théorème 12.2 avec une estimée plus faible que celle Berry-Esseen.

On peut estimer (12.5) en supposant seulement l'existence des variances. Soit $\varepsilon > 0$.

$$f_\delta^{(2)}\big(U_k^*\big) - f_\delta^{(2)}\big(U_k\big) = \big[f_\delta^{(2)}\big(U_k^*\big) - f_\delta^{(2)}\big(U_k\big)\big]\,\big[I_{\{|Y_k|\le\varepsilon\}} + I_{\{|Y_k|>\varepsilon\}}\big]\,.$$

Le premier terme est estimé par

$$\begin{aligned}\big|\, f_\delta^{(2)}\big(U_k^*\big) - f_\delta^{(2)}\big(U_k\big)\big| I_{\{|Y_k|\le\varepsilon\}} &\le \Big|\int_{U_k}^{U_k^*} f_\delta^{(3)}(x)\,dx\Big| I_{\{|Y_k|\le\varepsilon\}} \\ &\le A\delta^{-3}|Y_k| I_{\{|Y_k|\le\varepsilon\}} \le A\delta^{-3}\varepsilon\,;\end{aligned}$$

le deuxième terme par

$$\big|\, f_\delta^{(2)}\big(U_k^*\big) - f_\delta^{(2)}\big(U_k\big)\big|\, I_{\{|Y_k|>\varepsilon\}} \le 2\delta^{-2} B\, I_{\{|Y_k|\ge\varepsilon\}}$$

avec

$$B := \sup_s |h^{(2)}(s)|\,.$$

Comme $\mathrm{Var} S_n = 1$, le membre de droite de (12.6) peut être remplacé par

$$\begin{aligned}\frac{A\delta^{-3}\varepsilon}{2} + \delta^{-2} B \underbrace{\frac{1}{s_n^2}\sum_{k=1}^n \mathbb{E}\big((X_k-\mu_k)^2 I_{\{|X_k-\mu_k|\ge\varepsilon s_n\}}\big)}_{\equiv L(n;\varepsilon)} \\ + A\delta^{-3}\max_k \frac{\sigma_k}{s_n} + \delta\,.\end{aligned} \tag{12.7}$$

Par définition, la *condition de Lindenberg* est l'affirmation :

$$\forall\varepsilon\colon \quad \lim_{n\to\infty} L(n;\varepsilon) = 0\,.$$

Cette condition implique la *condition de Feller*, qui est l'affirmation

$$\lim_{n\to\infty}\max_k \frac{\sigma_k}{s_n} = 0\,.$$

En effet, pour tout ε

$$\begin{aligned}\frac{\sigma_k^2}{s_n^2} &= \frac{\mathbb{E}(Y_k^2 I_{\{|x|<\varepsilon s_n\}}) + \mathbb{E}(Y_k^2 I_{\{|x|\ge\varepsilon s_n\}})}{s_n^2} \\ &\le \frac{\varepsilon^2 s_n^2}{s_n^2} + \frac{1}{s_n^2}\sum_{j=1}^n \mathbb{E}(Y_j^2 I_{\{|x|\ge\varepsilon s_n\}}) \overset{n\to\infty}{\longrightarrow} \varepsilon^2\,.\end{aligned}$$

Si les X_j sont i.i.d., la condition de Lindenberg est vérifiée, car pour tout $\varepsilon > 0$

$$\lim_n \mathbb{E}\big((X_1-\mu_1)^2 I_{\{|X_1-\mu_1|\ge\varepsilon\sigma\sqrt{n}\}}\big) = 0\,.$$

Ceci démontre la première partie du théorème 12.2. □

Théorème 12.3 *Soit $X_1, X_2, \ldots$ des v.a. indépendantes telles que $\mathbb{E}(X_j) = \mu_j$ et $\mathrm{Var}(X_j) = \sigma_j^2$, $0 < \sigma_j^2 < \infty$. On pose $s_n^2 := \sigma_1^2 + \cdots + \sigma_n^2$. Si la condition de Lindenberg est vérifiée, alors*

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \Big| P\Big(\frac{1}{s_n} \sum_{k=1}^{n} \big(X_k - \mu_k\big) \le x\Big) - \Phi(x)\Big| = 0\,.$$

Preuve Si la condition de Lindenberg est vérifiée, il suffit de prendre dans (12.7) les limites $n \to \infty$, puis $\varepsilon \to 0$ et enfin $\delta \to 0$. □

12.5 Convergence faible

Dans cette dernière section on formalise une notion déjà rencontrée plusieurs fois. Cette notion est utilisée également dans la partie consacrée à la statistique mathématique. La thèse du théorème 12.2 est une affirmation *sur la fonction de répartition* F_n de $(\sigma\sqrt{n})^{-1}(S_n - \mathbb{E}(S_n))$:

$$\lim_{n\to\infty} \sup_{t} |F_n(t) - \Phi(t)| = 0\,.$$

Le résultat (12.1), qui se généralise aux v.a. i.i.d. qui possèdent une variance non nulle, montre que sous ces hypothèses il n'y a pas convergence ponctuelle de la suite

$$\frac{(S_n(\omega) - \mathbb{E}(S_n)}{\sigma\sqrt{n}}\,, \quad n \ge 1\,.$$

Le théorème 12.2 conduit naturellement à introduire la définition 12.1.

Définition 12.1 *Une suite de fonctions de répartition F_n, $n \ge 1$,* converge faiblement *vers la fonction de répartition F si et seulement si pour tout point de continuité t de F on a*

$$\lim_{n\to\infty} F_n(t) = F(t)\,.$$

Une suite de v.a. X_n, $n \ge 1$, converge en loi (ou en distribution) *vers une v.a. X si et seulement si la suite des fonctions de répartition F_{X_n} des X_n converge faiblement vers la fonction de répartition F_X de X. Cette convergence est notée $X_n \xrightarrow{\mathcal{L}} X$.*

Remarque 12.2 La terminologie $X_n \xrightarrow{\mathcal{L}} X$ n'est pas très heureuse (mais très utile !) puisqu'il ne s'agit pas d'une convergence des v.a.. Un *point de continuité* de F est un point $t \in \mathbb{R}$ tel que $F(t-) = F(t)$. Comme le nombre de points de discontinuité est au plus dénombrable, l'ensemble complémentaire des points de continuité de F est dense dans $\mathbb{R}$. □

Exemple 12.4 Les v.a. $Z_{2n}/2n$ de la section 11.2 convergent en loi vers une v.a. Y dont la loi est celle de l'arc-sinus. Dans la section 12.1, $\tilde{S}_n \xrightarrow{\mathcal{L}} X$ où $X \sim N(0, Dt)$. □

Exemple 12.5 On considère la v.a. X de fonction de répartition F, telle que $P(X = a) = 1$, et les v.a. X_n de fonctions de répartition F_n, telles que $P(X_n = a + 1/n) = 1$. Si $t \neq a$, $\lim_n F_n(t) = F(t)$, i.e. X_n converge en loi vers X. Dans cet exemple $F(a) = 1 \neq \lim_n F_n(a) = 0$. □

Proposition 12.1 *La convergence en loi a les propriétés suivantes.*

1) Si a est un point de continuité de F_X,

$$X_n \xrightarrow{\mathcal{L}} X \quad \Longrightarrow \quad \lim_{n\to\infty} P(X_n < a) = P(X < a) = P(X \leq a)\,.$$

2) Soit X une v.a. constante telle que $P(X = a) = 1$. Alors

$$\lim_{n\to\infty} P(|X_n - a| \geq \varepsilon) = 0 \quad \forall\, \varepsilon > 0 \quad \Longleftrightarrow \quad X_n \xrightarrow{\mathcal{L}} X\,.$$

Cette proposition permet d'énoncer la LGN en utilisant la convergence en loi : si les v.a. X_k sont i.i.d. et si l'espérance existe, alors

$$\frac{1}{n}\sum_k X_k \xrightarrow{\mathcal{L}} \mathbb{E}(X_1)\,.$$

Preuve 1) Soit a un point de continuité de F_X ; par définition de la convergence, $\lim_n P(X_n \leq a) = P(X \leq a)$. D'autre part, pour tout $\varepsilon > 0$, il existe $a^* < a$ tel que a^* est un point de continuité de F_X et $F_X(a^*) \geq F_X(a) - \varepsilon$ (voir remarque 12.2). Le résultat suit des inégalités

$$P(X \leq a) - 2\varepsilon \leq P(X_n \leq a^*) \leq P(X_n < a) \leq P(X_n \leq a)$$

qui sont valables pour n grand, et du fait que $\varepsilon > 0$ est arbitraire.

2) On a les inégalités suivantes

$$\begin{aligned} P(|X_n - a| > \varepsilon) &= P(X_n < a - \varepsilon) + P(X_n > a + \varepsilon) \\ &\leq F_{X_n}(a - \varepsilon) + 1 - F_{X_n}(a + \varepsilon) \\ &= P(X_n \leq a - \varepsilon) + P(X_n > a + \varepsilon) \\ &\leq P(|X_n - a| \geq \varepsilon)\,. \end{aligned}$$

Le résultat découle du fait que ces inégalités sont vraies pour tout ε positif et que $1 - F_{X_n}(a + \varepsilon) \geq 0$. □

Si $t < 1$, $\lim_n F_{U_n}(t) = 0$; et si $t > 1$, $\lim_n F_{U_n}(t) = 1$. On en conclut que $U_n \xrightarrow{\mathcal{L}} Y$ avec $P(Y = 1) = 1$.
Si l'on pose $W_n := n(1 - U_n)$,

$$F_{W_n}(s) = P\Big(U_n \geq 1 - \frac{s}{n}\Big) = 1 - F_{U_n}\Big(1 - \frac{s}{n}\Big)\,.$$

Si $s \leq 0$, $F_{W_n}(s) = 0$ pour tout n, et si $0 \leq s \leq n$ on obtient

$$\lim_{n\to\infty} F_{W_n}(s) = 1 - \lim_{n\to\infty} \Big(1 - \frac{s}{n}\Big)^n \to 1 - \mathrm{e}^{-s}\,.$$

W_n converge en loi vers une v.a. exponentielle de paramètre 1. □

Proposition 12.2 *Soit X une v.a. constante telle que $P(X = \mu) = 1$ et X_n, $n \geq 1$, une suite de v.a. telles que $\mathbb{E}(X_n) = \mu_n$ et $\mathrm{Var}(X_n) = \sigma_n^2$. On suppose que*

$$\lim_{n\to\infty} \mu_n = \mu \quad \textit{et} \quad \lim_{n\to\infty} \sigma_n^2 = 0\,.$$

Alors X_n converge en loi vers X.

Preuve En effet, par hypothèse

$$\begin{aligned}\mathbb{E}(|X_n - \mu|^2) &= \mathbb{E}(X_n - \mu_n + \mu_n - \mu)^2 \\ &= \mathbb{E}(X_n - \mu_n)^2 + (\mu_n - \mu)^2 \to 0\,.\end{aligned}$$

On obtient le résultat en utilisant l'inégalité de Markov et la proposition 12.1,

$$P(|X_n - \mu| \geq \varepsilon) \leq \frac{\mathbb{E}(|X_n - \mu|^2)}{\varepsilon^2} \to 0\,.$$

□

Soit $X_n \xrightarrow{\mathcal{L}} X$ avec $P(X = \mu) = 1$; si φ est continue et bornée, alors

$$\lim_{n\to\infty} \mathbb{E}(\varphi(X_n)) = \mathbb{E}(\varphi(X)) = \varphi(\mu)\,.$$

En effet, $|\varphi(x)| \leq C < \infty$ et

$$\forall\, \varepsilon \;\; \exists\, \delta > 0 \;\text{ tel que }\; \big(|x - \mu| \leq \delta \implies |\varphi(x) - \varphi(\mu)| \leq \varepsilon\big)\,.$$

Par conséquent

$$\begin{aligned}\big|\mathbb{E}\big(\varphi(X_n) - \varphi(\mu)\big)\big| &\leq \mathbb{E}\big(|\varphi(X_n) - \varphi(\mu)|\, I_{\{|X_n - \mu| \leq \delta\}}\big) \\ &\quad + \mathbb{E}\big(|\varphi(X_n) - \varphi(\mu)|\, I_{\{|X_n - \mu| > \delta\}}\big) \\ &\leq \varepsilon P(\{|X_n - \mu| \leq \delta\}) + 2C\, P(\{|X_n - \mu| > \delta\}) \\ &\leq \varepsilon + 2C\, P(\{|X_n - \mu| > \delta\}) \overset{n\to\infty}{\longrightarrow} \varepsilon\,.\end{aligned}$$

Comme ε est arbitraire, l'affirmation est vérifiée.

De façon beaucoup plus générale on a le théorème suivant (sans démonstration), qui énonce une affirmation équivalente à la convergence en loi.

Théorème 12.4 *Soit X et X_n, $n \geq 1$, des v.a. réelles. X_n converge en loi vers X si et seulement si pour toute fonction $\varphi : \mathbb{R} \to \mathbb{R}$ continue et bornée,*

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

Un corollaire immédiat de ce théorème est que si $g : \mathbb{R} \to \mathbb{R}$ est une fonction continue alors

$$X_n \xrightarrow{\mathcal{L}} X \implies g(X_n) \xrightarrow{\mathcal{L}} g(X) .$$

En effet, pour toute fonction $\varphi : \mathbb{R} \to \mathbb{R}$ continue et bornée, la fonction $\varphi \circ g : \mathbb{R} \to \mathbb{R}$ est aussi continue et bornée.

Résumé Dans les chapitres 10 et 12 on a établit deux résultats fondamentaux de la théorie des probabilités pour des v.a. indépendantes et identiquement distribuées.

1) La LGN : $\dfrac{\sum_{n=1}^{N} X_n}{N} \xrightarrow{\mathcal{L}} \mathbb{E}(X)$ si $\mathbb{E}(|X|)$ existe.

2) Le TLC : $\dfrac{\sum_{n=1}^{N} \left[X_n - \mathbb{E}(X_n)\right]}{\sqrt{N}} \xrightarrow{\mathcal{L}} Y$ avec $Y \sim N(0,1)$ si $\mathrm{Var} X_n = 1$.

Un point *essentiel* dans ce genre de résultats de type asymptotique est l'étude de la vitesse avec laquelle les limites sont obtenues, car on veut utiliser ces résultats comme *théorèmes d'approximation* (n fini). L'inégalité de Hoeffding donne une telle information pour la LGN. Lorsque ces théorèmes s'appliquent et que la vitesse de convergence est rapide, on a le cas du *hasard bénin* selon la terminologie de B. Mandelbrot (1924-2010). Cette forme du hasard est celle qu'on maîtrise bien. On a aussi donné un sens précis au fait que si des v.a. X_i sont i.i.d. et possédent une espérance et une variance σ non nulle, alors la somme S_n de ces v.a. fluctue autour de $\mathbb{E}(S_n)$ sur l'échelle $\sigma\sqrt{n}$, et le comportement de ces fluctuations *pour cette classe de v.a. est universel*. Par contre, si les variances des X_i ou les espérances des X_i n'existent pas, le comportement de S_n peut être très différent.

12.6 Exercices

Exercice 12.1 Calculer la probabilité de l'événement $P(Y \geq 90)$ si Y est la somme de 100 v.a. de Bernoulli i.i.d. $X_1, \ldots, X_n$, $P(X_i = 1) = 0{,}8$. Estimer cette probabilité en utilisant la deuxième partie du théorème 12.2. Calculer l'erreur relative de cette estimation.

Exercice 12.2 On lance 10^6 fois, de façon indépendante, une pièce de monnaie équilibrée. On note N^+ le nombre de Piles obtenus. Classer par ordre décroissant de probabilité les événements suivants.

1) $E_1 := \{N^+ \geq 6 \cdot 10^5\}$.
2) $E_2 := \{N^+ \geq 500500\}$.

3) $E_3 := \{N^+ = 5 \cdot 10^5\}$.

Justifier votre réponse en donnant des estimations des probabilités de ces événements.

Exercice 12.3 On considère une v.a. X de Poisson de paramètre $\lambda = 100$.
a) Estimer la probabilité $P(X > 120)$.
b) Quelle est l'estimation de l'erreur donnée par l'inégalité de Berry-Esseen ?
Indication : la somme de v.a. indépendantes de Poisson est une v.a. de Poisson.

Exercice 12.4 On doit déterminer la somme $\sum_{n=1}^{N} a_n$ où chaque $a_n \in [1, 10]$. Les a_n sont connus avec une précision ε,

$$a_n = b_n + \varepsilon_n \quad \text{avec} \quad -\varepsilon \leq \varepsilon_n \leq \varepsilon\,.$$

On calcule $\sum_{n=1}^{N} b_n$ à la place de $\sum_{n=1}^{N} a_n$.
a) Quelle est l'erreur maximale que l'on peut faire ?
b) Quelle est l'erreur à laquelle on s'attend si les erreurs ε_n sont indépendantes et uniformément distribuées sur l'intervalle $[-\varepsilon, \varepsilon]$?
c) Quelle est la probabilité de faire une erreur plus grande que $\sqrt{N}\varepsilon$?

Exercice 12.5 On suppose que les v.a. X_k, $k \geq 1$, sont i.i.d. avec moyenne μ et variance σ^2. Soit Z_n des v.a. telles que $Z_n \xrightarrow{\mathcal{L}} Y$, où Y est une v.a. constante, $P(Y = a) = 1$. Sous ces conditions montrer que

$$\lim_{n\to\infty} P\Big(\frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} \big(X_k - \mu\big) + Z_n \leq x\Big) = \Phi(x + a) \quad \forall\, x \in \mathbb{R}\,.$$

Indication : utiliser la proposition 12.1.

Chapitre 13

Estimation ponctuelle

Dans les quatre derniers chapitres on applique la théorie développée jusqu'ici à des thèmes classiques de la statistique mathématique. Les ouvrages [Mo], [Ar] et [Pe] couvrent une matière beaucoup plus vaste avec beaucoup plus de profondeur.

La statistique s'applique à un ensemble très vaste de domaines allant des sciences de base aux sciences humaines. Elle est fondamentale pour toutes les branches qui se basent sur des observations. L'exploration, la modélisation et l'analyse des données, ainsi que la manière d'obtenir celles-ci, sont des problématiques importantes de la statistique, qui ne sont pas abordées dans ce livre.

La statistique a des liens très forts avec la théorie des probabilités dont elle utilise les concepts et les résultats, comme la loi des grands nombres, le théorème de la limite centrale etc. Une question de base est celle de l'analyse de résultats obtenus en général sous la forme de valeurs $\mathbf{x} \in \mathbb{R}^n$ de v.a. $X_1, \ldots, X_n$. Une différence essentielle, par rapport aux chapitres précédents, est que le mécanisme de l'expérience aléatoire dont proviennent les résultats $\mathbf{x}$ n'est pas complètement connu. La modélisation de l'expérience aléatoire est incomplète en ce sens que la mesure de probabilité fait partie d'une famille de lois indexées par un paramètre θ. Dans les chapitres 14, 15 et 16 on suppose que la loi décrivant l'expérience est celle correspondant à une valeur déterminée θ_* du paramètre θ, mais qui est inconnue. Le mécanisme aléatoire de l'expérience est donc fixé, mais on ne le connaît pas entièrement. Le problème de l'*inférence* est d'estimer la valeur θ_* à partir des données expérimentales.

13.1 Modèle statistique

Pour formaliser cette situation on définit un *modèle statistique* comme un quadruplet $(\Omega, \mathcal{F}, P_\theta, \Theta)$ où Θ est un ensemble, qui est le domaine des valeurs possibles du paramètre θ ; pour tout $\theta \in \Theta$, $(\Omega, \mathcal{F}, P_\theta)$ est un espace de probabilité. La famille des lois de probabilité $\{P_\theta \colon \theta \in \Theta\}$ est donnée. Dans ce chapitre, après avoir donné des exemples de modèles statistiques, on introduit les notions de base dont on a besoin par la suite. Enfin, on étudie en détail un modèle statistique très important dans la pratique, le modèle de Gauss.

