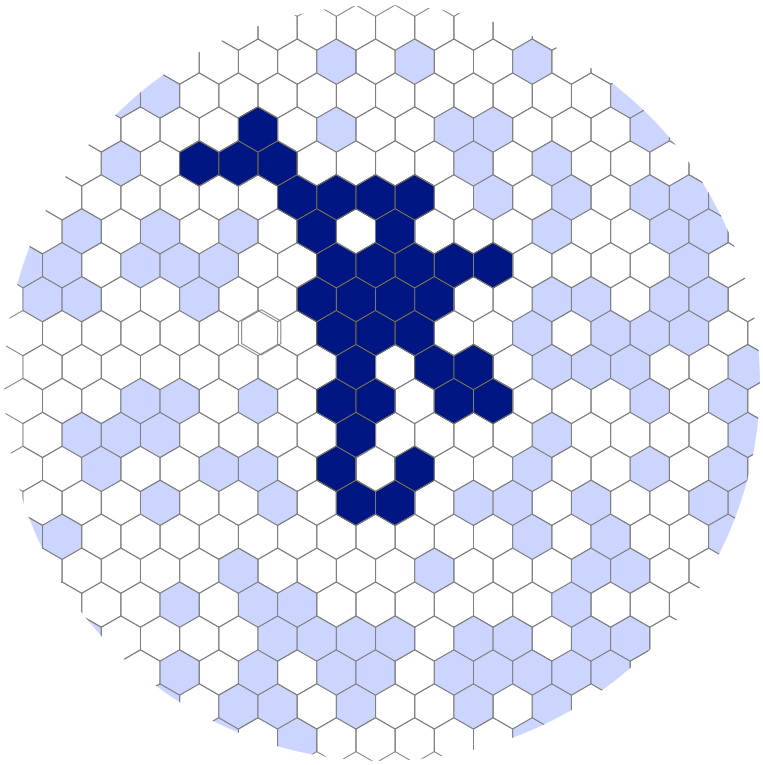
Probabilités et Statistique

— Yvan Velenik —



UNIVERSITÉ
DE GENÈVE

— Version du 4 mars 2026 —

Dernière version téléchargeable à l'adresse

www.unige.ch/math/folks/velenik/cours.html

Table des matières

0 Introduction

Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means.

Bertrand Russell

Misunderstanding of probability may be the greatest of all impediments to scientific literacy.

Stephen Jay Gould

Si la théorie des probabilités a été originellement motivée par l'analyse des jeux de hasard, elle occupe aujourd'hui une place centrale dans la plupart des sciences. Tout d'abord, de par ses applications pratiques : en tant que base des statistiques, elle permet l'analyse des données recueillies lors d'une expérience, lors d'un sondage, etc. ; elle a également conduit au développement de puissants algorithmes stochastiques pour résoudre des problèmes inabordables par une approche déterministe ; elle possède en outre de nombreuses applications directes : fiabilité, assurances, finance, moteurs de recherche, *deep learning*, etc. D'un côté plus théorique, elle permet la modélisation de nombreux phénomènes, aussi bien en sciences naturelles (physique, chimie, biologie, etc.) qu'en sciences humaines (économie, sociologie, par exemple) et dans d'autres disciplines (médecine, climatologie, informatique, réseaux de communication, traitement du signal, etc.). Elle s'est même révélée utile dans de nombreux domaines de mathématiques pures (algèbre, théorie des nombres, combinatoire, etc.) et appliquées (EDP, par exemple). Finalement, son intérêt intrinsèque lui assure une place de choix en mathématiques, et sa versatilité lui confère un des spectres les plus larges, allant des problèmes les plus appliqués aux questions les plus abstraites.

Le concept de probabilité est aujourd'hui familier à tout un chacun. Nous sommes constamment confrontés à des événements dépendant d'un grand nombre de facteurs hors de notre contrôle ; puisqu'il nous est impossible dans ces conditions de prédire exactement quel en sera le résultat, on parle de phénomènes aléatoires. Ceci ne signifie pas nécessairement qu'il y ait quelque chose d'intrinsèquement aléatoire à l'œuvre, mais simplement que l'information à notre disposition n'est que partielle. Quelques exemples : le résultat d'un jeu de hasard (pile ou face, jet de dé, roulette, loterie, etc.) ; la durée de vie d'un atome radioactif, d'un individu ou d'une ampoule électrique ; le nombre de gauchers dans un échantillon de personnes tirées au hasard ; le bruit dans un système de communication ; la fréquence d'accidents de la route ; le nombre de connexions à un serveur informatique ; le nombre d'étoiles doubles dans une région du ciel ; la position d'un grain de pollen en suspension dans l'eau ; l'évolution du cours de la bourse ; etc.

Le développement d'une théorie mathématique permettant de modéliser de tels phénomènes aléatoires a occupé les scientifiques depuis plusieurs siècles. Motivé initialement par l'étude des jeux de hasard, puis par des problèmes d'assurances, le domaine d'application de la théorie s'est ensuite immensément élargi. La plus ancienne trace écrite contenant les détails d'un calcul de probabilités est le poème *De Vetula*, généralement

attribué à Richard de Fournival[1] et datant du milieu du XIII^{e} siècle. Les premières publications consacrées à ce sujet remontent à G. Cardano[2] avec son livre *Liber De Ludo Aleæ* (publié en 1663, mais probablement achevé près d'un siècle plus tôt), ainsi qu'à Kepler[3] et Galilée[4]. Toutefois, il est généralement admis que la théorie mathématique des probabilités débute réellement en 1654 avec la correspondance entre Pascal[5] et Fermat[6], ainsi que l'adresse du premier à l'Académie Parisienne lors de laquelle il introduit ce qu'il appelle alors « la géométrie du hasard ». La théorie fut ensuite développée par de nombreuses personnes, dont Huygens[7], J. Bernoulli[8], de Moivre[9], D. Bernoulli[10], Euler[11], Gauss[12] et Laplace[13]. La théorie moderne des probabilités est fondée sur l'approche axiomatique de Kolmogorov[14], basée sur la théorie de la mesure de Borel[15] et Lebesgue[16]. Grâce à cette approche, la théorie a alors connu un développement très rapide tout au long du XX^{e} siècle.

0.1 Modélisation des phénomènes aléatoires

Le but de la théorie des probabilités est de fournir un modèle mathématique pour décrire les phénomènes aléatoires. Sous sa forme moderne, la formulation de cette théorie contient trois ingrédients : l'univers, les événements et la mesure de probabilité.