Exemple 13.1 On reprend l'exemple 4.4 du canal de transmission de la section 4.1. On connaît le résultat de l'expérience qui est la lettre b_k transmise par le canal, mais non la lettre envoyée. Ici la lettre envoyée joue le rôle du paramètre θ et donc $\Theta = \mathbb{A}$, l'alphabet d'entrée. Pour chaque valeur a_j de θ on a une mesure de probabilité qui est une ligne de la matrice stochastique $\mathbf{M} : \mathbf{M}_{jk} = P(b_k|a_j)$ et

$$P(b_k|a_j) = P(\text{la lettre } b_k \text{ est reçue} \,|\, \text{la lettre } a_j \text{ est envoyée}) .$$

Une manière d'estimer la valeur de θ est de construire une fonction de décision φ, telle que $\varphi(k)$ est un maximum global de la fonction $j \mapsto P(b_k|a_j)$:

$$P(b_k|a_{\varphi(k)}) \geq P(b_k|a_i) \quad \forall\, i\,. \tag{13.1}$$

La fonction de décision φ est un exemple d'estimateur ponctuel, plus précisément ici un estimateur de maximum de vraisemblance à cause des inégalités (13.1). Un estimateur est une v.a. car c'est une fonction du résultat d'une expérience aléatoire ; la valeur de cette v.a. pour un résultat de l'expérience donne la valeur estimée du paramètre.

Dans l'exemple 4.4, on a aussi considéré le cas où l'on a une information sur les lettres envoyées dans le canal sous la forme d'une mesure de probabilité sur $\mathbb{A}$; $P(A_j)$ donne la probabilité que la lettre a_j soit envoyée dans le canal (mesure de probabilité a priori). On construit une fonction de décision, qui tient compte de cette information, en utilisant la formule de Bayes pour calculer la mesure de probabilité a posteriori. Cette approche importante de la statistique, appelée *statistique bayésienne*, qui utilise une information a priori sur le paramètre θ, sous la forme d'une mesure de probabilité définie sur l'espace des paramètres Θ, n'est pas considérée dans ce livre. Dans cet exemple les fonctions de décision sont les mêmes dans les deux cas, si la mesure de probabilité $P(A_j)$ est uniforme. □

Exemple 13.2 On considère la mesure d'une grandeur scalaire m_* (exemple 2.2 section 2.1). Le point de vue adopté est que cette grandeur a une valeur objective qu'on veut déterminer à partir de mesures expérimentales. Le désaccord entre les observations est attribué à des incertitudes de nature aléatoire. Les données empiriques sont modélisées par des v.a. $X_1, \ldots, X_n$ dont on *postule* la loi conjointe sans préciser certains paramètres. Un choix standard est de postuler que les v.a. X_i sont indépendantes, $X_i \sim N(m, \sigma^2)$, et que les paramètres m et σ^2 sont inconnus. Dans cet exemple

$$\Theta = \{\theta = (m, \sigma^2) :\ m \in \mathbb{R} \text{ et } \sigma^2 \in \mathbb{R}^+\} .$$

Les quantités

$$\overline{x} := \frac{1}{n}(x_1 + \cdots x_n) \quad \text{et} \quad s^2 := \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2$$

sont des estimations de m et σ^2. Elles sont calculées sur la base des données empiriques $(x_1, \ldots, x_n)$ (valeurs des v.a. $X_1, \ldots, X_n$) obtenues lors des n mesures ;

par conséquent $\overline{x}$ et s^2 sont les valeurs des v.a.

$$\overline{X} := \frac{1}{n}(X_1 + \cdots X_n) \quad \text{et} \quad S_n^2 := \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$$

pour ces données empiriques. □

Exemple 13.3 On considère une urne $\mathcal{U}(N, R)$ contenant N jetons dont R sont rouges et $N-R$ blancs. On tire au hasard n jetons sans remplacement. On indique le nombre r de jetons rouges obtenus. Cette expérience est décrite par l'espace fondamental $\Omega := \{0, 1, \ldots, n\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, et la mesure de probabilité qui est la *loi hypergéométrique*,

$$\mathcal{H}_i(n; N, R)(r) := \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}, \quad r = 0, \ldots, n\,.$$

1) Dans une population dont le nombre total N d'individus est connu on veut estimer la taille R d'une sous-population spécifique à partir d'un échantillon de taille n. Ici $\Theta = \{0, 1, \ldots, N\}$, la liste des valeurs possibles de R ; θ est le paramètre cherché R et $P_\theta = \mathcal{H}_i(n; N, \theta)$. On effectue un tirage de n jetons et on obtient le résultat r. Une estimation naturelle du paramètre R est donnée par la valeur $\hat{\theta}_1(r)$ de la v.a.

$$r \mapsto \hat{\theta}_1(r) := \lfloor \frac{rN}{n} \rfloor\,.$$

(Si n est grand on s'attend à ce que $r/n \simeq R/N$).

2) Estimer le nombre total N d'individus d'une population en connaissant la taille R d'une sous-population spécifique. Cette fois R est connu, mais N est inconnu. $\Theta = \{R, R+1, \ldots\}$ est la liste des valeurs possibles de N ; le paramètre cherché $\theta = N$ et $P_\theta = \mathcal{H}_i(n; \theta, R)$. Une estimation naturelle du paramètre N est donnée par la valeur $\hat{\theta}_2(r)$ de la v.a.

$$r \mapsto \hat{\theta}_2(r) := \lfloor \frac{nR}{r} \rfloor\,.$$

□

13.2 Définitions de base

Un modèle statistique $(\Omega, \mathcal{F}, P_\theta, \Theta)$ est donné ; on écrit $\mathbb{E}_\theta(X)$, ou $\mathrm{Var}_\theta(X)$, si l'espérance de X, ou la variance de X, est calculée avec la mesure de probabilité P_θ du modèle. On suppose que les informations à disposition sont exprimées

sous la forme de n nombres réels $\mathbf{x} = (x_1, \ldots, x_n)$ (il est facile d'adapter les définitions qui suivent à d'autres situations si nécessaire) ; $\mathbf{x}$ est un *échantillon de longueur* n qui est la valeur d'une v.a. vectorielle notée $\mathbf{X} = (X_1, \ldots, X_n)$. L'*espace des échantillons* est l'ensemble de tous les échantillons possibles de l'expérience ; il est noté Σ. La loi de $\mathbf{X}$ (loi conjointe des $X_1, \ldots, X_n$) est une mesure de probabilité sur Σ. On utilise la convention suivante :

1) Si $\mathbf{X}$ est discrète, $f(\mathbf{x}; \theta) := P_\theta(\mathbf{X} = \mathbf{x})$.
2) Si $\mathbf{X}$ est continue, $f(\mathbf{x}; \theta)$ est la densité de probabilité de la loi conjointe P_θ de $\mathbf{X}$.

Exemple 13.4 a) Les v.a. $X_1, \ldots, X_n$ sont i.i.d., $X_i \sim N(m, \sigma^2)$ et $\theta = (m, \sigma^2)$;

$$f(\mathbf{x}; m, \sigma^2) := \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\Big(- \sum_{j=1}^{n} \frac{(x_j - m)^2}{2\sigma^2} \Big) .$$

b) Les v.a. $X_1, \ldots, X_n$ sont i.i.d., $X_i \sim \pi_\lambda$ et $\theta = \lambda$;

$$f(\mathbf{x}; \lambda) = P_\lambda\big(\mathbf{X} = \mathbf{x}\big) = \mathrm{e}^{-n\lambda} \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} .$$

c) Les v.a. $X_1, \ldots, X_n$ sont i.i.d., $P(X_i = 1) = 1 - P(X_i = 0) = p$;

$$f(\mathbf{x}; p) = P_p\big(\mathbf{X} = \mathbf{x}\big) = \prod_i p^{x_i} (1-p)^{1-x_i} = p^{\sum_i x_i} (1-p)^{n - \sum_i x_i} .$$

Le paramètre est $\theta = p$. □

Il y a deux sortes d'estimation d'un paramètre du modèle sur la base d'un échantillon $\mathbf{x}$. Soit l'estimation est donnée par une valeur unique, et dans ce cas on parle d'*estimation ponctuelle*, soit on donne un ensemble de valeurs, souvent un intervalle, et on parle d'*estimation par intervalle*. Ce deuxième type d'estimation est discuté au chapitre 15

Définition 13.1 *On introduit les notions suivantes.*

1) Une statistique *T est une v.a. qui ne dépend que des observations. C'est une v.a. qui est de la forme*

$$T = g(X_1, \ldots, X_n) \equiv g(\mathbf{X})$$

où g est une fonction donnée définie sur $\mathbb{R}^n$.

2) Un estimateur ponctuel *du paramètre θ est une statistique à valeur dans Θ, elle est notée $\hat{\theta}$. La valeur $\hat{\theta}(\mathbf{x})$ est l'*estimation ponctuelle *de θ pour l'échantillon $\mathbf{x} \in \Sigma$.*

3) Le biais *de $\hat{\theta}$ est $b_\theta(\hat{\theta}) := \mathbb{E}_\theta(\hat{\theta}) - \theta$. L'*estimateur *$\hat{\theta}$ est* non biaisé *si et seulement si $b_\theta(\hat{\theta}) = 0$ pour tout θ.*

4) Le carré moyen de l'erreur *est* $\mathrm{CME}_\theta(\hat{\theta}) := \mathbb{E}_\theta\big((\hat{\theta} - \theta)^2\big)$.

Remarque 13.1 On vérifie que

$$\begin{aligned}\mathrm{CME}_\theta(\hat\theta) &= \mathbb{E}_\theta\big((\hat\theta-\theta)^2\big) = \mathbb{E}_\theta\big([\hat\theta-\mathbb{E}_\theta(\hat\theta)+\mathbb{E}_\theta(\hat\theta)-\theta]^2\big)\\ &= \mathbb{E}_\theta\big([\hat\theta-\mathbb{E}_\theta(\hat\theta)]^2\big) + \big(\mathbb{E}_\theta(\hat\theta)-\theta\big)^2\\ &\quad + 2\underbrace{\mathbb{E}_\theta\big(\hat\theta-\mathbb{E}_\theta(\hat\theta)\big)}_{=0}\big(\mathbb{E}_\theta(\hat\theta)-\theta\big)\\ &= \mathbb{E}_\theta\big([\hat\theta-\mathbb{E}_\theta(\hat\theta)]^2\big) + \big(\mathbb{E}_\theta(\hat\theta)-\theta\big)^2\\ &= \mathrm{Var}_\theta(\hat\theta) + b_\theta(\hat\theta)^2\,.\end{aligned}$$

Si $\hat\theta$ est non biaisé, $\mathrm{CME}_\theta(\hat\theta) = \mathrm{Var}_\theta(\hat\theta)$. □

Il n'y a pas qu'une seule façon d'estimer un paramètre. Pour comparer les différents estimateurs d'un même paramètre il faut choisir des critères de comparaison. Par exemple, lorsque $\hat\theta_1$ et $\hat\theta_2$ sont non biaisés, l'estimateur $\hat\theta_1$ est *meilleur* que l'estimateur $\hat\theta_2$ si

$$\mathrm{Var}_\theta(\hat\theta_1) \le \mathrm{Var}_\theta(\hat\theta_2) \quad \forall\,\theta\in\Theta\,.$$

Un estimateur non biaisé est d'autant meilleur que sa variance est petite.

La fonction $(\mathbf{x},\theta)\mapsto f(\mathbf{x};\theta)$ est une fonction de deux variables ; elle est interprétée de deux manières différentes :

1. si θ est fixé, $\mathbf{x}\mapsto f(\mathbf{x};\theta)$ définit une mesure de probabilité sur Σ ;
2. si $\mathbf{x}$ est fixé, la fonction $\theta\mapsto f(\mathbf{x};\theta)$, qui est définie sur Θ, est appelée *fonction de vraisemblance* ; elle est notée ci-dessous par $L_{\mathbf{x}}(\,\cdot\,)$ pour bien la distinguer du premier cas.

Exemple 13.5 Dans le premier cas de l'exemple 13.3 le paramètre est $\theta = R$, et la fonction de vraisemblance est

$$R\mapsto L_r(R) = \mathcal{H}_i(n;N,R)(r)\,.$$

Dans le deuxième cas le paramètre est $\theta = N$, et la fonction de vraisemblance est

$$N\mapsto L_r(N) = \mathcal{H}_i(n;N,R)(r)\,.$$

□

Une classe importante d'estimateurs est celle où $\hat\theta(\mathbf{x})$ est un maximum global de la fonction de vraisemblance $L_{\mathbf{x}}$.

Définition 13.2 *Un estimateur $\hat\theta$ est un* estimateur de maximum de vraisemblance *si et seulement si pour chaque $\mathbf{x}$ la valeur $\hat\theta(\mathbf{x})$ de l'estimateur est un maximum global de $\theta\mapsto L_{\mathbf{x}}(\theta)$:*

$$L_{\mathbf{x}}\big(\hat\theta(\mathbf{x})\big) \ge L_{\mathbf{x}}(\theta) \quad \forall\,\theta\in\Theta\,,\ \forall\,\mathbf{x}\in\mathbb{R}^n\,.$$

Exemple 13.6 L'estimateur de l'exemple 13.1, les estimateurs de l'exemple 13.2 (sous certaines hypothèses, voir section 13.3), ceux de l'exemple 13.3 (voir exercice) sont des estimateurs de maximum de vraisemblance. Dans le cas de l'exemple 13.4 c) l'estimateur de maximum de vraisemblance de p est obtenu en cherchant le maximum global de la fonction $p \mapsto f(\mathbf{x}; p)$, ou de la fonction $p \mapsto \ln f(\mathbf{x}; p)$ puisque le logarithme est une fonction strictement monotone. Par conséquent, pour $\mathbf{x}$ donné, $\hat{p}(\mathbf{x})$ est solution de l'équation

$$\frac{d}{dp}\Big(\big(\sum_i x_i\big) \ln p + \big(n - \sum_i x_i\big) \ln(1-p) \Big) = 0\,.$$

Un calcul simple donne $\hat{p}(\mathbf{x}) = \dfrac{1}{n}\sum_i x_i$. Cet estimateur est non biaisé car $\mathbb{E}_p(\hat{p}) = p$. □

Pour un échantillon $\mathbf{x} = (x_1, \ldots, x_n)$ donné on obtient une estimation du paramètre en calculant $\hat{\theta}(\mathbf{x})$. Cette valeur dépend en général de la longueur n de l'échantillon car l'estimateur dépend de n, i.e. $\hat{\theta} = \hat{\theta}_n$. Une propriété naturelle d'un « bon » estimateur est que l'estimation est d'autant « meilleure » que la longueur de l'échantillon est grande.

Définition 13.3 *On suppose que pour chaque $n \in \mathbb{N}$ on a un estimateur $\hat{\theta}_n$. La suite $\hat{\theta}_n$, $n \geq 1$, des estimateurs est* consistante *si pour tout θ*

$$\lim_{n\to\infty} P_\theta\big(|\hat{\theta}_n - \theta| \geq \varepsilon\big) = 0 \quad \forall\, \varepsilon > 0\,.$$

Remarque 13.2 Sous des conditions assez générales, si les X_i sont i.i.d., les estimateurs de maximum de vraisemblance forment une suite consistante d'estimateurs. □

Proposition 13.1 *Si pour une suite d'estimateurs $\hat{\theta}_n$, $n \geq 1$, $\mathbb{E}_\theta(\hat{\theta}_n) = \theta_n$ et $\mathrm{Var}_\theta\, \hat{\theta}_n = \sigma_n^2$ de sorte que*

$$\lim_n \theta_n = \theta \quad \textit{et} \quad \lim_n \sigma_n^2 = 0\,,$$

alors la suite $\hat{\theta}_n$, $n \geq 1$, est consistante.

Preuve Découle directement des propositions 12.2 et 12.1. □

13.3 Modèle de Gauss

Dans de nombreuses situations concrètes on utilise le *modèle de Gauss*. Dans ce modèle les échantillons $\mathbf{x}$ proviennent de v.a. i.i.d. $X_1, \ldots, X_n$ telles que $X_i \sim N(m, \sigma^2)$. On parle de *populations normalement distribuées*.

On détermine les estimateurs de maximum de vraisemblance pour les paramètres m et σ^2 et on donne les propriétés principales de ces estimateurs. La recherche des estimateurs de maximum de vraisemblance est simplifiée en examinant d'abord deux cas particuliers.

1) On suppose que σ^2 est connu. La fonction de vraisemblance sous cette hypothèse est

$$m \mapsto L_{\mathbf{x}}(m) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\Big(-\sum_{j=1}^{n} \frac{(x_j - m)^2}{2\sigma^2} \Big) .$$

Il faut minimiser $\sum_{j=1}^{n}(x_j - m)^2$. L'estimateur de maximum de vraisemblance est la v.a.

$$\frac{1}{n}\sum_{j=1}^{n} X_j \equiv \overline{X} .$$

Cet estimateur est non biaisé.

2) On suppose que m est connu. La fonction de vraisemblance est

$$\sigma^2 \mapsto L_{\mathbf{x}}(\sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\Big(-\sum_{j=1}^{n} \frac{(x_j - m)^2}{2\sigma^2} \Big) .$$

En prenant le logarithme et en dérivant par rapport à σ^2 on obtient

$$\begin{aligned} \frac{d}{d\sigma^2} \ln L_{\mathbf{x}}(\sigma^2) &= \frac{d}{d\sigma^2}\Big(-\frac{n}{2}\big(\ln(2\pi) + \ln \sigma^2 \big) - \frac{1}{2\sigma^2}\sum_j (x_j - m)^2 \Big) \\ &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_j (x_j - m)^2 . \end{aligned}$$

L'estimateur de maximum de vraisemblance de σ^2 est la v.a.

$$\frac{1}{n}\sum_{j=1}^{n}(X_j - m)^2 .$$

Cet estimateur est aussi non biaisé.

Le cas général, où les deux paramètres m et σ^2 sont inconnus, s'obtient facilement à partir des deux cas précédents en remarquant que l'estimateur du premier cas ne dépend pas de la variance σ^2. Pour trouver le maximum de la fonction de vraisemblance $(m, \sigma^2) \mapsto L_{\mathbf{x}}(m, \sigma^2)$ il suffit de maximiser cette fonction d'abord par rapport à m, puis de maximiser l'expression obtenue par rapport à σ^2. Par conséquent la paire de v.a.

$$\Big(\overline{X}, \frac{1}{n}\sum_j (X_j - \overline{X})^2 \Big) \tag{13.2}$$

est celle des estimateurs de maximum de vraisemblance du modèle. Le premier estimateur $\overline{X}$ est non biaisé puisque

$$\mathbb{E}_{m,\sigma^2}(\overline{X}) = m\,.$$

Ce n'est pas le cas de l'estimateur de la variance. En effet

$$\begin{aligned}\sum_j \mathbb{E}_{m,\sigma^2}\big((X_j-\overline{X})^2\big) &= \sum_j \mathbb{E}_{m,\sigma^2}\big((X_j-m+m-\overline{X})^2\big)\\ &= \sum_j \mathbb{E}_{m,\sigma^2}\big((X_j-m)^2\big) + n\mathbb{E}_{m,\sigma^2}\big((m-\overline{X})^2\big)\\ &\quad - 2\mathbb{E}_{m,\sigma^2}\big((\overline{X}-m)\sum_j (X_j-m)\big)\\ &= \sum_j \mathbb{E}_{m,\sigma^2}\big((X_j-m)^2\big) - n\mathbb{E}_{m,\sigma^2}\big((\overline{X}-m)^2\big)\\ &= n\sigma^2 - n\mathrm{Var}_{m,\sigma^2}\overline{X} = (n-1)\sigma^2\,.\end{aligned}$$

C'est pourquoi on remplace souvent cet estimateur par l'estimateur

$$S_n^2 := \frac{1}{n-1}\sum_{k=1}^{n}\big(X_k-\overline{X}_n\big)^2$$

qui est non biaisé.

Définition 13.4 *La* loi du khi-carré à n degrés de liberté, *notée* χ^2_n, *est la loi de*

$$U_n := Z_1^2 + \cdots + Z_n^2$$

où $Z_1,\ldots,Z_n$ *sont i.i.d.,* $Z_i \sim N(0,1)$.

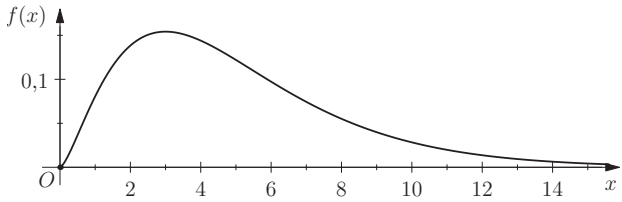


Figure 13.1 – Densité de la loi χ^2_5 du khi-carré à 5 degrés de liberté.

Si la v.a. U_n a une loi du khi-carré, son espérance est la somme des variances des Z_i et sa variance la somme des variances des Z_i^2,

$$\mathbb{E}(U_n) = n\mathbb{E}(Z_1^2) = n \quad \text{et} \quad \mathrm{Var}U_n = n\mathrm{Var}Z_1^2 = 2n\,.$$

En effet, $\mathrm{Var} Z_1^2 = \mathbb{E}(Z_1^4) - \mathbb{E}(Z_1^2)^2 = \mathbb{E}(Z_1^4) - 1$ et

$$\begin{aligned}\mathbb{E}(Z_1^4) &= \frac{1}{\sqrt{2\pi}} \int t^4 \mathrm{e}^{-t^2/2}\, dt \\ &= \frac{1}{\sqrt{2\pi}} \int t^3 \Big(- \frac{d}{dt} \mathrm{e}^{-t^2/2} \Big) dt \\ &= \frac{3}{\sqrt{2\pi}} \int t^2 \mathrm{e}^{-t^2/2}\, dt = 3\,.\end{aligned}$$

Par le TLC on obtient

$$(2n)^{-1/2}(U_n - n) \xrightarrow{\mathcal{L}} N(0,1)\,.$$

Calcul de la densité de Z_1^2.

$$P(Z_1^2 \le t) = P(-\sqrt{t} \le Z_1 \le \sqrt{t}) = F_{Z_1}(\sqrt{t}) - F_{Z_1}(-\sqrt{t})\,.$$

En dérivant par rapport à t on obtient la densité de U_1 : la densité est nulle si $t < 0$, sinon

$$\begin{aligned}\frac{d}{dt} P(Z_1^2 \le t) &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t/2} \Big(\frac{1}{2\sqrt{t}} \Big) + \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t/2} \Big(\frac{1}{2\sqrt{t}} \Big) \\ &= \frac{1}{\sqrt{2\pi t}} \mathrm{e}^{-t/2} \quad t \ge 0\,.\end{aligned}$$

La loi de Z_1^2 est une loi gamma de paramètres $\big(\frac{1}{2}, \frac{1}{2}\big)$. Par conséquent la loi du khi-carré à n degrés de liberté est une loi gamma de paramètres $\big(\frac{n}{2}, \frac{1}{2}\big)$ (proposition 6.4),

$$f_{\chi_n^2}(y) = \Gamma\Big(\frac{n}{2}\Big)^{-1} \frac{1}{2} \mathrm{e}^{-\frac{y}{2}} \Big(\frac{y}{2}\Big)^{\frac{n}{2}-1} I_{\mathbb{R}^+}(y)\,.$$

Proposition 13.2 *Soit $X_1, \ldots, X_n$ i.i.d. de loi $N(m, \sigma^2)$, $\sigma^2 > 0$. Alors les v.a.*

$$\overline{X}_n := \frac{1}{n} \sum_{k=1}^n X_k \quad \textit{et} \quad S_n^2 := \frac{1}{n-1} \sum_{k=1}^n \big(X_k - \overline{X}_n\big)^2$$

sont indépendantes,

$$\overline{X}_n \sim N\big(m, \frac{\sigma^2}{n}\big) \quad \textit{et} \quad \frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2\,.$$

Preuve On introduit les v.a. $Z_i := \sigma^{-1}(X_i - m)$, $i = 1, \ldots, n$, qui sont i.i.d. et de loi $N(0,1)$, ainsi qu'une base orthonormale de $\mathbb{R}^n$ composée des vecteurs $\mathbf{n}_i$ avec $\mathbf{n}_1 := ((\sqrt{n})^{-1}, \ldots, (\sqrt{n})^{-1})$. Soit $\mathbf{U}$ la matrice orthogonale dont les lignes sont les vecteurs $\mathbf{n}_i$. On pose

$$\mathbf{Z} := (Z_1, \ldots, Z_n) \quad \text{et} \quad \mathbf{Y} := \mathbf{U}\mathbf{Z}\,.$$

Chaque v.a. $Y_i \sim N(0,1)$ car $\mathbb{E}(Y_i) = 0$ et

$$\mathbb{E}(Y_i^2) = \sum_j \sum_k \mathbf{n}_i(j)\mathbf{n}_i(k)\mathbb{E}(Z_j Z_k) = \langle \mathbf{n}_i | \mathbf{n}_i \rangle = 1\,.$$

En particulier

$$Y_1 = \frac{\sqrt{n}}{\sigma}\big(\overline{X}_n - m\big) \quad \text{et} \quad \langle \mathbf{n}_1 | \mathbf{Z} \rangle\ \mathbf{n}_1 = \Big(\frac{\overline{X}_n - m}{\sigma}, \dots, \frac{\overline{X}_n - m}{\sigma}\Big)\,.$$

Par définition

$$\begin{aligned}
\mathbf{Z} - \langle \mathbf{n}_1 | \mathbf{Z} \rangle\ \mathbf{n}_1 &= \Big(Z_1 - \frac{\overline{X}_n - m}{\sigma}, \dots, Z_n - \frac{\overline{X}_n - m}{\sigma}\Big) \\
&= \Big(\frac{X_1 - \overline{X}_n}{\sigma}, \dots, \frac{X_n - \overline{X}_n}{\sigma}\Big) \\
&= \sum_{k=2}^n \langle \mathbf{n}_k | \mathbf{Z} \rangle\ \mathbf{n}_k = \sum_{k=2}^n Y_k\, \mathbf{n}_k\,.
\end{aligned}$$

Les v.a. Y_i sont indépendantes car leur loi conjointe est donnée par

$$\begin{aligned}
P(\mathbf{Y} \in A) &= P(\mathbf{Z} \in \mathbf{U}^{-1}A) \\
&= \frac{1}{(\sqrt{2\pi})^n} \int_{\mathbf{U}^{-1}A} \exp\Big(-\frac{z_1^2 + \dots + z_n^2}{2}\Big)\, dz_1 \cdots dz_n \\
&= \frac{1}{(\sqrt{2\pi})^n} \int_A \exp\Big(-\frac{y_1^2 + \dots + y_n^2}{2}\Big)\, dy_1 \cdots dy_n\,,
\end{aligned}$$

puisque $\mathbf{U}$ est orthogonale. On en déduit que

$$\sum_{k=1}^n \Big(\frac{X_k - \overline{X}_n}{\sigma}\Big)^2 = \Big\langle \sum_{i=2}^n Y_i \mathbf{n}_i \Big| \sum_{j=2}^n Y_j \mathbf{n}_j \Big\rangle = \sum_{j=2}^n Y_j^2\,.$$

Par conséquent la v.a. $\dfrac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$ et elle est indépendante de Y_1, et donc aussi de $\overline{X}_n$. □

Définition 13.5 *Soit $X \sim N(0,1)$ et $U_n \sim \chi^2_n$ deux v.a. indépendantes. La* loi de Student à n degrés de liberté *est la loi de la v.a.*

$$\sqrt{n}\frac{X}{\sqrt{U_n}}\,.$$

Cette loi est notée t_n. Student est le pseudonyme de W. Gosset (1876-1937).

La densité de la loi de Student à n degrés de liberté est égale à

$$f_{t_n}(s) = \frac{\Gamma\big(\frac{n+1}{2}\big)}{\Gamma\big(\frac{n}{2}\big)\sqrt{n\pi}}\Big(1 + \frac{s^2}{n}\Big)^{-\frac{n+1}{2}}\,.$$

Si $n = 1$, t_1 est la loi de Cauchy de paramètre 1. La loi t_n est symétrique, et de l'expression de sa densité on déduit facilement que $t_n \xrightarrow{\mathcal{L}} N(0,1)$.

Remarque 13.3 Le calcul de la densité de la loi t_n peut se faire par un calcul direct. On calcule d'abord la loi de $V := \sqrt{U_n}$,

$$F_V(t) = P(U_n \leq t^2) \quad \text{si } t > 0\,, \quad F_V(t) = 0 \quad \text{si } t \leq 0\,,$$

puis celle de X/V,

$$P\Big(\frac{X}{V} \leq t\Big) = P(X \leq tV) = \int_{\{(x,v)\colon x \leq tv\}} f_X(x) f_V(v)\, dxdv\,.$$

On a utilisé l'indépendance des v.a. X et V. Par définition

$$P\Big(\sqrt{n}\frac{X}{\sqrt{U_n}} \leq t\Big) = P\Big(\frac{X}{V} \leq \frac{t}{\sqrt{n}}\Big)\,.$$

En dérivant par rapport à t on obtient la densité de la loi t_n. Détails voir [Pe] pp. 80-81. □

Proposition 13.3 *Si $X_1, \ldots, X_n$ sont i.i.d. et $X_i \sim N(m, \sigma^2)$, alors la loi de*

$$T_n := \sqrt{n}\frac{\overline{X}_n - m}{S_n}$$

est une loi de Student t_{n-1} à $n-1$ degrés de liberté.

Preuve Ce résultat découle de la proposition 13.2. Soit

$$Z := \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \quad \text{et} \quad Y := \frac{(n-1)S_n^2}{\sigma^2}\,.$$

Les v.a. Z et Y sont indépendantes, $Z \sim N(0,1)$, $Y \sim \chi^2_{n-1}$ et

$$\sqrt{n}\frac{\overline{X}_n - m}{S_n} = \sqrt{n}\frac{\overline{X}_n - m}{\sigma}\frac{\sigma}{S_n} = \sqrt{n-1}\frac{Z}{\sqrt{Y}}\,.$$

□

Remarque 13.4 Les résultats ci-dessus sont établis pour le modèle de Gauss. Si les X_i sont i.i.d. et possèdent une espérance et une variance σ^2, mais ne sont pas de loi gaussienne, alors la suite des estimateurs $\overline{X}_n$, $n \geq 1$, est une suite consistante d'estimateurs de l'espérance des X_i (loi des grands nombres). Ce qui compte c'est la loi de $\overline{X}_n$, et le TLC assure que

$$\frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sqrt{\mathrm{Var}\overline{X}_n}} = \sqrt{n}\frac{\overline{X}_n - m}{\sigma} \xrightarrow{\mathcal{L}} Z\,, \quad Z \sim N(0,1)\,.$$

On peut aussi montrer que

$$S_n^2 \xrightarrow{\mathcal{L}} \sigma^2\,,\ S_n \xrightarrow{\mathcal{L}} \sigma\,,\ T_n \xrightarrow{\mathcal{L}} Z \text{ avec } Z \sim N(0,1)\,.$$

Par conséquent, en ce qui concerne les estimateurs $\overline{X}_n$ et T_n on a des résultats similaires à ceux du modèle de Gauss lorsque $n \to \infty$. □

13.4 Exercices

Exercice 13.1 On considère le modèle de l'urne $\mathcal{U}(N, R)$ dans le cas où le paramètre R est connu et le paramètre N est inconnu (exemple 13.3).
a) Donner explicitement la fonction de vraisemblance et déterminer l'estimateur de maximum de vraisemblance.
b) Estimer la taille N de la population totale avec cet estimateur en utilisant les données suivantes : $R = 1000$, la taille de l'échantillon est 1000 et dans cet échantillon il y a $r = 100$ jetons rouges.
Indication : pour trouver l'estimateur de maximum de vraisemblance, examiner le rapport

$$\frac{\mathcal{H}_i(n; N, R)(r)}{\mathcal{H}_i(n; N-1, R)(r)}$$

et montrer que ce rapport est plus grand que 1, puis plus petit que 1 lorsque N croît.

Exercice 13.2 Fréquence d'émission radioactive non connue. On suppose que les émissions ont lieu à des intervalles indépendants et que l'intervalle de temps entre deux émissions est bien modélisé par une v.a. exponentielle de paramètre θ inconnu, $f_\theta(t) = \theta \mathrm{e}^{-\theta t} I_{\mathbb{R}^+}(t)$. On mesure n intervalles de temps successifs ; l'intervalle de temps du $i^{\text{ième}}$ intervalle est donné par la valeur d'une v.a. X_i comme ci-dessus.
a) Déterminer l'estimateur de maximum de vraisemblance de θ.
b) Calculer le biais de cet estimateur.

Exercice 13.3 Soit X_k, $1 \leq k \leq n$, n v.a. i.i.d. d'espérance $\mathbb{E}(X_i) = \mu$ et variance $\mathrm{Var}(X_i) = \sigma^2$. La moyenne empirique des X_i est $\overline{X}_n$ et la variance empirique est S_n^2. Montrer que S_n^2 converge en loi vers σ^2.
Indication : introduire les v.a. $Y_i := X_i - \mu$ et écrire $(n-1)S_n^2$ à l'aide des Y_i. Utiliser la LGN.

Exercice 13.4 On considère n v.a. X_k, $1 \leq k \leq n$, qui sont indépendantes et uniformément distribuées sur l'intervalle $[0, \theta]$, où θ est un paramètre inconnu. Dans cet exercice et le suivant on examine le problème de l'estimation de θ. Pour modéliser ce problème on introduit l'espace de probabilité $\Omega = (\mathbb{R}^+)^n$ et les mesures de probabilités P_θ définies par les densités

$$p_\theta(x_1, \ldots, x_n) = \begin{cases} \theta^{-n} & \text{si } x_i \leq \theta, \forall\ i \leq n \\ 0 & \text{sinon.} \end{cases}$$

L'espace des paramètres $\Theta = (0, \infty)$.

a) Trouver la fonction de vraisemblance $L_{\mathbf{x}}(\theta)$.
b) Montrer que l'estimateur de maximum de vraisemblance $\hat{\theta} = \max_{k=1}^n X_k$.
c) On pose $\hat{\theta}_m := \hat{\theta}$ si l'échantillon $\mathbf{X} = (X_1, \ldots, X_m)$ est de taille m ; montrer que c'est une suite consistante d'estimateurs.

Exercice 13.5 a) Calculer le biais de l'estimateur de maximum de vraisemblance $\hat{\theta}$ de l'exercice 13.4. Comment modifier $\hat{\theta}$ afin d'avoir un estimateur non biaisé ?
b) Construire un estimateur non biaisé de θ à partir de la moyenne empirique $\overline{X}$. Est-ce que cet estimateur est meilleur que l'estimateur non biaisé construit sous a) ?

Chapitre 14

Méthode des moindres carrés

On reprend l'exemple 13.2 de la section 13.1. La grandeur physique scalaire a une valeur objective m_*. Une mesure de cette grandeur donne la valeur x. Celle-ci est interprétée comme la valeur d'une v.a.

$$X = m_* + Z$$

où Z est une v.a. qui modélise les incertitudes (ou erreurs) provenant de la mesure expérimentale de m_*. On distingue les *erreurs systématiques* qui ont invariablement le même signe et même grandeur sous des conditions identiques, et les *erreurs accidentelles* ou *aléatoires* qui n'ont ni le même signe ni la même grandeur sous des conditions identiques [1]. Les erreurs systématiques sont détectées en comparant les résultats obtenus avec un « cas connu » (étalon). Par exemple un mauvais étalonnage d'un instrument de mesure conduit à des erreurs systématiques. Pour interpréter les résultats des mesures on *fait des hypothèses sur la loi de* Z. Dans une mesure il faut distinguer la *précision* et *l'exactitude*. La précision se rapporte au degré de soin apporté lors de la mesure. La variance de Z nous renseigne sur la précision de la mesure : plus cette variance est petite, plus la mesure est précise. Si la précision n'est pas connue à l'avance (ce qui est le cas général), la variance de Z est un paramètre inconnu à déterminer. L'exactitude se rapporte à la justesse des résultats, absence d'erreur systématique. L'absence d'erreur systématique se traduit par $\mathbb{E}(Z) = 0$, i.e. $\mathbb{E}(X) = m_*$. Si les n mesures sont faites dans des conditions *identiques*, on peut supposer que les Z_i sont des v.a. indépendantes et identiquement distribuées. Dans cette formulation de cette expérience de physique il y a donc (au moins) deux paramètres à estimer, $\mathbb{E}(X)$ et $\mathrm{Var}X = \mathrm{Var}Z$. Par la suite on suppose que $\mathbb{E}(Z) = 0$ (absence d'erreur systématique).

Le point de vue adopté ici est que les paramètres inconnus ne sont pas des quantités aléatoires. L'aspect aléatoire, *qui est essentiel pour utiliser les méthodes de la statistique*, intervient uniquement lorsqu'on obtient l'échantillon $\mathbf{x}$.

1. Le terme « erreur », qui est couramment utilisé, peut prêter à confusion. Alors qu'il s'agit typiquement dans le cas d'une erreur systématique d'une mauvaise calibration d'un appareil de mesure, dans le cas d'une erreur accidentelle il s'agit de l'incertitude due au fait que l'on ne contrôle pas complètement les conditions dans lesquelles les mesures sont faites. Il ne s'agit pas dans ce second cas d'une erreur due à une fausse manipulation d'un appareil lors du processus de mesure.

14.1 Principe général

La méthode des moindres carrés est un procédé pour estimer des paramètres. Son intérêt est qu'elle repose sur des hypothèses faibles concernant les lois des v.a. et qu'elle est facile à utiliser à cause de son caractère géométrique. Le principe de la méthode est le suivant. Soit une relation de la forme

$$x = \alpha + \beta \mathrm{e}^t$$

entre deux quantités x et t. A partir de n mesures de x et t, (x_1, t_1), ..., (x_n, t_n), on veut estimer les paramètres α et β qui interviennent de manière linéaire dans la relation considérée. S'il n'y avait aucune incertitude on aurait

$$\forall\, i \quad x_i = \alpha + \beta \mathrm{e}^{t_i}$$

et à partir de deux relations avec $t_1 \neq t_2$ on obtiendrait

$$\beta = \frac{x_1 - x_2}{\mathrm{e}^{t_1} - \mathrm{e}^{t_2}} \quad \text{et} \quad 2\alpha = x_1 + x_2 - \beta\big(\mathrm{e}^{t_1} + \mathrm{e}^{t_2}\big)\,.$$

En présence d'incertitudes (erreurs) on peut seulement estimer les paramètres α et β. Si a et b sont des estimations de α et β, on compare

$$x_i \textit{ valeur observée (mesurée)}$$

et

$$\hat{x}_i := a + b\mathrm{e}^{t_i} \textit{ valeur ajustée} \text{ (pour } a, b \text{ et } t_i).$$

Par définition le *résidu* est la différence entre ces valeurs,

$$r_i := x_i - \hat{x_i}\,.$$

Gauss et Legendre (1752-1833) ont proposé (indépendamment) de prendre comme estimations de α et β les valeurs $\hat{\alpha}$ et $\hat{\beta}$ qui minimisent la somme des carrés des résidus,

$$\sum_i r_i^2 = \|\mathbf{x} - \hat{\mathbf{x}}\|^2\,,$$

où $\|\cdot\|$ est la norme euclidienne de $\mathbb{R}^n$. Dans l'exemple ci-dessus, $\hat{\alpha} = \hat{\alpha}(\mathbf{x}, \mathbf{t})$ et $\hat{\beta} = \hat{\beta}(\mathbf{x}, \mathbf{t})$ sont déterminés par la condition

$$\forall (a, b)\colon \quad \sum_{i=1}^{n} \big[x_i - (a + b\mathrm{e}^{t_i})\big]^2 \geq \sum_{i=1}^{n} \big[x_i - (\hat{\alpha} + \hat{\beta}\mathrm{e}^{t_i})\big]^2\,.$$

Le vecteur $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ est appelé *vecteur d'observation* et le vecteur $\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_n) \in \mathbb{R}^n$, tel que $\hat{x}_i := (\hat{\alpha} + \hat{\beta}\mathrm{e}^{t_i})$, *vecteur d'ajustement*. Sur le même schéma on peut traiter les problèmes où les paramètres inconnus interviennent linéairement. Dans les sections 14.2 et 14.3, on analyse plus en détail des cas simples importants, en faisant des hypothèses supplémentaires afin d'avoir des informations sur les estimateurs déterminés par cette méthode.