0.1.1 Univers.

Il s'agit d'un ensemble, noté habituellement Ω, dont les éléments correspondent à tous les résultats possibles de l'expérience aléatoire que l'on cherche à modéliser. On l'appelle également l'**espace des observables**, ou encore l'**espace échantillon**.

Exemple 0.1.

1. Un tirage à pile ou face : $\Omega = \{\mathsf{P}, \mathsf{F}\}$.
2. Deux tirages à pile ou face : $\Omega = \{\mathsf{PP}, \mathsf{PF}, \mathsf{FP}, \mathsf{FF}\}$.
3. Une suite de tirages à pile ou face se terminant à la première apparition d'un pile : $\Omega = \{\mathsf{P}, \mathsf{FP}, \mathsf{FFP}, \mathsf{FFFP}, \ldots\}$.
4. Une suite de lancers d'un dé : $\Omega = \{(a_k)_{k\geqslant 1} \,|\, a_k \in \{1, \ldots, 6\}, \forall k \geqslant 1\}$.
5. Taille d'une personne : $\Omega = \mathbb{R}^+$.

1. Richard de Fournival (1201, Amiens – 1260, Amiens ?), médecin, alchimiste, poète et érudit français.
2. Girolamo Cardano (1501, Pavie – 1576, Rome), parfois connu sous le nom de Jérôme Cardan, mathématicien, philosophe et médecin italien. Féru d'astrologie, certains ont affirmé qu'il avait prévu le jour de sa mort et qu'il avait cessé de s'alimenter avant cette date afin de rendre sa prédiction correcte.
3. Johannes Kepler (1571, Weil der Stadt – 1630, Ratisbonne), mathématicien, astronome et astrologue allemand.
4. Galilée ou Galileo Galilei (1564, Pise – 1642, Arcetri), physicien et astronome italien.
5. Blaise Pascal (1623, Clermont – 1662, Paris), mathématicien, physicien, philosophe, moraliste et théologien français. Auteur de nombreuses contributions majeures en mathématiques et en physique, il délaisse malheureusement ces dernières à la fin de 1654, à la suite d'une expérience mystique, et se consacre presque exclusivement à la réflexion philosophique et religieuse.
6. Pierre de Fermat (1601, Beaumont-de-Lomagne – 1665, Castres), juriste et mathématicien français.
7. Christiaan Huygens (1629, La Haye – 1695, La Haye), mathématicien, astronome et physicien néerlandais.
8. Jacques ou Jakob Bernoulli (1654, Bâle – 1705, Bâle), mathématicien et physicien suisse. Son ouvrage *Ars Conjectandi* contient la première preuve d'une version de la loi des grands nombres ; il lui a fallu 20 ans pour parvenir à une preuve qu'il juge satisfaisante.
9. Abraham de Moivre (1667, Vitry-le-François – 1754, Londres), mathématicien français.
10. Daniel Bernoulli (1700, Groningue – 1782, Bâle), médecin, physicien et mathématicien suisse.
11. Leonhard Euler (1707, Bâle – 1783, Saint-Pétersbourg), mathématicien et physicien suisse. Il est considéré comme le mathématicien le plus prolifique de tous les temps. Complètement aveugle pendant les dix-sept dernières années de sa vie, il produit presque la moitié de la totalité de son travail durant cette période.
12. Johann Carl Friedrich Gauss (1777, Brunswick – 1855, Göttingen), mathématicien, astronome et physicien allemand.
13. Pierre-Simon Laplace (1749, Beaumont-en-Auge – 1827, Paris), mathématicien, astronome et physicien français. Son ouvrage *Théorie Analytique des Probabilités*, publié en 1812 et considéré comme le premier livre fondamental sur la théorie des probabilités, a eu une influence majeure sur le développement de cette discipline.
14. Andreï Nikolaïevich Kolmogorov (1903, Tambov – 1987, Moscou), mathématicien russe.
15. Félix Édouard Justin Émile Borel (1871, Saint-Affrique – 1956, Paris), mathématicien et homme politique français.
16. Henri Léon Lebesgue (1875, Beauvais – 1941, Paris), mathématicien français.

6. Durée de vie d'une ampoule : $\Omega = \mathbb{R}^+$.
7. L'évolution du cours d'une action sur un intervalle de temps $[s,t]$: $\Omega = \mathcal{C}([s,t], \mathbb{R}^+)$, où l'on a noté $\mathcal{C}(A,B)$ l'ensemble des fonctions continues de A vers B.
8. La trajectoire d'un grain de pollen en suspension dans un fluide : $\Omega = \mathcal{C}(\mathbb{R}^+, \mathbb{R}^3)$. ⌟

Dans chaque cas, il ne s'agit que d'une modélisation de l'expérience correspondante. Il y a donc évidemment de nombreuses façons de choisir et d'encoder les différents résultats possibles d'une expérience aléatoire dans un ensemble Ω : dans le cinquième exemple, on pourrait limiter Ω à $[0,3]$ (mètres), voire à $\{1,2,\dots,3000\}$ (millimètres), sans perte de généralité ; dans le septième, on pourrait autoriser des discontinuités, etc.

0.1.2 Événements

Un événement est une propriété dont on peut dire si elle est vérifiée ou non une fois le résultat de l'expérience connu. Mathématiquement, un événement est identifié avec l'ensemble des résultats dans lesquels il est réalisé (un tel résultat est alors appelé une **réalisation** de l'événement).

Exemple 0.2. On lance deux fois un dé, $\Omega = \{(m,n) \in \{1,2,3,4,5,6\}^2\}$.

1. L'événement « le second lancer est un 6 » :
$$\{(m,6) \,|\, m \in \{1,2,3,4,5,6\}\}.$$
2. L'événement « le premier lancer est supérieur au second » :
$$\{(m,n) \in \Omega \,|\, m > n\}.$$
3. L'événement « la somme des deux lancers est paire » :
$$\{(m,n) \in \Omega \,|\, m+n \in 2\mathbb{N}\}.$$ ⌟

Exemple 0.3. On effectue une suite de lancers d'un dé :
$$\Omega = \{(a_k)_{k\geqslant 1} \,|\, a_k \in \{1,\dots,6\}, \forall k \geqslant 1\}.$$
L'événement « le 6 est sorti avant le 1 » correspond à
$$\big\{(a_k)_{k\geqslant 1} \in \Omega \,\big|\, \inf\{n \geqslant 1 \,|\, a_n = 6\} < \inf\{n \geqslant 1 \,|\, a_n = 1\}\big\}.$$
(Dans ce cas, il faudrait dire également comment interpréter les minima ci-dessus lorsque le 6 ou le 1 ne sortent jamais ; la convention usuelle est de poser $\inf \varnothing = +\infty$.) ⌟

Introduisons un peu de terminologie.