14.2 Mesure d'une quantité scalaire

Le cas le plus simple est celui où la quantité à mesurer est une constante, par exemple la masse d'un objet,

$$x = m_* \,.$$

Une mesure de x est interprétée comme la valeur d'une v.a. $X = m_* + Z$. Les paramètres du modèle statistique sont désignés par m et σ^2 (variance). On fait les hypothèses suivantes :

1) Pas d'erreurs systématiques, $\mathbb{E}_{m,\sigma^2}(X_i) = m$; $\mathrm{Var}_{m,\sigma^2} X_i = \sigma^2$.
2) $X_1, \ldots, X_n$ sont identiquement distribuées et non corrélées.

On construit deux estimateurs, un pour m et un autre pour σ^2. Soit un échantillon $\mathbf{x} \in \mathbb{R}^n$ de longueur n (vecteur d'observation). Si $\hat{m}$ est l'estimation de m, le vecteur d'ajustement $\hat{\mathbf{m}} = (\hat{m}, \ldots, \hat{m})$ appartient au sous-espace E_1 engendré par le vecteur $\mathbf{d} = (1, \ldots, 1)$. Le principe de la méthode consiste à déterminer $\hat{m}$ de sorte que $\hat{\mathbf{m}} \in E_1$ et

$$\|\mathbf{x} - \hat{\mathbf{m}}\|^2 \leq \|\mathbf{x} - \mathbf{m}\|^2 \quad \forall\, \mathbf{m} = (m, \ldots, m) \in E_1 \subset \mathbb{R}^n \,.$$

L'estimation de m est donc déterminée par le vecteur $\hat{\mathbf{m}}$ qui minimise la distance de $\mathbf{x}$ à un vecteur de E_1 ; par conséquent l'estimation de m est $\hat{m}(\mathbf{x})$ où $\hat{m}(\mathbf{x})\mathbf{d}$ est la projection orthogonale de $\mathbf{x}$ sur E_1 : ($\|\mathbf{d}\| = \sqrt{n}$)

$$\hat{m}(\mathbf{x})\mathbf{d} = \left\langle \mathbf{x} \middle| n^{-\frac{1}{2}}\mathbf{d} \right\rangle (n^{-\frac{1}{2}}\mathbf{d}) \,.$$

L'estimateur de m est la moyenne empirique

$$\hat{m}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \overline{x} \,.$$

Par construction les vecteurs $(\overline{x}, \ldots, \overline{x})$ et $(x_1 - \overline{x}, \ldots, x_n - \overline{x})$ sont orthogonaux. Ce dernier vecteur appartient au sous-espace vectoriel de dimension $n-1$ orthogonal à l'espace E_1 engendré par $\mathbf{d}$.

En résumé, le résultat x_i d'une mesure est décomposé de deux manières différentes ; d'une part

$$x_i = m_* + z_i$$

où z_i est l'incertitude lors de la $i^{\text{ième}}$ mesure, et d'autre part

$$x_i = \hat{m} + r_i$$

où $\hat{m}$ est la valeur estimée de m_*, qui dépend de toutes les mesures. La méthode des moindres carrés détermine les vecteurs $\hat{\mathbf{m}}$ et $\mathbf{r} = (r_1, \ldots, r_n)$ de sorte que

$$\mathbf{x} = \hat{\mathbf{m}} + \mathbf{r} \quad \text{et} \quad \hat{\mathbf{m}} \perp \mathbf{r} \,.$$

Proposition 14.1 *Si les v.a. $X_1, \dots X_n$ sont identiquement distribuées et non corrélées, alors la v.a. $\overline{X}_n = \frac{1}{n}\sum_i X_i$ est un estimateur non biaisé de l'espérance et la v.a. $S_n^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$ est un estimateur non biaisé de la variance.*

Preuve Voir exercice 14.2. □

14.3 Régression linéaire simple

Soit t une température et x une longueur. La théorie prédit la relation $x = \alpha_* + \beta_* t$. Une telle relation est un exemple de régression linéaire simple. La régression est multiple si

$$x = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \cdots + \beta_r t_r \,;$$

les β_i sont des constantes (coefficients de la régression). Ce cas plus général peut être traité de manière analogue.

On mesure n fois la température et la longueur, $(t_1, x_1), \dots, (t_n, x_n)$. Souvent dans la pratique une des grandeurs, par exemple ici la température, peut être déterminée de façon beaucoup plus précise que l'autre grandeur. Dans ce cas on peut considérer que les incertitudes expérimentales proviennent essentiellement de la mesure des longueurs. Les rôles des grandeurs t et x sont alors différents ; t est appelée *variable explicative* et x *variable réponse*. Les résultats des mesures des longueurs sont interprétés comme les valeurs de v.a.

$$X_i = \alpha + \beta t_i + Z_i \,.$$

Pour simplifier on fait encore l'hypothèse que la précision des mesures, qui est donnée par la variance des Z_i, est la même pour chaque $i = 1, \dots, n$. On obtient un modèle avec paramètres α, β et σ^2. On suppose qu'il n'y a pas d'erreurs systématiques ;

$$\mathbb{E}_{\alpha,\beta,\sigma^2}(X_i) = \alpha + \beta t_i \quad \text{et} \quad \mathrm{Var}_{\alpha,\beta,\sigma^2}(X_i) = \sigma^2 \,.$$

On utilise la notation de la section 14.2. Soit E_2 le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les vecteurs $\mathbf{d}$ et $\mathbf{t} := (t_1, \dots, t_n)$. Si les mesures étaient parfaites le vecteur d'observation $\mathbf{x} = (x_1, \dots, x_n)$ serait donné, pour les variables explicatives $\mathbf{t}$, par la relation

$$\mathbf{x} = \alpha \mathbf{d} + \beta \mathbf{t} \in E_2 \,.$$

On estime les paramètres α et β en déterminant le vecteur de E_2 le plus proche de $\mathbf{x}$; ce vecteur est donné par la projection orthogonale de $\mathbf{x}$ sur E_2. Soient $\mathbf{n}_1 := n^{-\frac{1}{2}}\mathbf{d}$ et $\mathbf{n}_2 := \mathbf{t} - \langle \mathbf{t} | \mathbf{n}_1 \rangle \, \mathbf{n_1}$ un vecteur orthogonal à $\mathbf{n}_1$ ($\mathbf{n}_2$ n'est pas

normalisé) ; les vecteurs $\mathbf{n}_1$ et $\mathbf{n}_2$ engendrent E_2. La projection orthogonale de $\mathbf{x}$ sur E_2 est

$$\langle \mathbf{x}|\mathbf{n}_1\rangle\, \mathbf{n}_1 + \frac{1}{\|\mathbf{n}_2\|^2}\, \langle \mathbf{x}|\mathbf{n}_2\rangle\, \mathbf{n}_2 \equiv \hat{\alpha}\mathbf{d} + \hat{\beta}\mathbf{t}\,.$$

On obtient par un calcul simple $\hat{\beta}(\mathbf{x},\mathbf{t})$ et $\hat{\alpha}(\mathbf{x},\mathbf{t})$. Soit

$$\overline{t} := \frac{1}{n}(t_1 + \cdots + t_n)\,;$$

les estimateurs $\hat{\beta}$ et $\hat{\alpha}$ sont

$$\hat{\beta} = \frac{\sum_j (x_j - \overline{x})(t_j - \overline{t})}{\sum_j (t_j - \overline{t})^2} = \frac{\sum_i (t_i - \overline{t})x_i}{\sum_i t_i^2 - n\overline{t}^{\,2}}$$

et

$$\hat{\alpha} = \overline{x} - \hat{\beta}\overline{t}\,.$$

Proposition 14.2 *Si les v.a. X_i sont non corrélées et si pour tout i*

$$\mathbb{E}_{\alpha,\beta,\sigma^2}(X_i) = \alpha + \beta t_i \quad \textit{et} \quad \mathrm{Var}_{\alpha,\beta,\sigma^2}(X_i) = \sigma^2\,,$$

alors $\hat{\alpha}$ et $\hat{\beta}$ sont non biaisés ($\mathbb{E}_{\alpha,\beta,\sigma^2}(\hat{\alpha}) = \alpha$ et $\mathbb{E}_{\alpha,\beta,\sigma^2}(\hat{\beta}) = \beta$), et

$$\mathrm{Var}_{\,\alpha,\beta,\sigma^2}\hat{\alpha} = \frac{\sigma^2 \sum_i t_i^2}{n S_{\mathbf{tt}}} \quad \textit{et} \quad \mathrm{Var}_{\,\alpha,\beta,\sigma^2}\hat{\beta} = \frac{\sigma^2}{S_{\mathbf{tt}}}\,,$$

où l'on a posé

$$S_{\mathbf{tt}} = \sum_j (t_j - \overline{t})^2 = \sum_j t_j^2 - n\overline{t}^{\,2}\,.$$

L'estimateur S^2 de σ^2 est aussi non biaisé ($\mathbb{E}_{\alpha,\beta,\sigma^2}(S^2) = \sigma^2$),

$$S^2 := \frac{1}{n-2}\sum_i (X_i - \hat{\alpha} - \hat{\beta}t_i)^2\,.$$

Preuve La preuve se résume à un long calcul.

$$\begin{aligned}
S_{\mathbf{tt}}\mathbb{E}(\hat{\beta}) &= \sum_j (t_j - \overline{t})(\alpha + \beta t_j) = \alpha \sum_j (t_j - \overline{t}) + \beta \sum_j t_j (t_j - \overline{t}) \\
&= \beta\Big(\sum_j t_j^2 - \overline{t}\sum_j t_j\Big) = \beta S_{\mathbf{tt}} \quad (\text{car } \sum_j (t_j - \overline{t}) = 0).\\
S_{\mathbf{tt}}^2 \mathrm{Var}\,\hat{\beta} &= \sum_j (t_j - \overline{t})^2 \mathrm{Var} X_j \quad (\text{car } X_j \text{ non corrélées}) \\
&= \sigma^2 \sum_j (t_j - \overline{t})^2 = \sigma^2 S_{\mathbf{tt}}\,.\\
\mathbb{E}(\hat{\alpha}) &= \sum_j \frac{\alpha + \beta t_j}{n} - \overline{t}\beta = \alpha\,.
\end{aligned}$$

Dans les calculs qui suivent on utilise la relation $\mathbb{E}(Y^2) = \mathrm{Var}Y + (\mathbb{E}(Y))^2$ et

$$\sum_j (X_j - \overline{X})(\bar{t} - t_j) = -S_{\mathbf{tt}}\hat{\beta}\,.$$

$$\begin{aligned}
\mathrm{Var}\,\hat{\alpha} &= \mathrm{Var}\Big[\sum_j \Big(\frac{1}{n} - \bar{t}\,\frac{t_j - \bar{t}}{S_{\mathbf{tt}}}\Big)X_j\Big] \\
&= \frac{1}{n^2 S_{\mathbf{tt}}^2}\mathrm{Var}\Big[\sum_j \big(S_{\mathbf{tt}} - n\bar{t}(t_j - \bar{t})\big)X_j\Big] \\
&= \frac{\sigma^2}{n^2 S_{\mathbf{tt}}^2}\sum_j \Big(S_{\mathbf{tt}} - n\bar{t}(t_j - \bar{t})\Big)^2 \\
&= \frac{\sigma^2}{n^2 S_{\mathbf{tt}}^2}\sum_j \Big(S_{\mathbf{tt}}^2 - 2n\bar{t}S_{\mathbf{tt}}(t_j - \bar{t}) + n^2\bar{t}^{\,2}(t_j - \bar{t})^2\Big) \\
&= \frac{\sigma^2}{n^2 S_{\mathbf{tt}}^2}\Big(nS_{\mathbf{tt}}^2 + n^2\bar{t}^{\,2}S_{\mathbf{tt}}\Big) \\
&= \frac{\sigma^2}{nS_{\mathbf{tt}}}\Big(S_{\mathbf{tt}} + n\bar{t}^{\,2}\Big) \\
&= \frac{\sigma^2}{nS_{\mathbf{tt}}}\sum_i t_i^2 \quad (\text{car } S_{\mathbf{tt}} = \sum_j t_j^2 - n\bar{t}^{\,2})\,.
\end{aligned}$$

$$\begin{aligned}
\sum_j \Big(X_j - \hat{\alpha} - t_j\hat{\beta}\Big)^2 &= \sum_j \Big(X_j - \overline{X} + \bar{t}\hat{\beta} - t_j\hat{\beta}\Big)^2 \\
&= \sum_j (X_j - \overline{X})^2 + 2\sum_j (X_j - \overline{X})(\bar{t} - t_j)\hat{\beta} + S_{\mathbf{tt}}\hat{\beta}^2 \\
&= \sum_j (X_j - \overline{X})^2 - \hat{\beta}^2 S_{\mathbf{tt}}\,.
\end{aligned}$$

$$\begin{aligned}
\sum_j \mathbb{E}\Big(X_j - \hat{\alpha} - t_j\hat{\beta}\Big)^2 &= \sum_j \mathbb{E}\big(X_j - \overline{X}\big)^2 - S_{\mathbf{tt}}\Big(\frac{\sigma^2}{S_{\mathbf{tt}}} + \beta^2\Big) \\
&= \sum_j \mathbb{E}(X_j^2) - n\mathbb{E}(\overline{X}^2) - \sigma^2 - \beta^2 S_{\mathbf{tt}} \\
&= \sum_j \big[\sigma^2 + (\mathbb{E}(X_j))^2\big] - n\mathbb{E}(\overline{X}^2) - \sigma^2 - \beta^2 S_{\mathbf{tt}} \\
&= n\sigma^2 + \sum_j (\mathbb{E}(X_j))^2 \\
&\quad - n\frac{\sigma^2}{n} - n\big(\mathbb{E}(\overline{X})\big)^2 - \sigma^2 - \beta^2 S_{\mathbf{tt}}
\end{aligned}$$

$$
\begin{aligned}
&= (n-2)\sigma^2 + \beta^2 \sum_i t_i^2 - \beta^2 n \bar{t}^{\,2} - \beta^2 S_{\mathbf{tt}} \\
&= (n-2)\sigma^2 + \beta^2 \big[\sum_i t_i^2 - n\bar{t}^{\,2} \big] - \beta^2 S_{\mathbf{tt}} \\
&= (n-2)\sigma^2 \,.
\end{aligned}
$$

□

14.4 Modèle de Gauss et les moindres carrés

Dans le cas de la régression linéaire simple (section 14.3), si les v.a. Z_i sont i.i.d. et $Z_i \sim N(0, \sigma^2)$, on vérifie facilement que les estimateurs $\hat{\alpha}$ et $\hat{\beta}$ sont les estimateurs du maximum de vraisemblance du modèle. Comme ce sont des v.a. qui sont linéaires en X_i, ces estimateurs sont encore des v.a. gaussiennes avec espérances et variances données par la proposition 14.2. Le lien étroit entre la méthode des moindres carrés et le modèle de Gauss est dû à Gauss lui-même. Il a justifié la méthode des moindres carrés en faisant des hypothèses sur les v.a. qui modélisent les incertitudes lors d'une mesure expérimentale. Sous ces hypothèses, il en a déduit que les lois de ces v.a. sont gaussiennes. Néanmoins, en dernière instance c'est l'expérience qui permet de valider le choix de lois gaussiennes. Dans certaines situations ce choix est acceptable, dans d'autres il ne l'est pas. Selon des propos de G. Lippmann (1845-1921), rapportés par H. Poincaré (1854-1912) dans son livre *Calcul des Probabilités*[2],

« tout le monde croit que les erreurs suivent une loi normale, les expérimentateurs car ils pensent qu'il s'agit d'un théorème, et les mathématiciens qui pensent que c'est un fait expérimental. »

La remarque 13.4 est utile lorsque le nombre d'observations est grand. Lorsqu'on analyse la méthode des moindres carrés dans le modèle de Gauss on a des informations très précises sur les estimateurs.

Proposition 14.3 *Sous les hypothèses ci-dessus la v.a.*

$$S^2 = \frac{1}{n-2} \sum_i (X_i - \hat{\alpha} - \hat{\beta} t_i)^2$$

de la proposition 14.2 est indépendante des v.a. $\hat{\alpha}$ et $\hat{\beta}$. La loi de

$$(n-2)S^2/\sigma^2$$

est celle du khi-carré à $n-2$ degrés de libertés.

2. *Calcul des Probabilités*, p. 169, Editions J. Gabay (1987).

Preuve La preuve est analogue à celle de la proposition 13.3. Les v.a. $Z_i := \sigma^{-1}(X_i - \alpha - \beta t_i)$ sont i.i.d. et de loi $N(0,1)$. Soit une base orthonormale de $\mathbb{R}^n$ composée des vecteurs $\mathbf{m}_j$ tels que $\mathbf{m}_1 = \mathbf{n}_1$ et $\mathbf{m}_2 = \|\mathbf{n}_2\|^{-1}\mathbf{n}_2$ (on utilise ici et après les notations de la section 14.3). Soit $\mathbf{U}$ la matrice orthogonale dont les lignes sont les vecteurs $\mathbf{m}_j$. On pose $\mathbf{Y} := \mathbf{U}\mathbf{Z}$; les v.a. Y_i sont i.i.d. et de loi $N(0,1)$ (voir la preuve de la proposition 13.3). Le point principal est l'observation simple que le vecteur $\mathbf{v} := \alpha\mathbf{d} + \beta\mathbf{t} \in E_2$, ce qui implique

$$\hat{\alpha}\mathbf{d} + \hat{\beta}\mathbf{t} = \langle \mathbf{X}|\mathbf{m}_1\rangle\, \mathbf{m}_1 + \langle \mathbf{X}|\mathbf{m}_2\rangle\, \mathbf{m}_2 = \sigma\big(\, \langle \mathbf{Z}|\mathbf{m}_1\rangle\, \mathbf{m}_1 + \langle \mathbf{Z}|\mathbf{m}_2\rangle\, \mathbf{m}_2\big) + \mathbf{v}$$

et

$$\mathbf{X} - \hat{\alpha}\mathbf{d} + \hat{\beta}\mathbf{t} = \sum_{i=3}^{n} \langle \mathbf{X}|\mathbf{m}_i\rangle\, \mathbf{m}_i = \sigma \sum_{i=3}^{n} \langle \mathbf{Z}|\mathbf{m}_i\rangle\, \mathbf{m}_i\,.$$

On obtient de suite la loi de $(n-2)S^2/\sigma^2$ puisque

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \hat{\alpha} - \hat{\beta}t_i)^2 = \sum_{j=3}^{n} \langle \mathbf{Z}|\mathbf{m}_j\rangle^2 = \sum_{j=3}^{n} Y_j^2\,.$$

Comme les v.a. $\hat{\alpha}$ et $\hat{\beta}$ s'expriment uniquement à l'aide des v.a. Y_1 et Y_2, la v.a. S^2 est indépendante des v.a. $\hat{\alpha}$ et $\hat{\beta}$. □

14.5 Exercices

Exercice 14.1 Deux quantités mesurables x et y sont liées entre elles par la relation $y = \alpha + \beta x^2$ où α et β sont des paramètres inconnus. On mesure ces deux quantités et on trouve

$$(x_1, y_1) = (2, 3) \quad (x_2, y_2) = (4, 6) \quad (x_3, y_3) = (5, 7)\,.$$

Estimer α et β par la méthode des moindres carrés.