Définition 0.4. *Un singleton (c'est-à-dire un événement réduit à un unique élément de Ω) est appelé **événement élémentaire**. Sinon on parle d'**événement composite**. On appelle Ω l'**événement certain** et $\varnothing$ l'**événement impossible**. Si A est un événement, on appelle A^c l'**événement contraire** de A. Si A, B sont deux événements, on appelle $A \cap B$ l'**événement « A et B »**, et $A \cup B$ l'**événement « A ou B »**. Finalement, si $A \cap B = \varnothing$, A et B sont dits **disjoints**, ou **incompatibles**.*

0.1.3 Mesure de probabilité

Étant en possession d'une notion d'événements, on cherche ensuite à attribuer à chacun d'entre eux une probabilité, qui représente le degré de confiance que l'on a en sa réalisation. Les probabilités sont encodées sous forme de nombres réels compris dans l'intervalle $[0,1]$, avec l'interprétation que plus la probabilité est proche de 1, plus notre confiance dans la réalisation de l'événement est grande. Un événement de probabilité 1 est dit **presque certain** ou **presque sûr**.

Remarque 0.5. *Il est important de ne pas confondre un événement de probabilité* 1 *avec un événement certain, ou un événement de probabilité nulle avec un événement impossible. Par exemple, considérons l'expérience aléatoire consistant à tirer au hasard un nombre réel dans l'intervalle* $[0,1]$ *de façon uniforme (c'est-à-dire sans privilégier aucun de ces nombres). Alors, quel que soit* $x \in [0,1]$*, l'événement « le nombre tiré est* x *» a probabilité nulle (il doit avoir la même probabilité que chacun des autres nombres de l'intervalle, et il y en a une infinité). Or, à chaque tirage, un événement de ce type est réalisé ! En ce sens, un événement de probabilité nulle est un événement dont il est impossible de prédire qu'il aura lieu et non pas un événement ne pouvant pas se produire.*

D'un point de vue philosophique (épistémologique et ontologique), la notion de probabilité se révèle très subtile et est encore largement discutée aujourd'hui. Bien qu'intéressants, ces débats n'ont pas réellement d'impact sur la pratique du probabiliste (mais un peu plus sur celle du statisticien). Pour cette raison, nous ne nous intéresserons pas, dans ce cours, aux différentes interprétations existantes de la notion de probabilité. Nous nous contenterons d'en mentionner une, utile pour motiver certaines contraintes que nous imposerons à notre modèle plus tard : l'**approche fréquentiste**. Dans cette approche, on n'accepte d'associer de probabilité qu'à des événements reproductibles, c'est-à-dire correspondant à des expériences pouvant être répétées un nombre arbitraire de fois dans les mêmes conditions. On identifie alors la probabilité d'un événement avec la fréquence asymptotique à laquelle il est réalisé lorsque l'expérience est reproduite infiniment souvent. Cette notion a l'avantage d'être très intuitive et de donner, en principe, un algorithme permettant de déterminer empiriquement avec une précision arbitraire la probabilité d'un événement. Elle souffre cependant de plusieurs défauts : d'une part, une analyse un peu plus approfondie montre qu'il est fort difficile (si tant est que ce soit possible) d'éviter que cette définition ne soit circulaire, et d'autre part, elle est beaucoup trop restrictive, ne permettant pas, par exemple, de donner de sens à une affirmation du type « il y a 50% de chance pour que la Californie soit touchée par un séisme de magnitude supérieure à 7 sur l'échelle de Richter dans les 30 prochaines années ». Dans de telles affirmations, l'événement en question ne correspond pas à une expérience reproductible et la notion de probabilité n'a plus d'interprétation en termes de fréquence, mais uniquement en termes de quantification de notre degré de certitude subjectif quant à la réalisation de l'événement en question ; il existe diverses variantes de ce type d'interprétation de la notion de probabilité, souvent regroupées sous le terme de **probabilité bayésienne** ou **épistémique**.

Désirant modéliser les phénomènes aléatoires, il est important que les propriétés que l'on impose à la fonction attribuant à chaque événement sa probabilité soient naturelles. Une façon de déterminer un ensemble de bonnes conditions est de considérer l'interprétation fréquentiste. Répétons N fois une expérience, dans les mêmes conditions, et notons $f_N(A)$ la fréquence de réalisation de l'événement A (c'est-à-dire le nombre de fois N_A où il a été réalisé divisé par N). On a alors, au moins heuristiquement,

$$\mathbb{P}(A) = \lim_{N\to\infty} f_N(A).$$

On peut ainsi déduire un certain nombre de propriétés naturelles de $\mathbb{P}$ à partir de celles des fréquences. En particulier, $0 \leqslant f_N(A) \leqslant 1$, $f_N(\Omega) = 1$ et, si A et B sont deux événements disjoints, $N_{A\cup B} = N_A + N_B$, et donc $f_N(A \cup B) = f_N(A) + f_N(B)$. Il est donc raisonnable d'exiger qu'une mesure de probabilité possède les propriétés correspondantes :

1. $0 \leqslant \mathbb{P}(A) \leqslant 1$;
2. $\mathbb{P}(\Omega) = 1$;
3. Si $A \cap B = \varnothing$, alors $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

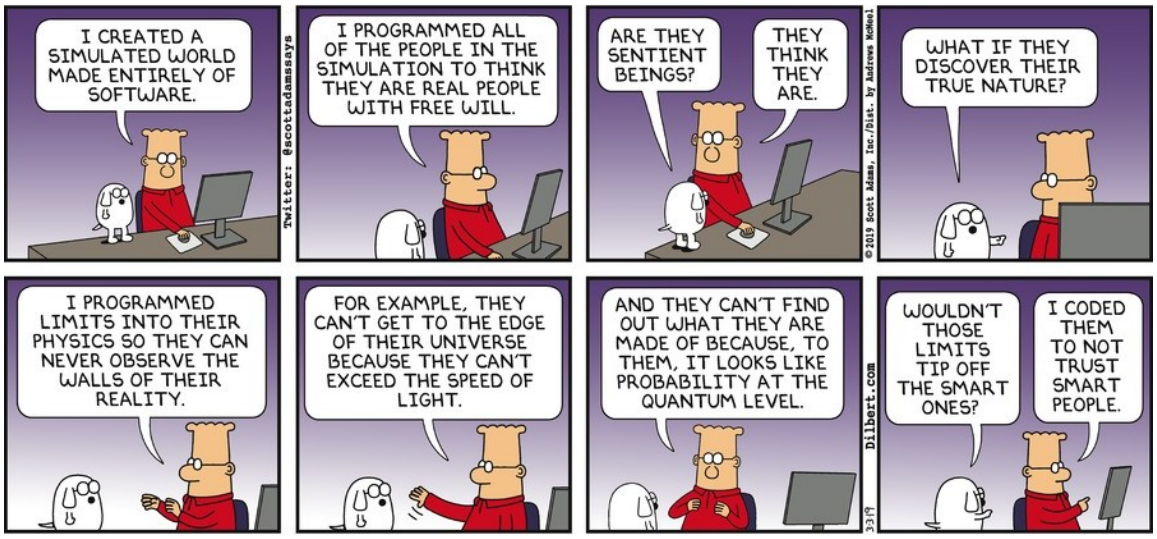
Exemple 0.6. On jette deux dés équilibrés. Il est alors naturel de prendre $\Omega = \{(n,m) \in \{1,2,3,4,5,6\}^2\}$. Les dés étant supposés équilibrés, la symétrie du problème fait qu'il n'y a aucune raison de penser un résultat plus vraisemblable qu'un autre (c'est le **principe d'indifférence**, originellement proposé par Laplace). On associe donc à chaque événement élémentaire $\{(n,m)\}$ la même probabilité $1/36$, ce qui conduit, par les propriétés ci-dessus, à définir la probabilité d'un événement A par $\mathbb{P}(A) = |A|/36$, où $|A|$ représente la cardinalité de A. On a ainsi, par exemple, que la probabilité que la somme des dés soit égale à 10 est donnée par $\mathbb{P}(\{(6,4),(5,5),(4,6)\}) = 3/36 = 1/12$. ⌟

Les conditions ci-dessus sont tout à fait naturelles et suffisent presque à construire la théorie des probabilités. Cependant, pour les mêmes raisons qu'en théorie de la mesure, il s'avère utile d'imposer une condition plus forte que 3., à savoir

3'. Si $(A_k)_{k\geqslant 1}$ est une collection d'événements deux à deux disjoints, alors

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Remarque 0.7. *Les conditions ci-dessus montrent que, sur un plan purement formel, la théorie des probabilités se réduit*[17] *à la théorie de la mesure restreinte à des mesures (positives) satisfaisant* $\mathbb{P}(\Omega) = 1$. *En particulier, comme cela a été vu dans le cours de théorie de la mesure, lorsque* Ω *est un ensemble infini non dénombrable, il n'est en général pas possible de définir la mesure* $\mathbb{P}$ *sur tous les sous-ensembles de* Ω, *mais uniquement sur une tribu* $\mathcal{F}$ *de sous-ensembles. Pratiquement, ceci n'aura cependant absolument aucune incidence, aucun des sous-ensembles exclus n'admettant de description explicite*[18].



©Scott Adams

17. Mais elle s'en distingue considérablement par les problèmes abordés. Le célèbre mathématicien Mark Kac aurait un jour déclaré : *probability theory is measure theory with a soul*. C'est dans cet esprit qu'est rédigé ce cours, avec une emphase sur les aspects probabilistes plutôt que sur les aspects purement analytiques.

18. On peut montrer que la preuve de l'existence d'ensembles non mesurables repose *nécessairement* sur l'axiome du choix. Par conséquent, il est impossible d'en expliciter un.

1 Probabilité, probabilité conditionnelle et indépendance

La théorie des probabilités n'est, au fond, que le bon sens réduit au calcul.

Pierre-Simon de Laplace

1.1 Espace de probabilité

1.1.1 Définition et propriétés élémentaires

Comme expliqué dans l'introduction, la théorie des probabilités est bâtie sur un triplet[1] :

- ▷ Un ensemble Ω, l'**univers**.
- ▷ Une tribu $\mathcal{F}$ sur Ω, la **tribu des événements**.
- ▷ Une mesure $\mathbb{P}$ sur l'espace mesurable $(\Omega, \mathcal{F})$ satisfaisant $\mathbb{P}(\Omega) = 1$, la **mesure de probabilité**.

Définition 1.1. *Un **espace de probabilité** est un triplet $(\Omega, \mathcal{F}, \mathbb{P})$, où $(\Omega, \mathcal{F})$ est un espace mesurable et $\mathbb{P}$ est une mesure sur $(\Omega, \mathcal{F})$ satisfaisant $\mathbb{P}(\Omega) = 1$.*

Les propriétés suivantes d'une mesure de probabilité sont des conséquences immédiates de la définition.

Lemme 1.2. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité.*

1. *Pour tout $A \in \mathcal{F}$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.*
2. *(Monotonicité) Pour tout $A \subset B \in \mathcal{F}$, $\mathbb{P}(A) \leqslant \mathbb{P}(B)$.*
3. *(Additivité finie) Soit $A_1, \dots, A_n$ une collection finie d'événements 2 à 2 disjoints. Alors, $\mathbb{P}\big(\bigcup_{i=1}^n A_i\big) = \sum_{i=1}^n \mathbb{P}(A_i)$.*
4. *(Sous-σ-additivité) Soit I un ensemble fini ou dénombrable et $(A_i)_{i\in I} \subset \mathcal{F}$. Alors, $\mathbb{P}\big(\bigcup_{i\in I} A_i\big) \leqslant \sum_{i\in I} \mathbb{P}(A_i)$.*
5. *Pour tout $A, B \in \mathcal{F}$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.*

Démonstration. *3.* Il suffit d'appliquer la propriété de σ-additivité à la collection $(B_k)_{k\geqslant 1}$ avec $B_k := A_k$ pour $1 \leqslant k \leqslant n$, et $B_k := \varnothing$ pour $k > n$, puisque $\mathbb{P}(\varnothing) = 0$.
1. $1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$.

1. Si nécessaire, l'appendice A.1 contient quelques rappels de théorie de la mesure.

2. $\mathbb{P}(B) = \mathbb{P}\big(A \cup (B \setminus A)\big) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geqslant \mathbb{P}(A)$.