Exercice 14.2 Démontrer la proposition 14.1.

Exercice 14.3 On suppose que les v.a. X_i sont identiquement distribuées, non corrélées, de moyenne μ et variance σ^2. Les v.a. $Y_i := X_i - \overline{X}$ sont appelées *déviations*.
a) Vérifier que la moyenne empirique $\overline{X}$ est non corrélée avec chaque Y_i.
b) X et Y sont deux v.a. positives et non corrélées. Quelle est la plus grande variance, $\mathrm{Var}(X+Y)$ ou $\mathrm{Var}(X-Y)$?

Exercice 14.4 On considère une régression linéaire $x = \alpha + \beta t$ où t est la variable explicative et x la variable réponse. On a observé

t	:	1	2	3	4	5	6	7
x	:	3	5	8	10	10	12	15

Donner la valeur des estimateurs $\hat{\alpha}$, $\hat{\beta}$ et S^2 de la proposition 14.2.

Exercice 14.5 On modélise la mesure expérimentale d'une quantité scalaire z par une v.a. réelle. La variance de cette v.a. donne des renseignements sur la précision de la mesure.
a) On suppose que $z = axy$ et qu'on peut mesurer de manière indépendante x et y de sorte qu'on obtient une v.a. $Z = aXY$, produit de deux v.a. indépendantes. Exprimer l'espérance et la variance de Z en fonction des espérances et des variances de X et Y. Donner un estimateur non biaisé de l'espérance de Z.
b) On suppose que $y = g(x)$ où g est une fonction continue et bornée. On peut mesurer expérimentalement x de sorte qu'on obtient une v.a. $Y = g(X)$. Comme estimateur de la moyenne de Y on utilise $g(\overline{X}_n)$ où $\overline{X}_n$ est la moyenne empirique des mesures de la quantité x. Justifier ce choix en montrant que cette suite d'estimateurs est consistante (on admet qu'il n'y a pas d'erreurs systématiques).
Indication : voir chapitre 12.5.

Chapitre 15

Estimation par intervalle

Les estimateurs considérés jusqu'ici sont des estimateurs ponctuels qui donnent pour chaque échantillon $\mathbf{x}$ une estimation $\hat{\theta}(\mathbf{x})$ du paramètre θ. Dans ce chapitre on examine une autre manière de faire des estimations ; on donne non plus une valeur estimée unique, mais un ensemble de valeurs. Lorsque le paramètre à estimer est réel, cet ensemble est en général un intervalle.

15.1 Domaine de confiance

Le cadre est le même que dans les chapitres précédents. Les données empiriques proviennent de n v.a. $X_1, \dots, X_n$ réelles qui forment un vecteur aléatoire $\mathbf{X} := (X_1, \dots, X_n)$. Pour analyser ces données empiriques on suppose :

1) la loi conjointe des X_i, $f(\mathbf{x}; \theta)$, est connue pour chaque $\theta \in \Theta$;
2) les données empiriques correspondent à des v.a. dont la loi conjointe est $f(\mathbf{x}; \theta_*)$; la valeur θ_* est fixe mais inconnue.

Pour estimer la valeur inconnue du paramètre on construit pour chaque échantillon $\mathbf{x}$ un sous-ensemble $C(\mathbf{x}) \subset \Theta$ tel que

$$P_{\theta_*}(\{\mathbf{y} \colon C(\mathbf{y}) \ni \theta_*\}) = 1 - \alpha$$

avec α petit. Cela signifie que la probabilité de l'événement « la valeur θ_* inconnue appartient à $C(\mathbf{x})$ », calculée avec P_{θ_*}, est $1 - \alpha$.

Définition 15.1 *Soit* $0 < \alpha < 1$. *Un* domaine de confiance *de* niveau de confiance $1 - \alpha$, *ou* seuil de confiance α, *est une application qui à chaque* $\mathbf{x} \in \Sigma$ *associe un sous-ensemble* $C(\mathbf{x}) \subset \Theta$ *tel que pour tout* θ

$$P_\theta\big(\{\mathbf{y} \colon C(\mathbf{y}) \ni \theta\}\big) = 1 - \alpha \,. \tag{15.1}$$

Le seuil α *est petit, par exemple* 0,05 *ou* $\alpha = 0{,}01$.

Lorsque le paramètre est réel on construit en général un intervalle dont les bords sont spécifiés par deux v.a. $\tau_\pm$,

$$C(\mathbf{x}) = [\tau_-(\mathbf{x}), \tau_+(\mathbf{x})] \subset \Theta \,.$$

Remarque 15.1 Jusqu'ici les événements étaient exprimés sous la forme $\{X \in B\}$ où X est une v.a. et $B \subset \Omega$ est un ensemble donné. Ici c'est l'intervalle $[\tau_-(\cdot), \tau_+(\cdot)]$ qui est aléatoire. Dans l'approche exposée dans ce livre, le paramètre θ n'est pas une quantité aléatoire.

$$P_\theta([\tau_-, \tau_+] \ni \theta) = \int f(\mathbf{x};\theta) I_{\{\mathbf{y}:\, [\tau_-(\mathbf{y}),\tau_+(\mathbf{y})] \ni \theta\}}(\mathbf{x})\, d\mathbf{x}\,.$$

Une fois l'expérience réalisée, $\mathbf{x}$ est connu ; *pour n'importe quel* θ on peut dire sans ambiguïté si l'affirmation $C(\mathbf{x}) \ni \theta$ est vraie ou non. □

Dans la définition 15.1 le point important est que l'affirmation (15.1) est vraie uniformément en θ. Elle est donc vraie pour θ_*. L'interprétation probabiliste d'un domaine de confiance de niveau $1-\alpha$ est la suivante :

> *La probabilité, calculée avec* P_{θ_*}*, que lors d'une réalisation de* $\mathbf{X}$ *le domaine de confiance contienne* θ_**, est égale à* $1-\alpha$ *:*
>
> $$P_{\theta_*}\big(\{\mathbf{x}\colon C(\mathbf{x}) \ni \theta_*\}\big) = 1-\alpha\,.$$

Ce qui est particulier ici c'est le fait que, malgré la remarque 15.1, on ne peut pas *savoir* si l'événement $\{\mathbf{x}\colon C(\mathbf{x}) \ni \theta_*\}$ est réalisé ou non lorsque l'échantillon est connu. En effet, la valeur (non aléatoire) θ_* du paramètre est inconnue. Cependant, cet événement est bien défini, ainsi que sa probabilité, qui est égale à $1-\alpha$, lorsqu'elle est calculée par rapport à la mesure de probabilité P_{θ_*} du modèle. Une fois que l'échantillon est obtenu cet événement est soit vrai, soit faux. Ceci est illustré dans la remarque 15.2.

Le principe de construction d'un domaine de confiance est simple : pour chaque $\theta \in \Theta$ on choisit un événement $A(\theta)$ qui ne dépend que des observations et tel que $P_\theta(A_\theta) = 1-\alpha$. L'événement $A(\theta)$ est « typique » pour la mesure de probabilité P_θ. Ce choix n'est pas univoque, mais souvent il y en a un qui est naturel et qui s'impose. Une fois ce choix fait, lorsqu'on obtient des données expérimentales $\mathbf{x}$, on examine pour ces données expérimentales si l'événement $A(\theta)$ est réalisé ou non. Si $A(\theta)$ est réalisé, θ fait partie du domaine de confiance :

$$C(\mathbf{x}) := \{\theta \in \Theta :\ A(\theta) \text{ est réalisé pour } \mathbf{x}\}\,.$$

Ainsi, par construction,

$$\theta \in C(\mathbf{x}) \iff \mathbf{x} \in A(\theta)\,.$$

Par conséquent

$$\forall \theta\colon\ P_\theta(C(\mathbf{X}) \ni \theta) = P_\theta(\mathbf{X} \in A(\theta)) = P_\theta(A(\theta)) = 1-\alpha\,.$$

15.1.1 Intervalle de confiance pour le modèle de Gauss

On construit des intervalles de confiance pour les deux paramètres m et σ^2 du modèle. La construction de ces intervalles est facilitée par l'existence de v.a. pivot. On donne d'abord une définition.

Définition 15.2 *Soit* $0 < p < 1$. *Le* p-quantile *d'une fonction de répartition* F *est la valeur* $q(p)$ *telle que* $F(q(p)) = p$.

Exemple 15.1 La médiane d'une v.a. est le 1/2-quantile de sa fonction de répartition. Pour une loi $N(0,1)$,

$$q_{N(0,1)}(0{,}975) = 1{,}96\,.$$

Pour une loi de Student t_n

n	:	10	30	60
$q_{t_n}(0{,}975)$	:	2,228	2,042	2

□

a) On suppose que σ^2 est connu. L'estimateur de maximum de vraisemblance du paramètre m est $\overline{X}_n \sim N(m, \sigma^2/n)$. Pour tout $n \geq 1$ la v.a.

$$Y_n := \frac{\overline{X}_n - m}{\sqrt{\mathrm{Var}\overline{X}_n}} \sim N(0,1)$$

a une loi qui ne dépend pas du paramètre m. Une telle v.a. est appelée *v.a. pivot* car sa loi ne dépend pas du paramètre qu'on estime. On introduit le quantile $z_\alpha = q_{N(0,1)}(1-\alpha)$ qui est caractérisé par

$$P(Y > z_\alpha) = \alpha \quad \text{si } Y \sim N(0,1).$$

$$z_{0,01} = 2{,}33 \qquad z_{0,025} = 1{,}96 \qquad z_{0,05} = 1{,}645\,.$$

Un choix naturel pour $A(m)$ est l'événement

$$\begin{aligned} A(m) :&= \left\{\mathbf{x}\colon \left|\frac{\overline{X}_n(\mathbf{x}) - m}{\sqrt{\mathrm{Var}\overline{X}_n}}\right| \leq z_{\alpha/2}\right\} \\ &= \left\{\mathbf{x}\colon m - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \overline{X}_n(\mathbf{x}) \leq m + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right\} \end{aligned}$$

de sorte que $P_m(A(m)) = 1-\alpha$. Pour simplifier l'écriture on pose $\overline{x}_n = \overline{X}_n(\mathbf{x})$. L'intervalle de confiance est par définition

$$\begin{aligned} I(\mathbf{x}) &= \{m\colon A(m) \text{ est réalisé pour } \mathbf{x}\} \\ &= \left\{m \in \mathbb{R}\colon \overline{x}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq m \leq \overline{x}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right\} \qquad (15.2) \\ &= [\tau_-(\mathbf{x}), \tau_+(\mathbf{x})] \quad \text{où} \quad \tau_\pm(\mathbf{x}) = \overline{x}_n \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\,. \end{aligned}$$

Remarque 15.2 L'exemple suivant illustre la signification d'un intervalle de confiance. Soit X une v.a. de loi $N(0,1)$. On a à disposition N valeurs indépendantes de X. Sachant que X a une loi $N(m,1)$ on estime la valeur du paramètre $m = \theta$ à partir de ces N valeurs de X à disposition. Ici $\theta_* = 0$. Pour estimer θ on calcule pour chacune de ces valeurs deux intervalles de confiance (15.2)

avec $\alpha = 0.05$ et $\alpha = 0.32$ (ici $n = 1$). Si $\alpha = 0.05$ l'intervalle a une longueur $2 \times 1{,}96 = 3.92$ et si $\alpha = 0.32$ la longueur et approximativement 2. La probabilité que chaque intervalle contienne la valeur 0 est de 0.95 si $\alpha = 0.5$ et de 0,68 si $\alpha = 0.32$. Si N est grand la LGN, qui relie la probabilité et la fréquence relative d'un événement, implique que l'on s'attend à obtenir environ 5% des intervalles de confiance ne contenant pas $\theta_* = 0$, si $\alpha = 0.05$ et 32%, si $\alpha = 0.32$. On peut constater ce fait dans la figure 15.1 où l'on a reporté les intervalles de confiance pour $N = 100$. □

b) Si m et σ^2 sont inconnus, on construit de la même façon un intervalle de confiance pour m en utilisant la proposition 13.3. En effet, pour chaque $n \geq 1$ T_n est une v.a. pivot puisque sa loi ne dépend pas de m et σ^2. Ce qui change dans (15.2), c'est que σ est remplacé par son estimateur $S_n(\mathbf{x})$ et $z_{\alpha/2}$ est remplacé par $t_{\alpha/2,n-1}$, où $t_{\alpha,n}$ est défini par

$$P(Y > t_{\alpha,n}) = \alpha \quad \text{si } Y \sim t_n.$$

Par exemple pour $\alpha = 0{,}05$, l'intervalle de confiance est

$$\begin{aligned} I(\mathbf{x}) &= \Big\{ m \in \mathbb{R} \colon \overline{x}_n - \frac{S_n(\mathbf{x})}{\sqrt{n}} t_{0{,}025,\, n-1} \leq m \leq \overline{x}_n + \frac{S_n(\mathbf{x})}{\sqrt{n}} t_{0{,}025,\, n-1} \Big\} \\ &= [\tau_-(\mathbf{x}), \tau_+(\mathbf{x})] \quad \text{où} \quad \tau_\pm(\mathbf{x}) = \overline{x}_n \pm \frac{s_n}{\sqrt{n}} t_{0{,}025,\, n-1}\,. \end{aligned}$$

avec

$$S_n^2(\mathbf{x}) = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x}_n)^2\,.$$

Remarque 15.3 Les physiciens adoptent une autre convention pour donner une estimation de la précision d'une mesure. Les résultats sont énoncés sous la forme

$$\overline{x}_n - \frac{s_n}{\sqrt{n}} \leq m \leq \overline{x}_n + \frac{s_n}{\sqrt{n}} \quad \text{avec } s_n := S_n(\mathbf{x}).$$

Cela revient à choisir $t_{\alpha/2,\, n-1} = q_{t_{n-1}}(1 - \frac{\alpha}{2}) = 1$. Si $n = 10$, cette convention correspond à choisir pour le modèle de Gauss $\alpha = 0{,}34$! Si l'on fait 100 séries de 10 mesures chacune, on a une situation qui ressemble à celle de la deuxième colonne de la figure 15.1. □

c) Pour construire des intervalles de confiance pour la variance on utilise la v.a. pivot (proposition 13.2)

$$(n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1}\,.$$

Soit $\chi^2_{\alpha,n}$ défini par

$$P(Y > \chi^2_{\alpha,n}) = \alpha \quad \text{si } Y \sim \chi^2_n.$$

L'événement

$$B(\sigma^2) := \Big\{ \mathbf{x} \colon \chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S_n^2(\mathbf{x})}{\sigma^2} \leq \chi^2_{\alpha/2,n-1} \Big\}$$

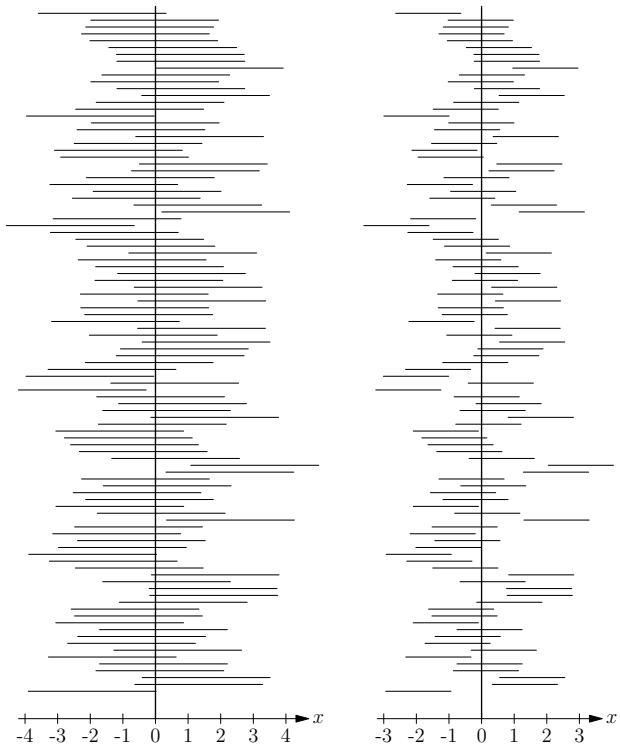


Figure 15.1 Simulation de 100 intervalles de confiance pour une v.a. de loi $N(0,\,1)$ (voir remarque 15.2). A gauche le niveau est 0,95. A droite, c'est la convention des physiciens qui est utilisée (voir remarque 15.3), ce qui correspond à $\alpha = 0{,}32$ à la place de $\alpha = 0{,}05$. Dans la simulation plus du tiers des intervalles, construits avec la convention des physiciens, ne contiennent pas la vraie valeur du paramètre $\theta_* = 0$.

vérifie $P_{m,\sigma^2}\big(B(\sigma^2)\big) = 1-\alpha$. Par conséquent on obtient l'intervalle de confiance de niveau $1-\alpha$

$$I(\mathbf{x}) = \Big\{\sigma^2 \colon \frac{(n-1)S_n^2(\mathbf{x})}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)S_n^2(\mathbf{x})}{\chi^2_{1-\alpha/2,n-1}}\Big\}\,.$$

15.1.2 Intervalle de confiance pour des v.a. de Bernoulli

Soit $X_1, \ldots, X_n$ des v.a. i.i.d. de Bernoulli de paramètre p qui est inconnu. L'estimateur de maximum de vraisemblance du paramètre p est (voir exemple 13.6)

$$\hat{p}_n(\mathbf{x}) = \frac{1}{n}\big(\sum_i x_i\big) \equiv \overline{x}_n\,.$$

Cet estimateur est non biaisé et sa variance $\operatorname{Var}\hat{p}_n = \frac{p(1-p)}{n}$ converge vers 0 si n diverge. La suite de ces estimateurs est donc consistante (proposition 13.1).

Soit

$$Y_n := \frac{n^{1/2}\big(\overline{X}_n - p\big)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}\,.$$

On sait par la LGN et par le TLC que

$$\overline{X}_n \xrightarrow{\mathcal{L}} p \quad \text{et} \quad \frac{n^{1/2}\big(\overline{X}_n - p\big)}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} Y\,,\ Y \sim N(0,1)\,.$$

On peut aussi montrer que $Y_n \xrightarrow{\mathcal{L}} Y$. *Asymptotiquement* on obtient une v.a. pivot Y. La v.a. Y_n est fréquemment utilisée pour la construction d'un intervalle de confiance de la manière suivante. On procède comme dans le cas a) du modèle gaussien en prenant

$$A(p) := \big\{\mathbf{x}\colon |Y_n| \le z_{\alpha/2}\big\}\,.$$

On a donc remplacé le quantile $q(1-\alpha/2)$ de la loi de Y_n par celui de la loi limite $N(0,1)$. À cause de cette approximation on obtient un intervalle de confiance

$$I(\mathbf{x}) = \Big\{p\,:\, \overline{x}_n - z_{\alpha/2}\frac{\sqrt{\overline{x}_n(1-\overline{x}_n)}}{\sqrt{n}} \le p \le \overline{x}_n + z_{\alpha/2}\frac{\sqrt{\overline{x}_n(1-\overline{x}_n)}}{\sqrt{n}}\Big\}\,.$$

Le niveau de confiance est $1-\alpha$ *seulement* dans la limite $n \to \infty$.