4. Il suffit de considérer le cas dénombrable $(A_k)_{k\geqslant 1}$ (sinon on complète avec une infinité de copies de l'ensemble vide). Introduisons $B_1 \coloneqq A_1$ et, pour $k \geqslant 2$, $B_k \coloneqq A_k \setminus \bigcup_{i=1}^{k-1} B_i$. On a alors $\bigcup_{k=1}^{n} A_k = \bigcup_{k=1}^{n} B_k$ pour tout n, $B_i \cap B_j = \varnothing$ si $i \neq j$, et $B_k \subset A_k$ pour tout k. Par conséquent, $\mathbb{P}(\bigcup_{k=1}^{\infty} A_k) = \mathbb{P}(\bigcup_{k=1}^{\infty} B_k) = \sum_{k=1}^{\infty} \mathbb{P}(B_k) \leqslant \sum_{k=1}^{\infty} \mathbb{P}(A_k)$.

5. Suit de $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}\big(B \setminus (A \cap B)\big)$ et de $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}\big(B \setminus (A \cap B)\big)$. □

Le dernier point s'étend à toute collection finie d'événements : c'est le **principe d'inclusion-exclusion** :

Lemme 1.3. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. Alors, $\forall A_1, A_2, \dots, A_n \in \mathcal{F}$,*

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

*En outre, les sommes partielles des premiers termes du membre de droite fournissent alternativement un majorant et un minorant du membre de gauche (**inégalités de Bonferroni**[2]).*

Démonstration. Sera faite en exercices. □

La σ-additivité de la mesure de probabilité a également pour conséquence l'importante propriété de continuité suivante.

Lemme 1.4. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité.*

1. *Soit $(A_n)_{n\geqslant 1}$ une suite croissante d'événements, c'est-à-dire telle que $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$. Alors,*
$$\mathbb{P}(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n),$$
où $\lim_{n\to\infty} A_n \coloneqq \bigcup_{n=1}^{\infty} A_n$.
2. *Soit $(B_n)_{n\geqslant 1}$ une suite décroissante d'événements, c'est-à-dire telle que $B_1 \supseteq B_2 \supseteq B_3 \supseteq \cdots$. Alors,*
$$\mathbb{P}(\lim_{n\to\infty} B_n) = \lim_{n\to\infty} \mathbb{P}(B_n),$$
où $\lim_{n\to\infty} B_n \coloneqq \bigcap_{n=1}^{\infty} B_n$.

Démonstration. On pose $A_0 \coloneqq \varnothing$ et on observe que $\bigcup_{k\geqslant 1} A_k = \bigcup_{k\geqslant 1}(A_k \setminus A_{k-1})$. Alors,

$$\mathbb{P}(\lim_{n\to\infty} A_n) = \sum_{k=1}^{\infty} \mathbb{P}(A_k \setminus A_{k-1}) = \lim_{n\to\infty} \sum_{k=1}^{n} \big(\mathbb{P}(A_k) - \mathbb{P}(A_{k-1})\big) = \lim_{n\to\infty} \big(\mathbb{P}(A_{n+1}) - \mathbb{P}(A_0)\big) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

La seconde affirmation suit facilement, puisque la suite des complémentaires $(B_n^{\mathrm{c}})_{n\geqslant 1}$ est croissante. On peut donc appliquer la première partie pour obtenir

$$\mathbb{P}(\lim_{n\to\infty} B_n) = \mathbb{P}\Big(\bigcap_{n=1}^{\infty} B_n\Big) = 1 - \mathbb{P}\Big(\bigcup_{n=1}^{\infty} B_n^{\mathrm{c}}\Big) = 1 - \lim_{n\to\infty} \mathbb{P}(B_n^{\mathrm{c}}) = \lim_{n\to\infty} \mathbb{P}(B_n).$$ □

2. Carlo Emilio Bonferroni (1892, Bergame – 1960, Florence), mathématicien italien.

Remarque 1.5. *Soit $(A_n)_{n\geqslant 1}$ une famille d'événements deux à deux disjoints. La suite d'événements $B_N \coloneqq \bigcup_{n=1}^{N} A_n$ est croissante et $\lim_{N\to\infty} B_N = \bigcup_{n\geqslant 1} A_n$. Par conséquent, la propriété de continuité ci-dessus et l'additivité finie de $\mathbb{P}$ implique sa σ-additivité :*

$$\mathbb{P}\Big(\bigcup_{n\geqslant 1} A_n\Big) = \mathbb{P}\Big(\lim_{N\to\infty} B_N\Big) = \lim_{N\to\infty} \mathbb{P}(B_N) = \lim_{N\to\infty} \mathbb{P}\Big(\bigcup_{n=1}^{N} A_n\Big) = \lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

$\mathbb{P}$ est donc σ-additive si et seulement si elle est finiment additive et continue (au sens ci-dessus).

1.1.2 Espaces de probabilité discrets

On parle d'**espace de probabilité discret** lorsque Ω est fini ou dénombrable ; on considère alors généralement $\mathcal{F} \coloneqq \mathscr{P}(\Omega)$. Dans le cas d'un espace de probabilité discret $(\Omega, \mathcal{F}, \mathbb{P})$, la mesure de probabilité $\mathbb{P}$ est entièrement déterminée par la valeur qu'elle associe aux événements élémentaires, c'est-à-dire aux valeurs $\mathbb{P}(\{\omega\})$, $\omega \in \Omega$. En effet, on a alors, par additivité ou σ-additivité,

$$\forall A \in \mathcal{F}, \qquad \mathbb{P}(A) = \mathbb{P}\Big(\bigcup_{\omega\in A} \{\omega\}\Big) = \sum_{\omega\in A} \mathbb{P}(\{\omega\}).$$

La fonction de Ω dans $[0,1]$ donnée par $\omega \mapsto \mathbb{P}(\{\omega\})$ est appelée **fonction de masse**.

Exemple 1.6.

- ▷ Pour modéliser un dé équilibré, $\mathbb{P}(\{\omega\}) = \frac{1}{6}$, pour tout $\omega \in \Omega \coloneqq \{1,2,3,4,5,6\}$.
- ▷ Pour 5 lancers d'une pièce équilibrée, $\mathbb{P}(\{\omega\}) = 2^{-5}$, pour tout $\omega \in \Omega \coloneqq \{\mathsf{P},\mathsf{F}\}^5$.
- ▷ On est intéressé à la couleur d'une boule extraite d'une urne contenant 5 boules rouges et 1 boule verte. On prend $\Omega \coloneqq \{\mathrm{R},\mathrm{V}\}$. On a $\mathbb{P}(\{\mathrm{R}\}) = 5/6$ et $\mathbb{P}(\{\mathrm{V}\}) = 1/6$. ⌟

Dans les deux premiers exemples, tous les événements élémentaires ont même probabilité.

Définition 1.7. *On appelle **distribution de probabilité uniforme** sur un univers Ω fini, la mesure de probabilité définie par $\mathbb{P}(\{\omega\}) \coloneqq 1/|\Omega|$, pour tout $\omega \in \Omega$. On dit dans ce cas qu'il y a **équiprobabilité**.*

Évidemment, une telle distribution n'existe pas si Ω est infini dénombrable. Manifestement, lorsqu'il y a équiprobabilité, la probabilité d'un événement A est simplement donnée par le quotient entre le nombre de « cas favorables » et le nombre total de « cas possibles » :

$$\mathbb{P}(A) = \sum_{\omega\in A} \frac{1}{|\Omega|} = \frac{|A|}{|\Omega|}.$$

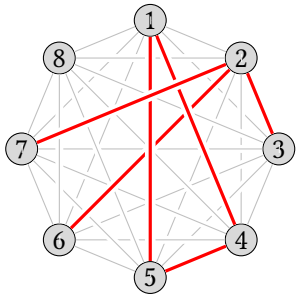
Ceci réduit de nombreux problèmes de calcul de probabilité dans ce contexte à de la combinatoire ; nous renvoyons au cours de mathématiques discrètes pour les notions et outils pertinents.

Exemple 1.8. Au poker, une main est constituée de 5 cartes tirées au hasard uniformément à partir d'un jeu de 52 cartes. Le nombre de mains possibles est donc de $\binom{52}{5}$. Calculons la probabilité d'avoir un brelan (c'est-à-dire trois cartes de même valeur et deux cartes de valeurs différentes des trois précédentes et distinctes entre elles). Il suffit de déterminer le nombre de mains correspondant à cet événement. Il y a $\binom{13}{1}$ choix pour la valeur du brelan, $\binom{4}{3}$ pour les couleurs correspondantes, $\binom{12}{2}$ pour les valeurs des deux cartes restantes et 4^2 choix pour leurs couleurs. La probabilité recherchée est donc donnée par

$$\frac{\binom{13}{1}\binom{4}{3}\binom{12}{2}4^2}{\binom{52}{5}} = \frac{88}{4\,165} \cong 2,1\%.$$ ⌟

Exemple 1.9. Nous allons à présent introduire un exemple d'espace de probabilité fini qui a été beaucoup étudié : le graphe aléatoire d'Erdős[3]–Rényi[4]. Soit $m \geqslant 0$ et $n \geqslant 1$ deux entiers. Le **graphe aléatoire** $\mathcal{G}(n, m)$ est l'espace de probabilité $(\Omega, \mathscr{P}(\Omega), \mathbb{P})$, où les éléments de Ω sont les graphes $G = (S, A)$ avec ensemble de sommets $S = \{1, \dots, n\}$ et ensemble d'arêtes $A \subset \big\{\{i, j\} \,\big|\, 1 \leqslant i < j \leqslant n\big\}$ satisfaisant $|A| = m$, et où $\mathbb{P}$ est la mesure de probabilité uniforme sur Ω. En d'autres termes, l'ensemble des n sommets est fixé, mais l'emplacement des m arêtes est aléatoire. Le graphe complet avec n sommets possédant $N \coloneqq \binom{n}{2}$ arêtes, on a $|\Omega| = \binom{N}{m}$.

À titre d'exemple, voici une réalisation du graphe aléatoire $\mathcal{G}(8, 6)$ (les arêtes présentes sont indiquées en rouge) :



⌟

Exemple 1.10. On lance un dé équilibré jusqu'à la première apparition d'un 6. Si l'on s'intéresse au nombre de lancers nécessaires, il est naturel de prendre $\Omega \coloneqq \mathbb{N}^\star \cup \{+\infty\}$, la valeur $+\infty$ correspondant à une expérience où le 6 ne sort jamais. Déterminons la probabilité à associer aux différentes valeurs de n. Pour un dé équilibré, on a $\mathbb{P}(\{n\}) = \frac{1}{6}(\frac{5}{6})^{n-1}$, pour tout $n \in \mathbb{N}^\star$. En effet, parmi les 6^n suites possibles de n lancers, seules 5^{n-1} ont un 6 au n-ième tirage mais aucun auparavant. En particulier,

$$\mathbb{P}(\{+\infty\}) = 1 - \sum_{n=1}^{\infty} \mathbb{P}(\{n\}) = 1 - \tfrac{1}{6} \sum_{n=1}^{\infty} (\tfrac{5}{6})^{n-1} = 0,$$

et la probabilité de ne jamais voir de 6 est nulle (mais certainement pas *impossible* : il y a une infinité non dénombrable de séries de lancers lors desquelles cela se produit !). ⌟

1.1.3 Mesures de probabilité à densité

Nous noterons $\mathcal{B}(\mathbb{R}^n)$ l'ensemble des boréliens de $\mathbb{R}^n$ et λ_n la mesure de Lebesgue sur $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ (on écrira parfois λ pour parler de λ_1 et on utilisera occasionnellement l'écriture $\int f(x)\,\mathrm{d}x \equiv \int f(x)\,\mathrm{d}\lambda(x)$). Plus généralement, si $A \subset \mathbb{R}^n$, on notera $\mathcal{B}(A)$ l'ensemble des boréliens de A.

Définition 1.11. *Soit $n \in \mathbb{N}^*$ et $f : \mathbb{R}^n \to \mathbb{R}$ une fonction mesurable positive, telle que $\int_{\mathbb{R}^n} f(x)\,\mathrm{d}\lambda_n(x) = 1$. Alors l'application $\mathbb{P} : \mathcal{B}(\mathbb{R}^n) \to [0, 1]$ définie par*

$$\forall A \in \mathcal{B}(\mathbb{R}^n), \qquad \mathbb{P}(A) \coloneqq \int_A f(x)\,\mathrm{d}\lambda_n(x)$$

*est une mesure de probabilité sur $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$[5]. Dans ce cas, on dit que la mesure $\mathbb{P}$ est **absolument continue** (par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$), et f est appelée la **densité de probabilité**[6] associée à $\mathbb{P}$.*

3. Pál Erdős (1913, Budapest – 1996, Varsovie), également orthographié Paul Erdős, Paul Erdös ou Paul Erdos, mathématicien hongrois.
4. Alfréd Rényi (1921, Budapest – 1970, Budapest), mathématicien hongrois.
5. La vérification est laissée en exercice.
6. En d'autres termes, la fonction f est la dérivée de Radon–Nikodym de $\mathbb{P}$ par rapport à λ_n.