Exemple 15.2 Un journal rapporte le résultat d'un sondage avant une votation : 52% de la population est favorable au projet soumis à votation avec une marge d'erreur de ±4%. Comment analyser ce sondage ? On va supposer que les personnes sondées ont été choisies au hasard et que l'usage du modèle ci-dessus est justifié. De plus, on suppose que l'intervalle de confiance donné par le journal est de niveau 0,95 (pratique courante). Dans le modèle avec les variables de Bernoulli, $X_i = 1$ si la personne i est favorable au projet. L'estimation de p est $\hat{p}_n = 0{,}52$ et l'intervalle de confiance de niveau 0,95 est

$$\Big(0{,}52 - 1{,}96\frac{\sqrt{0{,}52\cdot 0{,}48}}{\sqrt{n}},\, 0{,}52 + 1{,}96\frac{\sqrt{0{,}52\cdot 0{,}48}}{\sqrt{n}}\Big)\,.$$

La marge d'erreur est de 4% ; le nombre n de personnes interrogées est donc

$$1{,}96\frac{\sqrt{0{,}52\cdot 0{,}48}}{\sqrt{n}} \approx 0{,}04 \implies n = \frac{(1{,}96)^2(0{,}52)(0{,}48)}{(0{,}04)^2} = 599{,}29\,,$$

soit environ 600 personnes. □

15.2 Exercices

Exercice 15.1 Combien de personnes aurait-il fallu interroger dans l'exemple 15.2 pour qu'on obtienne une marge d'erreur de 0,02 ?
Indication : utiliser l'estimation de p obtenue par le journal.

Exercice 15.2 On a obtenu l'échantillon suivant pour des v.a. gaussiennes i.i.d.

42,70	43,48	43,63	42,78	43,18
42,56	42,76	42,87	42,95	43,39
43,01	43,06	41,60	43,20	43,10

a) Calculer la moyenne empirique et la variance empirique.
b) Donner un intervalle de confiance de niveau 0,95 pour la variance.

Exercice 15.3 Construire un intervalle de confiance de niveau $1-\alpha$ pour l'estimateur de maximum de vraisemblance $\hat{\theta}$ de l'exercice 13.4.
Indication : l'estimateur est $\max_{i=1}^{n} X_i$.

Exercice 15.4 a) Construire un intervalle de confiance de niveau $1-\alpha$ à partir de l'estimateur construit dans l'exercice 13.5.
b) Considérer le cas où n est grand et $\alpha = 0{,}05$.
Indication : l'estimateur est $\dfrac{2}{n}\sum_{i=1}^{n} X_i$.

Exercice 15.5 On considère une régression linéaire $x = \alpha + \beta t$ avec t comme variable explicative. Les v.a. Z_i, qui décrivent les incertitudes lors des mesures, sont i.i.d. et de loi $N(0, \sigma^2)$. On reprend les données de l'exercice 14.4.
a) Construire une v.a. pivot pour le paramètre α.
Indication : à partir des résultats des propositions 14.2 et 14.3 on peut construire une v.a. pivot de loi t_{n-2}.
b) Construire un intervalle de confiance de niveau 0,95 pour α.

Chapitre 16

Test

Dans ce dernier chapitre on expose brièvement quelques notions de base concernant les tests statistiques. On suppose qu'on a un modèle statistique ; l'expérience aléatoire qu'on étudie est décrite par $(\Omega, \mathcal{F}, P_{\theta_*})$ pour une valeur θ_* fixe, mais inconnue du paramètre θ. On veut déterminer dans quelle mesure une propriété postulée du modèle est compatible avec les résultats expérimentaux $\mathbf{x}$. En d'autres termes, si l'on suppose que la propriété postulée est vraie, est-ce que les résultats expérimentaux $\mathbf{x}$ sont *compatibles avec l'hypothèse* dans le sens que *les résultats expérimentaux peuvent être « expliqués » par le mécanisme aléatoire de l'expérience.* Un test de signification permet de juger si une différence est réelle ou si elle est due à une variation aléatoire.

16.1 Test de signification, p-valeur

Dans un modèle statistique on appelle *hypothèse* H_0 une affirmation sur la valeur θ_* du paramètre θ. *Une hypothèse ne dépend pas des observations.* Elle est juste ou fausse. Formellement une hypothèse est identifiée à un sous-ensemble $\Theta_0 \subset \Theta$:

$$\Theta_0 := \{\theta \in \Theta \colon \text{l'affirmation } H_0 \text{ est vraie pour } \theta\}\,.$$

La négation de l'affirmation H_0 est appelée *hypothèse alternative* H_1. Elle est identifiée à $\Theta_1 = \Theta \backslash \Theta_0$.

Exemple 16.1 On lance n fois une pièce de monnaie. On veut tester l'hypothèse H_0 « la pièce est équilibrée ». Le modèle statistique est donné par

$$\Omega = \{0,1\}^n\,,\ P_\theta(\omega) = \prod_{j=1}^{n} \theta^{\omega_j} \prod_{j=1}^{n} (1-\theta)^{1-\omega_j} \quad \text{avec } 0 < \theta < 1.$$

Pile est codé par 1 et $P(Pile) = \theta$. Ici $\Theta_0 = \{\frac{1}{2}\}$ et $\Theta_1 = (0,1)\backslash\{\frac{1}{2}\}$ correspond à l'hypothèse alternative « le dé n'est pas équilibré ». □

Exemple 16.2 On fait n mesures d'une quantité scalaire. On se place dans le cas du modèle de Gauss et on suppose que σ^2 est connu.

a) On teste l'hypothèse H_0 « la quantité vaut m_0 ». Dans ce cas $\Theta_0 = \{m_0\}$ et $\Theta_1 = \mathbb{R}\backslash\{m_0\}$.
b) On teste l'hypothèse H_0 « la quantité est inférieure ou égale à m_0 ». Dans ce cas $\Theta_0 = \{m \in \mathbb{R}\colon m \leq m_0\}$ et $\Theta_1 = \{m \in \mathbb{R}\colon m > m_0\}$. □

Pour analyser l'hypothèse H_0 on peut faire un *test de signification*. Ce test permet de juger la signification de l'écart entre les résultats expérimentaux et les résultats théoriques, *si l'on suppose que* H_0 *est vraie*. Le résultat est donné sous la forme d'une p-valeur (voir définition 16.1). Les tests les plus simples font intervenir une *statistique de test*, telle que les grandes valeurs de X indiquent que les résultats obtenus sous l'hypothèse H_0 sont difficilement explicables uniquement par le mécanisme aléatoire du modèle de l'expérience. Dans le cas de l'exemple 16.1 on pose $S := \#$piles de l'échantillon et on choisit la statistique de test

$$X := \left|S - \frac{n}{2}\right| \qquad \left(\mathbb{E}_{\frac{1}{2}}(S) = \frac{n}{2}\right).$$

Pour analyser l'exemple 2 on introduit la v.a.

$$Z := \frac{\overline{X}_n - m_0}{\sigma}\sqrt{n}\,.$$

Dans le cas 16.2 a) on choisit $X = |Z|$. Le cas 16.2 b) est un peu différent. On s'intéresse ici aux valeurs de $\overline{X}_n > m_0$ puisque ce sont ces valeurs qui peuvent nous indiquer si $m_* > m_0$. Dans ce cas on prend $X = Z$; ce choix est compatible avec H_0 et ne favorise pas les valeurs $\overline{X}_n > m_0$ (voir aussi l'exemple 16.6).

Définition 16.1 *La* p-valeur *du test est la probabilité, calculée lorsque l'hypothèse* H_0 *est vraie, que la statistique de test* X *prenne la valeur observée ou une autre valeur indiquant un écart plus extrême par rapport à l'hypothèse* H_0.

Exemple 16.3 Dans le cas de l'exemple 16.1 on suppose qu'on a fait 10 lancers et qu'on a trouvé les résultats $1, 0, 0, 0, 1, 0, 0, 0, 0, 0$. La valeur de X pour cet échantillon est 3 ; par conséquent la p-valeur sous H_0 est

$$\begin{aligned}\mathrm{Prob}(X \geq 3) &= P_{\frac{1}{2}}(S \in \{0, 1, 2, 8, 9, 10\}) \\ &= \frac{2}{2^{10}}\left(\binom{10}{0} + \binom{10}{1} + \binom{10}{2}\right) = 0{,}11\,.\end{aligned}$$

La p-valeur n'est pas très grande, mais le test de signification ne remet pas en question l'hypothèse H_0. □

Exemple 16.4 a) Pour l'exemple 16.2 a), sous l'hypothèse H_0, la p-valeur est calculée avec la mesure de probabilité P_{m_0}. Si $|X| = 3{,}09$,

$$P_{m_0}(|Z| \geq 3{,}09) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-3{,}09} \mathrm{e}^{-\frac{t^2}{2}}\,dt + \frac{1}{\sqrt{2\pi}}\int_{3{,}09}^{\infty} \mathrm{e}^{-\frac{t^2}{2}}\,dt = 0{,}002\,.$$

b) Dans le cas 16.2 b), si $X = 1{,}28$, la p-valeur sous H_0 vaut

$$\begin{aligned}\sup_{m\le m_0} P_m(Z \ge 1{,}28) &= \sup_{m\le m_0} P_m\Big(\frac{\overline{X}_n - m}{\sigma}\sqrt{n} \ge 1{,}28 + \frac{(m_0-m)\sqrt{n}}{\sigma}\Big)\\ &= \sup_{m\le m_0}\Big(1 - \Phi\Big(1{,}28 + \frac{(m_0-m)\sqrt{n}}{\sigma}\Big)\Big)\\ &= 1 - \Phi(1{,}28) = 0{,}1\,.\end{aligned}$$

Si on ne connaît pas σ^2 on utilise $T_n \sim t_{n-1}$ à la place de Z. □

Dans l'exemple 16.4 a) la p-valeur est très petite. En général on considère que l'écart est *statistiquement significatif* lorsque la p-valeur est inférieure à une valeur petite appelée *seuil de signification*. Très souvent ce niveau est fixé à 0,05. Par conséquent dans le cas 16.4 a) on doute fortement de l'hypothèse H_0 sur la base des résultats obtenus.

Exemple 16.5 Fréquence d'émission radioactive non connue. On suppose que les émissions ont lieu à des intervalles indépendants et que l'intervalle de temps entre deux émissions est bien modélisé par une v.a. exponentielle de paramètre θ inconnu, $f_\theta(t) = \theta \mathrm{e}^{-\theta t} I_{\mathbb{R}^+}(t)$. On mesure n intervalles de temps successifs, $x_1, \dots, x_n$. La loi de $X := X_1 + \cdots + X_n$, où X_i est la v.a. donnant la longueur du $i^{\text{ième}}$ intervalle de temps, est une loi d'Erlang (loi gamma de paramètres (n, θ)) de densité

$$f_X(t) = \theta^n \frac{t^{n-1}}{(n-1)!}\mathrm{e}^{-\theta t} I_{\mathbb{R}^+}(t)\,.$$

La v.a. $2\theta X \sim \chi^2_{2n}$; en effet

$$P(2\theta X \le t) = P(X \le t/(2\theta))\,,$$

et par conséquent la loi de $2\theta X$ est une loi gamma de paramètres $(n, 1/2)$. L'estimateur de maximum de vraisemblance de θ est $\hat\theta = n/X$ et

$$2\theta X = 2n\frac{\theta}{\hat\theta} \sim \chi^2_{2n} \quad \text{avec} \quad P_\theta\Big(\theta > \frac{\hat\theta}{2n}\chi^2_{\alpha,2n}\Big) = \alpha\,.$$

Si une théorie prédit $\theta \le \theta_0$, et si après avoir fait 10 mesures on obtient $\hat\theta(\mathbf{x}) = 1{,}6\theta_0$, est-ce que cet écart est statistiquement significatif ? Pour répondre à cette question on calcule la p-valeur du test sous l'hypothèse $H_0 = \{\theta \le \theta_0\}$,

$$\begin{aligned}\sup_{\theta\le\theta_0} P_\theta(\hat\theta \ge 1{,}6\theta_0) &= \sup_{\theta\le\theta_0}\big(1 - P_\theta(\hat\theta < 1{,}6\theta_0)\big)\\ &= 1 - \inf_{\theta\le\theta_0} P_\theta\Big(\frac{\hat\theta}{20}\frac{20\theta}{1{,}6\theta_0} \le \theta\Big)\,.\end{aligned}$$

On identifie $\chi^2_{\alpha,20}$ à $(20\theta)/(1{,}6\theta_0)$; α est minimal si $\theta = \theta_0$, et par conséquent on obtient

$$\sup_{\theta\le\theta_0} P_\theta(\hat\theta \ge 1{,}6\theta_0) = 1 - \alpha_0 \quad \text{avec} \quad \chi^2_{\alpha_0,20} = 20/1{,}6 = 12{,}5\,.$$

Cela correspond à $\alpha_0 \approx 0.9$. Le résultat expérimental peut être considéré statistiquement consistant avec la théorie ; ce résultat expérimental n'est donc pas statistiquement significatif, car il peut être « expliqué » par le mécanisme aléatoire de l'expérience en supposant que la théorie est correcte. □

16.2 Erreurs de première et deuxième espèce

Dans cette section on reprend la même problématique sous un angle un peu différent. On formalise la situation où, à l'issue du test, on décide de rejeter ou non l'hypothèse H_0. Toute décision est liée à un risque qui est quantifié par le niveau du test. Ce risque dépend entre autres du choix de la statistique utilisée pour faire le test. La région critique définie ci-dessous est la région pour laquelle on décide que les valeurs empiriques sont statistiquement significatives ; dans ce cas on rejette H_0. La marche à suivre est la suivante.

1) On formule une hypothèse H_0 ($\Theta_0 \subset \Theta$).
2) On choisit un test, ici une statistique de test X.
3) On fixe *le niveau du test* $1-\alpha$, α petit, par exemple $\alpha = 0{,}05$ ou $\alpha = 0{,}01$.
4) On détermine une *région critique* du test qui est un sous-ensemble $\mathcal{R} \subset \Sigma$ de l'espace des échantillons tel que
$$\sup_{\theta \in \Theta_0} P_\theta(\mathcal{R}) \leq \alpha\,.$$
5) On fait le test ; si l'échantillon obtenu $\mathbf{x} \in \mathcal{R}$, alors on rejette H_0. Dans le cas contraire on ne rejette pas H_0.

Exemple 16.6 On reprend l'exemple 16.2. Dans le cas *a*) on teste l'affirmation $m_* = m_0$ contre $m_* \neq m_0$. On choisit le seuil $\alpha = 0{,}05$.

$$P_{m_0}\big(|Z| \geq 1{,}96\big) = 0{,}05\,.$$

La région critique est $\big(\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i\big)$

$$\begin{aligned}\mathcal{R} &= \{\mathbf{x} : |Z(\mathbf{x})| \geq 1{,}96\} \\ &= \Big\{\mathbf{x} : \overline{x}_n \leq m_0 - 1{,}96\frac{\sigma}{\sqrt{n}}\Big\} \cup \Big\{\mathbf{x} : \overline{x}_n \geq m_0 + 1{,}96\frac{\sigma}{\sqrt{n}}\Big\}\,.\end{aligned}$$

Dans le cas *b*) on teste l'affirmation $m_* \leq m_0$ i.e. $H_0 := \{\theta : \theta \leq m_0\}$; pour tout $\theta \leq m_0$

$$\begin{aligned}\alpha = P_\theta\Big(\sqrt{n}\frac{\overline{X}_n - \theta}{\sigma} \geq z_\alpha\Big) &= P_\theta\Big(\sqrt{n}\frac{\overline{X}_n - m_0}{\sigma} \geq z_\alpha + \underbrace{\sqrt{n}\frac{\theta - m_0}{\sigma}}_{\leq 0 \text{ sous } H_0}\Big) \\ &\geq P_\theta\Big(\sqrt{n}\frac{\overline{X}_n - m_0}{\sigma} \geq z_\alpha\Big)\end{aligned}$$

On peut prendre comme région critique si $\alpha = 0{,}05$ ($z_{0,05} = 1{,}645$)

$$\mathcal{R} := \Big\{ \mathbf{x} : \overline{x}_n \geq m_0 + 1{,}645 \frac{\sigma}{\sqrt{n}} \Big\} .$$

□

Dans cette procédure de décision on peut faire deux sortes d'erreurs :

erreur de première espèce : $\theta_* \in \Theta_0$ et $\mathbf{x} \in \mathcal{R}$;

erreur de deuxième espèce : $\theta_* \notin \Theta_0$ et $\mathbf{x} \notin \mathcal{R}$.

Par convention on minimise l'erreur de première espèce. On contrôle les probabilités d'erreurs de première espèce *uniformément* dans le paramètre θ ; si $\theta^* \notin \Theta_0$ on ne peut pas faire d'erreur de première espèce.

La logique du test est un *argument par contradiction*. L'hypothèse H_0 exprime une position défavorable vis-à-vis des faits qu'on souhaite mettre en évidence. Le choix de H_0 est donc important. Les deux hypothèses H_0 et H_1 ne sont pas traitées de la même façon. On prend donc le risque de considérer qu'un événement $\{\mathbf{X} \in \mathcal{R}\}$ de probabilité inférieure à α, *si l'hypothèse* H_0 *est vraie*, ne se produira pas. Le point de vue adopté est que si l'événement $\{\mathbf{X} \in \mathcal{R}\}$ se produit, on a des raisons statistiques pour mettre en doute H_0 : on *rejette* H_0. Dans le cas contraire *on ne rejette pas* H_0.

On définit

$$\beta(\theta) := P_\theta(\mathbf{X} \in \mathcal{R}) .$$

La fonction $\beta(\theta)$, restreinte à $\Theta \backslash \Theta_0$, est appelée *puissance du test*. Pour un seuil donné, on préfère un test avec grande puissance, afin de minimiser aussi l'erreur de deuxième espèce. En effet, si $\theta_* \in \Theta_1$, la probabilité de faire une erreur de deuxième espèce est $1 - \beta(\theta_*)$. Il est possible que $\beta(\theta_*)$ soit petit et donc $P_{\theta_*}(\mathbf{X} \notin \mathcal{R})$ grand ; la probabilité de faire une erreur de deuxième espèce peut être grande.

Exemple 16.7 Dans l'exemple 16.6 a)

$$\begin{aligned}
\beta(\theta) &= P_\theta\Big(\Big| \frac{\overline{X}_n - m_0}{\sigma} \sqrt{n} \Big| \geq 1{,}96 \Big) \\
&= P_\theta\Big(\frac{\overline{X}_n - \theta}{\sigma} \sqrt{n} \geq 1{,}96 + \frac{(m_0 - \theta)\sqrt{n}}{\sigma} \Big) \\
&+ P_\theta\Big(\frac{\overline{X}_n - \theta}{\sigma} \sqrt{n} \leq -1{,}96 + \frac{(m_0 - \theta)\sqrt{n}}{\sigma} \Big) \\
&= 1 - \Phi\Big(1{,}96 + \frac{m_0 - \theta}{\sigma} \sqrt{n} \Big) + \Phi\Big(-1{,}96 + \frac{m_0 - \theta}{\sigma} \sqrt{n} \Big) .
\end{aligned}$$

Dans l'exemple 16.6 b)

$$\beta(\theta) = P_\theta\Big(\frac{\overline{X}_n - m_0}{\sigma} \sqrt{n} \geq 1{,}645 \Big) = 1 - \Phi\Big(1{,}645 + \frac{m_0 - \theta}{\sigma} \sqrt{n} \Big) .$$

Lorsque H_1 est vraie, $\beta(\theta) \geq \alpha$. De plus $\lim_{\theta\to\infty} \beta(\theta) = 1$. La probabilité de faire une erreur de deuxième espèce n'est pas nécessairement petite. La puissance de $\beta(\theta)$ dépend de n ; la probabilité de faire une erreur de deuxième espèce diminue lorsque n augmente.

Exemple 16.8 On teste l'hypothèse H_0 « un dé est équilibré ». Le test du χ^2 de Pearson (1857-1936) permet de tester si une v.a. X discrète, prenant les valeurs $1, \ldots, k$, a la loi $P(X = j) = p_j$. On mesure l'écart entre les fréquences empiriques $N_n(\mathbf{x}; i) = \#\{j : x_j = i\}$ et np_i au moyen de la statistique

$$T_n(\mathbf{x}) := \sum_{j=1}^{k} \frac{\Big(N_n(\mathbf{x}; j) - np_j\Big)^2}{n\, p_j}\,.$$

Pour employer la statistique T_n il faut connaître sa loi. Lorsque le nombre des observations est grand, à la place de calculer cette loi, on procède comme pour l'intervalle de confiance pour les v.a. de Bernoulli en utilisant le

Théorème 16.1 *Si les v.a. $X_1, \ldots, X_n$ sont i.i.d.,à valeur dans $\{1, \ldots, k\}$, et si $P(X_1 = i) = p_i$, $i = 1, \ldots, k$, alors la v.a.*

$$\sum_{j=1}^{k} \frac{\Big(N_n(\,\cdot\,; j) - np_j\Big)^2}{n\, p_j}$$

converge en loi vers une v.a. Z telle que $Z \sim \chi^2_{k-1}$.

Si n est grand on remplace la loi de T_n par celle du χ^2_{k-1} et on détermine une région critique (asymptotiquement de niveau $1 - \alpha$) par

$$\mathcal{R} = \{\mathbf{x} : T_n(\mathbf{x}) \geq \chi^2_{\alpha,k-1}\}\,.$$

Dans le cas présent, si $\alpha = 0{,}05$, $\chi^2_{0{,}05,\,5} = 11{,}07$.

On procède à 120 lancers du dé et on obtient l'échantillon

i	:	1	2	3	4	5	6
$N_{120}(\mathbf{x}; i)$	:	26	20	16	27	15	16

On ne rejette pas H_0 car la p-valeur du test vaut $P(T_n \geq 7{,}10) = 0{,}22$,

$$T_{120}(\mathbf{x}) = \frac{1}{20}\Big(26^2 + 20^2 + 16^2 + 27^2 + 15^2 + 16^2\Big) - 120 = 7{,}10\,.$$

□

Remarque 16.1 Dans l'expression de T_n on connaît la loi de $N_n(\cdot; j)$ sous l'hypothèse H_0. C'est une loi binomiale de paramètres $n = 120$ et $p = 1/6$. Par conséquent $\mathrm{Var} N_{120}(\cdot; j) = 100/6$ et

$$\mathbb{E}(T_{120}) = \sum_{j=1}^{6} \frac{1}{20} \mathrm{Var} N_{120}(\cdot; j) = 5\,.$$

Si la loi de T_{120} peut être approximée par celle du χ^2_5 alors la variance de T_{120} est approximativement le double de $\mathbb{E}(T_{120})$, soit 10, et l'écart-type est approximativement 3,16. Expérimentalement on a trouvé une valeur $7{,}10 \leq 5 + 3{,}16$, en bon accord avec l'hypothèse H_0. □

16.3 Exercices

Exercice 16.1 On lance un dé 1200 fois et on obtient

i	:	1	2	3	4	5	6
$N_{1200}(\mathbf{x}; i)$	:	175	215	220	190	170	230

Est-ce que le dé est équilibré ?

Exercice 16.2 On considère des v.a. i.i.d. X_i, $X_i \sim N(m,1)$ où la valeur de m est inconnue, $m = 0$ ou $m = 0{,}5$. On teste l'hypothèse $H_0 = \{m = 0\}$ à partir d'un échantillon de ces v.a. de taille n. Donner un test de niveau 0,95 et calculer la probabilité de faire une erreur de deuxième espèce pour ce test.

Exercice 16.3 On lance un dé à six faces 300 000 fois. On a obtenu 101 351 fois le chiffre 5 ou le chiffre 6.
a) Calculer la fréquence relative de cet événement.
b) Est-ce que le dé est équilibré ?
Indication : utiliser la v.a. de test Y_n du paragraphe 15.1.2 et estimer la p-valeur. Alternativement utiliser l'inégalité de Hoeffding. Comparer les deux méthodes.

Exercice 16.4 On reprend les données de l'exercice 14.4. Les v.a. Z_i qui décrivent les incertitudes lors des mesures sont i.i.d. et de loi $N(0, \sigma^2)$.
a) Construire une v.a. pivot pour le paramètre β.
Indication : à partir des résultats des propositions 14.2 et 14.3 on peut construire une v.a. pivot de loi t_{n-2}.
b) Tester l'hypothèse $\beta = 0$ contre $\beta > 0$. Choisir 0,95 comme niveau du test.

Solutions de quelques exercices

Exercice 3.9 Le nombre total de symboles est Nn. La question a) est équivalente à compter les rangements de r boules indistinguables dans Nn boîtes, au plus une boule par boîte. La probabilité qu'on ait r_1 boules dans les boîtes 1 à N, r_2 boules dans les boîtes $N+1$ à $2N$ etc est

$$\frac{\binom{N}{r_1}\cdots\binom{N}{r_n}}{\binom{Nn}{r}} = \frac{r!}{r_1!\cdots r_n!}\underbrace{\frac{(N!)^n(Nn-r)!}{(N-r_1)!\cdots(N-r_n)!(Nn)!}}_{=C(N)}\,.$$

Lorsque N diverge, en utilisant la formule de Stirling, on montre sans difficulté que la fraction $C(N)\to n^{-r}$.

Exercice 3.10 a) La probabilité qu'aucune boîte ne contienne plus d'une boule est

$$\frac{[M]_n}{M^n} = \prod_{j=1}^{n-1}\Big(1-\frac{j}{M}\Big) \le \exp\Big(-\sum_{j=1}^{n-1}\frac{j}{M}\Big) = \exp\Big(-\frac{n(n-1)}{2M}\Big)\,.$$

Une borne inférieure est donnée par

$$\exp\Big(-\sum_{j=1}^{n-1}\frac{j}{M}-\sum_{j=1}^{n-1}\frac{j^2}{M^2}\Big)\,.$$

En utilisant l'identité

$$1+2^2+3^2+\cdots+n^2 = \frac{n(n+1)(2n+1)}{6}\,,$$

$$\prod_{j=1}^{n-1}\Big(1-\frac{j}{M}\Big) \ge \exp\Big(-\frac{n(n-1)}{M}\Big[\frac{1}{2}-\frac{1}{6M}+\frac{n}{3M}\Big]\Big) \ge \exp\Big(-\frac{2n(n-1)}{3M}\Big)$$

si $2n\le M$. Si $c=\sqrt{\frac{3\ln 2}{2}}$, la probabilité qu'aucune boîte ne contienne plus d'une boule est au moins $1/2$.
b) On montre l'inégalité.

$$\begin{aligned}\exp(-x-x^2)\le 1-x &\iff x+x^2\ge \ln\frac{1}{1-x} = x+\frac{x^2}{2}+\frac{x^3}{3}+\cdots\\ &\iff \frac{x^2}{2}\ge x^2\Big(\frac{x}{3}+\frac{x^2}{4}+\cdots\Big)\,.\end{aligned}$$

Si $x \le \frac{1}{2}$ la parenthèse est inférieure à $\frac{1}{3}(\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots) = \frac{1}{3}$.

Exercice 5.2 Si n est impair le moment d'ordre n est nul. Si $n = 2m$, le moment d'ordre $2m$ est égal à

$$\sigma^{2m} = \#\text{appariements de } 2m \text{ objets} = 1 \cdot 3 \cdots (2m-1)\sigma^{2m}\,.$$

Exercice 5.7 Soit $\Omega = [0,x]^k$ et P la mesure de probabilité de densité

$$f(\mathbf{x}) := \Big(\int_0^x dt h(t)\Big)^{-k} \prod_{i=1}^k h(x_i)\,.$$

Soit $A_\pi := \{\mathbf{x}\colon x_{\pi(1)} \ge x_{\pi(2)} \ge \cdots \ge x_{\pi(k)}\}$ où π est une permutation de $1,\dots,k$. L'union des A_π est Ω, et par symétrie $P(A_\pi)$ ne dépend pas de π. Si la permutation π est l'identité, $P(A_\pi)$ s'écrit

$$\Big(\int_0^x dt h(t)\Big)^{-k} \cdot \int_0^x dt_1\, h(x_1) \int_0^{x_1} dx_2\, h(x_2) \cdots \int_0^{x_{k-1}} dx_k\, h(x_k) = \frac{1}{k!}\,.$$

Exercice 6.3 Pour une réalisation ω, l'application $j \mapsto S_n(\omega)$ est non décroissante. Si $T_r > n$, alors $S_n < r$, et si $S_n < r$, alors $T_r > n$. Ceci établit $\{T_r > n\} = \{S_n < r\}$ et donc $P(T_r > n) = P(S_n < r)$. La distribution de S_n est une distribution binomiale $\mathcal{B}_i(n,p)$. Calcul de $P(T_r = k)$;

$$\{\, T_r = k \,\} = \{\, S_{k-1} = r-1 \,\} \cap \{\, X_k = 1 \,\}\,.$$

Ces deux derniers événements sont indépendants ; (pour $k \ge r$)

$$P(T_r = k) = P(S_{k-1} = r-1)P(X_k = 1) = \binom{k-1}{r-1} p^r (1-p)^{k-r}\,.$$

$\{\, T_r < \infty \,\} = \bigcup_{n\ge 1}\{T_r \le n\}$; l'ensemble complémentaire est

$$\bigcap_{n\ge 1}\{T_r > n\} = \bigcap_{n\ge 1}\{S_n < r\}\,.$$

Comme $\lim_n P(S_n < r) = 0$, on a $1 - P(T_r < \infty) = 0$. En effet (si $q = 1-p$)

$$\lim_n \sum_{k\le r-1} \binom{n}{k} p^k q^{n-k} < \lim_n \sum_{k\le r-1} n^k q^{n-r+1} < \lim_{n\to\infty} r n^r q^{n-r} = 0\,.$$

Exercice 6.4 On suppose que c'est l'urne U_0 qu'on découvre vide ; il y a k boules dans l'urne U_1. On note cet événement par $E_0(k)$. Si $E_0(k)$ est vrai, on a choisi $N+1$ fois l'urne U_0 et on a fait $(N+1)+(N-k)$ tirages ; en utilisant les v.a. données en indication de l'exercice

$$T_{N+1} = 2N - k + 1\,.$$

Comme $p = 1/2$,

$$P(E_0(k)) = P(T_{N+1} = 2N - k + 1) = \binom{2N-k}{N} 2^{-(2N-k+1)}\,.$$

De façon similaire on définit $E_1(k)$. La probabilité cherchée est

$$P(E_0(k)) + P(E_1(k)) = \binom{2N-k}{N} 2^{-(2N-k)}\,.$$

Exercice 7.5 $\mathbb{E}(X) = \frac{x}{\lambda}$ et $\mathrm{Var}X = \frac{x}{\lambda^2}$. En effet, en utilisant le changement de variables $y = \lambda t$,

$$\frac{1}{\Gamma(x)} \int_0^\infty \lambda t \mathrm{e}^{-\lambda t} (\lambda t)^{x-1}\, dt = \frac{1}{\lambda\Gamma(x)} \int_0^\infty \mathrm{e}^{-y} (y)^x\, dy = \frac{\Gamma(x+1)}{\lambda\Gamma(x)} = \frac{x}{\lambda}\,;$$

$$\mathbb{E}(X^2) = \frac{1}{\Gamma(x)} \int_0^\infty \lambda t^2 \mathrm{e}^{-\lambda t} (\lambda t)^{x-1}\, dt = \frac{\Gamma(x+2)}{\lambda^2\Gamma(x)} = \frac{x(x+1)}{\lambda^2}\,.$$

$\{T_n \le x\}$ si et seulement si le nombre d'émissions jusqu'au temps x est au moins n, i.e. $\{T_n \le x\} = \{N(x) \ge n\}$. Par conséquent

$$P(T_n \le x) = \sum_{j=n}^{\infty} \frac{\mathrm{e}^{-\lambda x}(\lambda x)^j}{j!}\,.$$

En dérivant par rapport à x on obtient la densité de la loi de T_n :

$$(-\lambda) \sum_{j \ge n} \frac{\mathrm{e}^{-\lambda x}(\lambda x)^j}{j!} + \lambda \sum_{j \ge n} \frac{\mathrm{e}^{-\lambda x}(\lambda x)^{j-1}}{(j-1)!} = \lambda \frac{\mathrm{e}^{-\lambda x}(\lambda x)^{n-1}}{(n-1)!}\,.$$

La loi de T_n est une loi d'Erlang ou loi $\gamma_{n,\lambda}$.

Exercice 7.7 a) $P(X_1 = X_2) = \sum_{i=1}^n p_i^2$. b) Par l'inégalité de Cauchy-Schwarz

$$\sum_{i=1}^n p_i^2 \le 1 = \sum_{i=1}^n p_i \le \Big(\sum_{i=1}^n p_i^2\Big)^{1/2} \Big(\sum_{i=1}^n 1\Big)^{1/2}\,.$$

c) Si $p_i = 1/n$ pour tout i, $P(X_1 = X_2) = 1/n$ et si $p_1 = 1$, $p_j = 0$, $j \ne 1$, $P(X_1 = X_2) = 1$. Dans les autres cas les inégalités sont strictes.

Exercice 8.1 b) On pose $q_i = n_i/n$. Du point a)

$$\binom{n}{n_1\,, n_2\,, \cdots\,, n_p} q_1^{n_1} \cdots q_p^{n_p} \le \sum_{\substack{n_1,\ldots,n_p \ge 0:\\ n_1+\cdots+n_p = n}} \binom{n}{n_1\,, n_2\,, \cdots\,, n_p} q_1^{n_1} \cdots q_p^{n_p} = 1\,.$$

Par conséquent

$$\binom{n}{n_1\,,n_2\,,\cdots\,,n_p} \le \prod_{i=1}^{p} \Big(\frac{n_i}{n}\Big)^{-n_i} = \exp\Big(-n\sum_{i=1}^{p}\frac{n_i}{n}\ln\frac{n_i}{n}\Big)\,.$$

Exercice 8.8 Par l'inégalité de Cauchy-Schwarz

$$\mathbb{E}(XI_{\{X>a\}}) \le \big(\mathbb{E}(X^2)\big)^{1/2}\big(\mathbb{E}(I_{\{X>a\}})\big)^{1/2}\,.$$

Par ailleurs

$$\mathbb{E}(XI_{\{X>a\}}) = \mathbb{E}(X) - \mathbb{E}(XI_{\{X\le a\}}) \ge \mathbb{E}(X) - a\,.$$

De ces inégalités on obtient

$$P(X>a) = \mathbb{E}(I_{\{X>a\}}) \ge \frac{\big(\mathbb{E}(X)-a\big)^2}{\mathbb{E}(X^2)}\,.$$

Exercice 8.9 a) $P(A_1\cup A_2) = P(A_1)+P(A_2)-P(A_1\cap A_2)$. On procède par récurrence. On suppose l'inégalité établie pour n.

$$\begin{aligned} P(A_1\cup\cdots\cup A_{n+1}) &= P(A_1\cup\cdots\cup A_n) + P(A_{n+1}) \\ &\quad - P([A_1\cup\cdots\cup A_n]\cap A_{n+1}) \\ &\ge P(A_1\cup\cdots\cup A_n) + P(A_{n+1}) - \sum_{j=1}^{n} P(A_j\cap A_{n+1}) \\ &\ge \sum_j P(A_j) - \sum_{i<j} P(A_i\cap A_j)\,. \end{aligned}$$

b) D'une part

$$P(\max_i X_i > t) = P(\{X_1\ge t\}\cup\cdots\cup\{X_n\ge t\}) \le nP(X_1\ge t)\,.$$

D'autre part, l'indépendance des v.a. X_i et le point a) impliquent

$$P(\{X_1\ge t\}\cup\cdots\cup\{X_n\ge t\}) \ge nP(X_1\ge t) - \frac{n(n-1)}{2}\big(P(X_1\ge t)\big)^2\,.$$

Exercice 8.10 a) voir exemple 12.6. b)

$$\begin{aligned} P(Y_n\le n) &= P(X_1\le n)\cdots P(X_n\le n) = (1-P(X_1>n))^n \\ &= \Big(1-\frac{nP(X_1>n)}{n}\Big)^n \overset{n\to\infty}{\longrightarrow} 1\,. \end{aligned}$$

Si $\mathbb{E}(X_1)$ existe, le résultat découle du lemme I.1 et de $\lim_n nP(X_1>n)=0$.
c) Par définition

$$p = P(Y_n\le\lambda_p) = (1-P(X_1>\lambda_p)^n\,,$$

et donc

$$P(X_1 > \lambda_p) = 1 - \mathrm{e}^{\frac{1}{n}\ln p} = -\frac{1}{n}\ln p + O(n^{-2}) = \frac{\ln(1/p)}{n} + O(n^{-2})\,.$$

Asymptotiquement, pour n grand, $\lambda_p \sim \ln n$ si $X_1 \sim \gamma_{1,1}$ et $\lambda_p \sim \sqrt{2\ln n}$ si $X_1 \sim N(0,1)$. Pour une loi de Pareto de paramètre α,

$$\lambda_p \sim \Big(\frac{n}{\ln(1/p)}\Big)^{1/\alpha}\,.$$

Si $0 < \alpha \leq 1$, on observe un comportement totalement différent de celui d'une v.a. possédant une espérance.

Exercice 9.6 On note $a_1, \dots a_{r+1}$ la longueur des séquences de Faces, et $b_1, \dots, b_r$ la longueur des séquences de Piles. Par définition $a_1 \geq 0$, $a_{r+1} \geq 0$; dans les autres cas $a_i \geq 1$ et $b_j \geq 1$. Si la longueur des séquences de Faces est fixée, le nombre de configurations avec k Piles et r séquences de Piles est égal au nombre de placement de k boules indistinguables dans k boîtes, au moins une boule par boîte. Ce nombre est $\binom{k-1}{r-1}$ (voir exemple 3.6). Pour compter le nombre de séquences possibles de Faces on définit $\bar{a}_1 := a_1 + 1$, $\bar{a}_{r+1} := a_{r+1} + 1$ et $\bar{a}_j := a_j$ sinon. On a $\sum_i \bar{a}_i = n - k + 2$; le nombre de séquences possibles de Faces est $\binom{n-k+1}{r}$ puisque $\bar{a}_i \geq 1$. Le nombre de configurations avec k Piles est $\binom{n}{k}$. La probabilité cherchée est donc

$$P(E|k \text{ Piles et } k-n \text{ Faces}) = \frac{\binom{k-1}{r-1}\binom{n-k+1}{r}}{\binom{n}{k}}\,.$$

Exercice 9.9 Pour le point a) utiliser par exemple une analyse avec diagramme en arbre ou l'automate ci-dessous.
b) L'automate a six états, $1, \dots, 6$. L'état initial est 1. Si l'on obtient Pile, on retourne à l'état 1, sinon on passe à l'état 2. De l'état 2 on passe à l'état 3 si l'on obtient Face, sinon on passe à l'état 4. Lorsqu'on atteint l'état 3, B est sûr de gagner : ou bien on atteint l'état final 5 si l'on obtient Pile, ou l'on retourne dans l'état 3 si l'on obtient Face. Lorsqu'on est dans l'état 4, on atteint l'état final 6 si l'on obtient Pile et on *retourne à l'état* 2 si l'on obtient Face. Lorsqu'on est dans l'état 2, la probabilité de gagner pour B est $1/2$, et la probabilité de revenir dans cet état est $1/4$. Par conséquent

$$P(B \text{ gagne}) = \frac{1}{2}\Big(1 + \frac{1}{4} + \frac{1}{16} + \cdots\Big) = \frac{2}{3}\,.$$

Exercice 9.10 Voir solution exercice 10.5.