Remarque 1.12. *1. Insistons sur le fait que $f(x)$ n'est **pas** une probabilité (en particulier, $f(x)$ peut être plus grande que 1).*

2. La densité de probabilité associée à une mesure de probabilité absolument continue $\mathbb{P}$ n'est pas unique : si f est une densité de probabilité pour $\mathbb{P}$ et g ne diffère de f que sur un ensemble de mesure de Lebesgue nulle, alors g est également une densité de probabilité pour $\mathbb{P}$. En effet, si $B = \{x \in \mathbb{R}^n \mid f(x) \neq g(x)\}$, alors

$$\int_A f \,\mathrm{d}\lambda_n = \int_{A\setminus B} f \,\mathrm{d}\lambda_n + \int_{A\cap B} f \,\mathrm{d}\lambda_n = \int_{A\setminus B} g \,\mathrm{d}\lambda_n = \int_A g \,\mathrm{d}\lambda_n,$$

puisque $\lambda_n(A \cap B) = 0$ implique que $\int_{A\cap B} f \,\mathrm{d}\lambda_n = \int_{A\cap B} g \,\mathrm{d}\lambda_n = 0$.
On vérifie facilement que c'est la seule obstruction à l'unicité. Parler de « la » fonction de densité associée à une mesure de probabilité $\mathbb{P}$ ne portera donc pas à conséquence.

Exemple 1.13. Soit $A \in \mathcal{B}(\mathbb{R}^n)$ tel que $\lambda_n(A) > 0$. La **mesure de probabilité uniforme** sur A est la mesure de probabilité de densité de probabilité

$$f(x) = \frac{\mathbf{1}_A(x)}{\lambda_n(A)}.$$ ⌟

1.2 Probabilité conditionnelle, formule de Bayes

De nombreuses affirmations prennent la forme « si B a lieu, alors la probabilité de A est p », où B et A sont des événements (tels « il neige demain », et « le bus sera à l'heure », respectivement). Afin de motiver la définition de la probabilité conditionnelle d'un événement A étant connue la réalisation d'un événement B, revenons à l'interprétation fréquentiste des probabilités. On considère deux événements A et B. On désire déterminer la fréquence de réalisation de l'événement A lorsque l'événement B a lieu. Une façon de procéder est la suivante : on répète l'expérience un grand nombre de fois N. On note N_B le nombre de tentatives lors desquelles B est réalisé, et $N_{A\cap B}$ le nombre de ces dernières tentatives lors desquelles A est également réalisé. La fréquence de réalisation de A *parmi les tentatives ayant donné lieu à* B est alors donnée par

$$\frac{N_{A\cap B}}{N_B} = \frac{N_{A\cap B}}{N}\frac{N}{N_B}.$$

D'après l'interprétation fréquentiste, lorsque N devient grand, le membre de gauche devrait converger vers la probabilité de A conditionnellement à la réalisation de l'événement B, alors que le membre de droite devrait converger vers $\mathbb{P}(A \cap B)/\mathbb{P}(B)$. Ceci motive la définition suivante.

Définition 1.14. *Soit $B \in \mathcal{F}$ un événement tel que $\mathbb{P}(B) > 0$. Pour tout $A \in \mathcal{F}$, la **probabilité conditionnelle de A sachant B** est la quantité*

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

Remarque 1.15. *On vérifie aisément que l'application $\mathbb{P}(\,\cdot \mid B) : \mathcal{F} \to \mathbb{R}$ est une mesure de probabilité sur $(\Omega, \mathcal{F})$ (et également sur $(B, \mathcal{F}_B)$, où $\mathcal{F}_B \coloneqq \{A \cap B \mid A \in \mathcal{F}\}$).*

Exemple 1.16. On lance deux fois un dé équilibré. Sachant que le résultat du premier lancer est un 3, quelle est la probabilité que la somme des deux résultats soit supérieure à 6 ? Ici, $B = \{(3, k) \mid k = 1, \dots, 6\}$, $A = \{(a, b) \in \{1, \dots, 6\}^2 \mid a + b > 6\}$ et $A \cap B = \{(3, 4), (3, 5), (3, 6)\}$. On a alors

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{|A \cap B|}{|B|} = \frac{3}{6} = \frac{1}{2}.$$ ⌟

Exemple 1.17. On suppose qu'à chaque naissance, il y a une probabilité $\frac{1}{2}$ que le nouveau-né soit un garçon et $\frac{1}{2}$ que ce soit une fille. On considère les problèmes suivants.

1. Parmi toutes les familles avec deux enfants dont au moins un est un garçon, on en sélectionne une au hasard uniformément. Quelle est la probabilité que l'autre enfant soit également un garçon ?
2. Parmi toutes les familles avec deux enfants dont l'aîné est un garçon, on en sélectionne une au hasard uniformément. Quelle est la probabilité que l'autre enfant soit également un garçon ?

Dans les deux cas, on considère pour Ω l'ensemble de toutes les possibilités pour les sexes des deux enfants. On a donc $\Omega = \{(\mathrm{G},\mathrm{G}),(\mathrm{F},\mathrm{F}),(\mathrm{F},\mathrm{G}),(\mathrm{G},\mathrm{F})\}$, où le premier membre de chaque paire représente le sexe de l'aîné et le second celui du cadet. L'intérêt de distinguer l'aîné et le cadet est que la mesure de probabilité décrivant notre problème devient uniforme : chacune de ces 4 possibilités a probabilité $1/4$. Dans les deux cas, nous sommes intéressés à la probabilité de l'événement $A \coloneqq \{(\mathrm{G},\mathrm{G})\}$.