Exercice 10.1 On utilise l'inégalité de Markov. On pose $Y := |\frac{S_n}{n} - p|$.

$$P(Y \geq \varepsilon) = P(Y^4 \geq \varepsilon^4) \leq \frac{\mathbb{E}(Y^4)}{\varepsilon^4}\,.$$

$$\mathbb{E}(Y^4) = \frac{1}{n^4} \sum_{i,j,k,l=1}^{n} \mathbb{E}[(X_i - p)(X_j - p)(X_k - p)(X_l - p)] \,.$$

Les seuls termes non nuls sont du type $\mathbb{E}[(X_i - p)^2(X_j - p)^2] = p^2(1-p)^2$ et $\mathbb{E}[(X_i - p)^4] = p(1-p)(p^3 + (1-p)^3)$. Il y a $3n(n-1)$ termes du premier type et n du second type. Donc

$$\mathbb{E}(Y^4) = \frac{p(1-p)}{n^4}[n(p^3 + (1-p)^3) + 3p(1-p)(n^2 - n)] \leq \frac{1}{4n^2} \,.$$

La convergence de la série $\sum_n \frac{1}{n^2}$ entraîne la loi forte des grands nombres.

Exercice 10.3 On définit récursivement, à partir des intervalles $I_0 := [0, 2^{-1})$ et $I_1 := [2^{-1}, 1)$, les intervalles I_{00} et I_{01} en divisant par deux I_0, et les intervalles I_{10} et I_{11} en divisant par deux I_1 ; on obtient $I_{00} := [0, 2^{-2})$, $I_{01} := [2^{-2}, 2^{-1})$, $I_{10} := [2^{-1}, 2^{-1} + 2^{-2})$, $I_{11} := [2^{-1} + 2^{-2}, 1)$. On répète cette opération à partir des nouveaux intervalles. Si, par convention, on exclut les développements en base 2, qui se terminent uniquement par des 1, alors

$$I_{\omega_0\omega_1\cdots\omega_k} = \{\omega\colon \text{le développement en base 2 de } \omega \text{ commence par } \omega_0\omega_1\cdots\omega_k\} \,;$$

$X_n(\omega) = \omega_n$ donne le $(n+1)^{\text{ième}}$ chiffre du développement en base 2 de ω. Il découle de ceci que X_n est une v.a. de Bernoulli de paramètre $1/2$, et que

$$P(X = \omega_0 \cdots X_k = \omega_k) = 2^{-k-1} = P(X_0 = \omega_0) \cdots P(X_k = \omega_k)$$

pour tout k. Soit $a_0 \cdots a_{r-1}$ un motif. On définit les v.a. Z_k, $k \geq 0$, par

$$Z_k(\omega) := \begin{cases} 1 & \text{si } a_0 \cdots a_{r-1} = \omega_k \cdots \omega_{k+r-1} \\ 0 & \text{sinon.} \end{cases}$$

On peut décomposer $N_n(\omega; m_r)$ en r sommes de v.a. i.i.d.

$$\frac{N_n(\omega; m_r)}{n} = \frac{1}{r} \sum_{p=0}^{r-1} \frac{r}{n} \sum_{\substack{j \geq 0 : \\ jr+p \leq n}} Z_{jr+p} \,.$$

On applique la loi forte des grands nombres pour chacune des r sommes.
Si $k = 5$, les probabilités conditionnelles sont nulles, sauf si $a_5 = b_0, \ldots, a_9 = b_4$. Si cette condition est vérifiée, les probabilités conditionnelles sont égales à 2^{-5}. L'état du système dynamique après 5 unités de temps est dans $I_{a_5\cdots a_9}$. Lorsque $k = 10$ les probabilités conditionnelles sont toujours égales à 2^{-10}. Cela signifie que la connaissance partielle de la condition initiale ω (connaissance des dix premiers chiffres du développement binaire de ω) ne permet *aucune* prédiction sur l'état du système après 10 unités de temps.

Exercice 10.4 a) Soit $a < x_0 < b$ et F la fonction de répartition de la loi de densité f. On a $F_Z = F$ car

$$P(Z \leq x_0) = P(X \leq x_0 | Y \leq f(X)) = \frac{P(X \leq x_0, Y \leq f(X))}{P(Y \leq f(X))} \,.$$

$$P(X \le x_0, Y \le f(X)) = \frac{1}{M(b-a)} \int_a^{x_0} f(u)\, du = \frac{F(x_0)}{M(b-a)}\,;$$

$$P(Y \le f(X)) = \frac{1}{M(b-a)} \int_a^{b} f(u)\, du = \frac{1}{M(b-a)}\,.$$

b) Soit D la région du carré $[0,1]^2$ située sous le graphe de g. On considère l'espace de probabilité $\Omega = [0,1]^2$ avec la mesure de probabilité uniforme et la v.a. $Y = I_D$. Soit $Y_1, \ldots, Y_n$ des v.a. i.i.d., $Y_i \stackrel{\mathcal{L}}{=} Y$. Par la LGN

$$\forall\, \varepsilon > 0 : \quad \lim_{n\to\infty} P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} Y_i - \int_0^1 g(u)\, du\Big| \ge \varepsilon\Big) = 0\,.$$

Exercices 9.10 et 10.5 Si le processus ne s'arrête pas après $i-1$ tirages, l'urne contient après ces tirages b boules noires et $a+i-1$ boules blanches. La probabilité conditionnelle cherchée est donc

$$\frac{a+i-1}{a+b+i-1}\,.$$

$$P(N > k) = \prod_{i=1}^{k} \frac{a+i-1}{a+b+i-1} = \frac{(a+k-1)!(a+b-1)!}{(a-1)!(a+b+k-1)!}\,.$$

Si $b > 1$, ce produit tend vers 0 au moins aussi vite que k^{-2}. Si $b = 1$, $\lim_k kP(N > k) = a$. $\mathbb{E}(N) = \sum_{k\ge 0} P(N > k) < \infty$ si et seulement si $b > 1$. Dans ce cas, on peut écrire

$$\frac{(a+k-1)!}{(a+b+k-1)!} = \frac{1}{b-1}\Big(\frac{(a+k-1)!}{(a+b+k-2)!} - \frac{(a+k)!}{(a+b+k-1)!}\Big)\,.$$

On en déduit

$$\mathbb{E}(N) = \sum_{k\ge 0} P(N > k) = \frac{a+b-1}{b-1}\,.$$

On introduit des v.a. tronquées N'_j et on pose $S_n = \sum_{i=1}^n N_i$, $S'_n = \sum_{i=1}^n N'_i$.

$$P(S_n \ne S'_n) \le nP(N_1 \ne N'_1)) = nP(N_1 > M) = n\frac{a}{a+M} \to 0$$

si $M = \lfloor an \ln n \rfloor$. Si $b = 1$,

$$P(N = k) = P(N > k-1) - P(N > k) = \frac{a}{(a+k)(a+k-1)}\,,$$

$$\mathbb{E}(N'_1) = \sum_{k=1}^{M} kP(N = k) = \sum_{j=1}^{M-1} \frac{a}{a+j} - \frac{Ma}{a+M} = a \ln \frac{M}{a} + O(1)\,,$$

$$\mathrm{Var}(N'_1) = \sum_{k=1}^{M} \frac{ak^2}{(a+k)(a+k-1)} \le aM\,.$$

On prend n suffisamment grand de sorte que

$$\Big|\frac{\mathbb{E}(S'_n)}{n\ln n} - a\Big| \le \frac{\varepsilon}{2}\,.$$

En utilisant le lemme 10.1

$$\begin{aligned} P\Big(\Big|\frac{S_n}{n\ln n} - a\Big| \ge \varepsilon\Big) &\le P\Big(\Big|\frac{S'_n - \mathbb{E}(S'_n)}{n\ln n}\Big| \ge \frac{\varepsilon}{2}\Big) + nP(N_1 \ne N'_1) \\ &\le \frac{4n\mathrm{Var}(N'_1)}{n^2(\ln n)^2\varepsilon^2} + \frac{1}{\ln n} \le \frac{4a^2}{\varepsilon^2\ln n} + \frac{1}{\ln n} \overset{n\to\infty}{\longrightarrow} 0\,. \end{aligned}$$

Exercice 11.3 a) Par le principe du miroir, un chemin qui arrive en k et passe par $j \ge r$ doit couper ou toucher la ligne horizontale à hauteur r. L'ensemble de ces chemins est en correspondance biunivoque avec les chemins qui partent de l'origine et arrivent en $k + 2(r-k) = 2r - k$;

$$P_n(\max \ge r) \equiv P(\max_{\ell=1}^{n} S_\ell \ge r) = P(S_n = 2r - k)\,.$$

b) La probabilité cherchée est donné par

$$P_n(\max \ge r) - P_n(\max \ge r+1) = P(S_n = 2r-k) - P(S_n = 2r-k+2)\,.$$

c) La probablité cherchée est donné par la somme sur $k = r, r-1, \ldots$ du résultat précédent,

$$\sum_{j\ge 0} \big(P(S_n = r+j)) - P(S_n = r+j+2)\big) = P(S_n = r) + P(S_n = r+1)\,.$$

Un des termes est nul, car $P(S_n = k) = 0$ si n et k n'ont pas la même parité.

Exercice 12.2 N^+ est une somme de $N = 10^6$ v.a. de Bernoulli d'espérance $1/2$ et de variance $1/4$. Le TLC indique que la somme N^+ appartient avec probabilité $\approx 0{,}95$ à l'intervalle

$$\mathbb{E}(N^+) \pm 2\sqrt{\mathrm{Var}(N^+)} = 5\cdot 10^5 \pm 10^3\,.$$

$\mathrm{Prob}(E_2) > \mathrm{Prob}(E_3) > \mathrm{Prob}(E_1)$. Pour estimer $\mathrm{Prob}(E_1)$, utiliser l'inégalité de Chebyshev ou celle de Hoeffding. Pour $\mathrm{Prob}(E_3)$, faire un calcul en utilisant la formule de Stirling.

Exercice 12.4 L'erreur maximale est $N\varepsilon$. Si les erreurs sont uniformément distribuées sur $[-\varepsilon, \varepsilon]$, $\mathbb{E}(\varepsilon_n) = 0$ et $\mathrm{Var}(\varepsilon_n) = \varepsilon^2/3$. Lorsque N est grand, par le TLC on s'attend à des erreurs de l'ordre $O(\sqrt{N}\varepsilon)$. La probabilité que l'erreur sur la somme dépasse $\sqrt{N}\varepsilon$ est égale à

$$P(|Z| \ge \sqrt{3}) = 0{,}084 \quad \text{où } Z \text{ est } N(0,1).$$

Exercice 12.5 En remplaçant X_k par $X_k - \mu$ et Z_n par $Z_n - a$, il suffit de montrer l'affirmation pour $\mu = 0$ et $a = 0$. Soit $Y_n := \dfrac{1}{\sigma\sqrt{n}} \displaystyle\sum_{k=1}^{n} X_k$. Par

hypothèse $\lim_n P(Y_n \le x) = \Phi(x)$, et pour tout $\varepsilon > 0$ et $\delta > 0$ il existe $N(\varepsilon, \delta)$ tel que pour tout $n \ge N(\varepsilon, \delta)$, $P\big(|Z_n| \ge \delta\big) \le \varepsilon$. Si n est suffisamment grand,

$$P\big(Y_n + Z_n \le x\big) \le P\big(\{Y_n + Z_n \le x\} \cdot \{|Z_n| < \delta\}\big) + \varepsilon \le P\big(Y_n \le x + \delta\big) + \varepsilon\,.$$

$$P\big(Y_n \le x - \delta\big) - \varepsilon \le P\big(\{Y_n \le x - \delta\} \cdot \{|Z_n| < \delta\}\big) \le P\big(Y_n + Z_n \le x\big)\,.$$

On obtient le résultat en prenant la limite $n \to \infty$, puis la limite $\varepsilon \to 0$ et finalement $\delta \to 0$ (utiliser la continuité de Φ).

Exercice 13.3 On introduit les déviations $Y_i := X_i - \mu$.

$$(n-1)S_n^2 = \sum_{j=1}^{n} Y_j^2 - n\overline{Y}_n^2\,.$$

Les v.a. Y_i sont i.i.d. ; par l'inégalité de Chebyshev $\overline{Y}_n^2 \xrightarrow{\mathcal{L}} 0$. Les v.a. Y_i^2 sont i.i.d. de moyenne σ^2, et par la LGN $\dfrac{1}{n}\sum_i Y_i^2 \xrightarrow{\mathcal{L}} \sigma^2$. Pour des v.a. U_n et V_n telles que $U_n \xrightarrow{\mathcal{L}} u$ et $V_n \xrightarrow{\mathcal{L}} v$, u et v des constantes,

$$P(|U_n + V_n - u - v| \ge 2\delta) \le P(|U_n - u| \ge \delta) + P(|V_n - v| \ge \delta) \overset{n\to\infty}{\longrightarrow} 0\,.$$

Exercice 13.4 La fonction de vraisemblance pour $\mathbf{x}$ donné s'écrit

$$\theta \mapsto L_{\mathbf{x}}(\theta) = \theta^{-n} \prod_{i=1}^{n} I_{[0,\theta]}(x_i)\,.$$

Elle est donc nulle tant que $\theta < \max_i x_i$; puis elle vaut θ^{-n}. L'estimateur de maximum de vraisemblance est $\max_i X_i$. Pour la suite voir exemple 12.6.

Exercice 14.5 a) En utilisant l'indépendance de X et Y, $\mathbb{E}(Z) = a\mathbb{E}(X)\mathbb{E}(Y)$,

$$\mathrm{Var}Z = a^2\big[\mathbb{E}(Y)\mathrm{Var}X + \mathbb{E}(X)\mathrm{Var}Y + \mathrm{Var}X\,\mathrm{Var}Y\big]\,.$$

Comme estimateur non biaisé de Z on peut prendre $a\overline{X}_n\overline{Y}_n$.
b) Si les v.a. X_i ont une espérance x, par la LGN les v.a. $\overline{X}_n$ convergent en loi vers la v.a. constante x. Par une variante du calcul qui montre que $\lim_n \mathbb{E}(g(\overline{X}_n)) = g(x)$, les v.a. $g(\overline{X}_n)$, $n \ge 1$, forment une suite consistante :

$$\forall\,\varepsilon\ \exists\,\delta > 0 \ \text{ tel que } \ \big(|t - x| \le \delta \implies |g(t) - g(x)| \le \varepsilon\big)\,,$$

$$\implies P(|g(\overline{X}_n) - g(x)| \ge \varepsilon) \le P(|\overline{X}_n - x| \ge \delta) \overset{n\to\infty}{\longrightarrow} 0\,.$$

Exercice 15.3 L'estimateur de maximum de vraisemblance est $\hat\theta = \max_i X_i$. Comme on a toujours $\hat\theta \le \theta$, un choix naturel pour $A(\theta)$ est

$$A(\theta) := \{x_\alpha(\theta) \le \hat\theta \le \theta\} \quad \text{avec} \quad P_\theta(A(\theta)) = 1 - \alpha\,.$$

On a $x_\alpha(\theta) = \alpha^{1/n}\theta$ car

$$1-\alpha = P_\theta(A(\theta)) = 1 - P_\theta(\hat\theta \le x_\alpha(\theta)) = 1 - x_\alpha(\theta)^n \theta^{-n}\,.$$

L'intervalle de confiance $I(\mathbf{x})$ est donné par

$$I(\mathbf{x}) := \{\theta : A_\theta \text{ est réalisé pour } \mathbf{x}\} = \{\theta : \hat\theta(\mathbf{x}) \le \theta \le \alpha^{-1/n}\,\hat\theta(\mathbf{x})\}\,.$$

Exercice 15.4 L'estimateur est $T := \frac{2}{n}\sum_i X_i$ tel que

$$\mathbb{E}_\theta(T) = \theta \quad \text{et} \quad \mathrm{Var}_\theta(T) = \theta^2/3n\,.$$

Pour n grand, le TLC implique que la loi de

$$Y := (T - \mathbb{E}_\theta(T))/\sqrt{\mathrm{Var}_\theta(T)}$$

est proche de la loi $N(0,1)$. Asymptotiquement, si $n \to \infty$, $P(|Y| \le 1{,}96) = 0{,}95$. On choisit de faire cette approximation et on prend

$$A(\theta) := \{|(T - \mathbb{E}_\theta(T))/\sqrt{\mathrm{Var}_\theta(T)}| \le 1{,}96\}\,.$$

Pour un échantillon $\mathbf{x}$, l'intervalle de confiance est donné par

$$I(\mathbf{x}) := \{\theta\text{: l'événement } A(\theta) \text{ est réalisé pour } \underline{x}\}\,.$$

$A(\theta)$ est réalisé si et seulement si

$$\sqrt{3n}\Big|\frac{T-\theta}{\theta}\Big| \le 1{,}96 \Longleftrightarrow \Big(1 - \frac{1{,}96}{\sqrt{3n}}\Big)\theta \le T \le \Big(1 + \frac{1{,}96}{\sqrt{3n}}\Big)\theta\,.$$

On obtient pour l'intervalle de confiance

$$I(\mathbf{x}) = \Big[\Big(1 + \frac{1{,}96}{\sqrt{3n}}\Big)^{-1} T(\mathbf{x})\,, \Big(1 - \frac{1{,}96}{\sqrt{3n}}\Big)^{-1} T(\mathbf{x})\Big]\,.$$

Exercice 15.5 La v.a. statistique $\hat\alpha$ est gaussienne et non biaisée. La variance est donnée dans la proposition 14.2. L'estimateur de la variance est donné par la v.a. S^2 de la proposition 14.3. On a donc

$$\frac{(\hat\alpha - \alpha)\sqrt{nS_{tt}}}{\sigma\sqrt{\sum_j t_j^2}} \sim N(0{,}1) \quad \text{et} \quad \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}\,.$$

Par la définition 13.5, la v.a.

$$Y := \frac{(\hat\alpha - \alpha)\sqrt{nS_{tt}}}{S\sqrt{\sum_j t_j^2}} \sim t_{n-2}\,.$$

La construction de l'intervalle de confiance est standard. Avec les données numériques $\alpha \in [1{,}57 - 1{,}80,\ 1{,}57 + 1{,}80]$.

Exercice 16.4 On procède comme pour l'exercice 15.5. On utilise la v.a. pivot

$$Z = \frac{(\hat\beta - \beta)\sqrt{S_{tt}}}{S} \sim t_{n-2}\,.$$

Sous l'hypothèse H_0 ($\beta = 0$) on rejette celle-ci dès que $Z > 2{,}01$ si le niveau est 0,95. Avec les données numériques $Z = 11{,}9$. On rejette donc H_0.

Index

Bibliographie

[Ar] S. F. Arnold, *Mathematical Statistics*, Prentice-Hall International (1990)

[Be] C. Berge, *Principes de combinatoire*, Dunod Paris (1968)

[Bi] P. Billingsley, *Probability and measure*, Wiley, New-York (1995)

[Br] L. Breiman, *Probability*, Addison-Wesley (1968)

[Ch] K.L. Chung, *Elementary Probability Theory with Stochastic Processes*, Undergraduate Texts in Mathematics, Springer, Berlin (1974)

[Fe1] W. Feller, *Introduction to Probability Theory and its Applications*, vol. I, Wiley, New-York (1968)

[Fe2] W. Feller, Introduction to Probability Theory and its Applications, vol. II, Wiley, New-York (1968)

[MiUp] M. Mitzenmacher, E. Upfal, *Probability and Computing*, Cambridge University Press, cambrideg (2005)

[Mo] S. Morgenthaler, *Introduction à la statistique*, 3^{e} édition, Presses polytechniques et universitaires romandes, Lausanne (1997)

[Pe] R. W. Pestman, *Mathematical Statistics*, 2nd edition de Gruyter, Berlin, New-York (2009)

[Si] Y. G. Sinai, *Probability Theory, An introductory course*, Springer Textbook Berlin, Heidelberg (1992).