Dans le premier problème, l'information qui nous est donnée est qu'au moins un des enfants est un garçon, ce qui revient à conditionner par l'événement $B \coloneqq \{(\mathrm{G},\mathrm{G}),(\mathrm{F},\mathrm{G}),(\mathrm{G},\mathrm{F})\}$. La probabilité cherchée est donc

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(\{(\mathrm{G},\mathrm{G})\})}{\mathbb{P}(\{(\mathrm{G},\mathrm{G}),(\mathrm{F},\mathrm{G}),(\mathrm{G},\mathrm{F})\})} = \frac{1}{3}.$$

Passons à présent au second problème. L'information obtenue est différente : on sait qu'un des enfants est un garçon, mais également qu'il s'agit de l'aîné, ce qui correspond à l'événement $C \coloneqq \{(\mathrm{G},\mathrm{G}),(\mathrm{G},\mathrm{F})\}$. On a donc

$$\mathbb{P}(A \mid C) = \frac{\mathbb{P}(\{(\mathrm{G},\mathrm{G})\})}{\mathbb{P}(\{(\mathrm{G},\mathrm{G}),(\mathrm{G},\mathrm{F})\})} = \frac{1}{2}. \qquad \lrcorner$$

Exemple 1.18. Considérons à présent une variante du premier problème de l'Exercice 1.17. Albert est invité par sa collègue Zoé. Il sait que celle-ci a deux enfants. Lorsqu'il sonne à la porte, un garçon vient lui ouvrir. Quelle est la probabilité que son autre enfant soit également un garçon ? Considérons l'univers $\Omega \coloneqq \{(\mathrm{G}_*,\mathrm{G}),(\mathrm{F}_*,\mathrm{F}),(\mathrm{F}_*,\mathrm{G}),(\mathrm{G}_*,\mathrm{F}),(\mathrm{G},\mathrm{G}_*),(\mathrm{F},\mathrm{F}_*),(\mathrm{F},\mathrm{G}_*),(\mathrm{G},\mathrm{F}_*)\}$, où l'étoile indique quel enfant ouvre la porte. Nous supposerons les 8 possibilités équiprobables. L'événement qui nous intéresse est $A \coloneqq \{(\mathrm{G}_*,\mathrm{G}),(\mathrm{G},\mathrm{G}_*)\}$. L'information dont Albert dispose au moment où un garçon ouvre la porte correspond à l'événement $B \coloneqq \{(\mathrm{G}_*,\mathrm{G}),(\mathrm{G}_*,\mathrm{F}),(\mathrm{G},\mathrm{G}_*),(\mathrm{F},\mathrm{G}_*)\}$. On a donc

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(\{(\mathrm{G}_*,\mathrm{G}),(\mathrm{G},\mathrm{G}_*)\})}{\mathbb{P}(\{(\mathrm{G}_*,\mathrm{G}),(\mathrm{G}_*,\mathrm{F}),(\mathrm{G},\mathrm{G}_*),(\mathrm{F},\mathrm{G}_*)\})} = \frac{1}{2}.$$

Ainsi, bien que ce problème ressemble superficiellement au premier problème de l'Exemple 1.17, la façon dont l'information est acquise influence le résultat. $\lrcorner$

Définition 1.19. *Une famille $(B_i)_{i\in I} \subset \mathcal{F}$, I fini ou dénombrable, est une **partition** de Ω si*

$$B_i \cap B_j = \varnothing, \text{ dès que } i \neq j, \qquad \text{et} \qquad \bigcup_{i\in I} B_i = \Omega.$$

En dépit de sa simplicité, le théorème suivant est crucialement important en théorie des probabilités.

Théorème 1.20. *Soit $(B_i)_{i\in I}$ une partition de Ω telle que $\mathbb{P}(B_i) > 0$, pour tout $i \in I$, et soit $A \in \mathcal{F}$.*

1. ***(Loi de la probabilité totale)***
$$\mathbb{P}(A) = \sum_{i\in I} \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i).$$
2. ***(Formule de Bayes[7])*** *Si $\mathbb{P}(A) > 0$,*
$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i)}{\sum_{j\in I} \mathbb{P}(A \mid B_j)\,\mathbb{P}(B_j)}.$$

7. Thomas Bayes (1702, Londres – 1761, Tunbridge Wells), mathématicien britannique et pasteur de l'Église presbytérienne.

Démonstration. Par σ-additivité,

$$\sum_{i\in I}\mathbb{P}(A\,|\,B_i)\,\mathbb{P}(B_i)=\sum_{i\in I}\mathbb{P}(A\cap B_i)=\mathbb{P}\big(\bigcup_{i\in I}(A\cap B_i)\big)=\mathbb{P}\big(A\cap\bigcup_{i\in I}B_i\big)=\mathbb{P}(A).$$

La seconde affirmation suit de l'observation que

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

et l'application de la loi de la probabilité totale au dénominateur. □

Remarque 1.21. *Dans la terminologie statistique, on appelle $\mathbb{P}(B_i)$ la probabilité **à priori** de B_i et $\mathbb{P}(B_i\,|\,A)$ la probabilité **à posteriori** de B_i (sachant A). La formule de Bayes donne donc un moyen de transformer les probabilités à priori en probabilités à posteriori.*

Exemple 1.22. On se donne deux urnes. La première contient deux boules blanches et trois boules noires ; la seconde trois blanches et quatre noires. Une boule est tirée au hasard de la première urne et placée dans la seconde. On tire ensuite au hasard une boule de la seconde urne : quelle est la probabilité qu'elle soit noire ?

Soit A l'événement « la boule tirée de la seconde urne est noire », et B l'événement « la boule déplacée de la première urne à la seconde est noire ». Puisque B et B^{c} forment une partition de Ω, une application de la loi de la probabilité totale donne

$$\mathbb{P}(A)=\mathbb{P}(A\,|\,B)\,\mathbb{P}(B)+\mathbb{P}(A\,|\,B^{\mathrm{c}})\,\mathbb{P}(B^{\mathrm{c}}).$$

À présent,

$$\mathbb{P}(A\,|\,B)=\mathbb{P}(A\,|\,\text{la }2^{\text{nde}}\text{ urne contient trois boules blanches et cinq noires})=\tfrac{5}{8};$$
$$\mathbb{P}(A\,|\,B^{\mathrm{c}})=\mathbb{P}(A\,|\,\text{la }2^{\text{nde}}\text{ urne contient quatre boules blanches et quatre noires})=\tfrac{1}{2}.$$

Puisque $\mathbb{P}(B)=\frac{3}{5}$ et $\mathbb{P}(B^{\mathrm{c}})=\frac{2}{5}$, on obtient $\mathbb{P}(A)=\frac{23}{40}$. ⌟

Exemple 1.23. Le test de dépistage d'un certain virus n'est pas infaillible :

- ▷ Faux positifs : 1 fois sur 100, il est positif, alors que l'individu n'est pas contaminé ;
- ▷ Faux négatifs : 2 fois sur 100, il est négatif, alors que l'individu est contaminé.

Il est donc important de répondre aux questions suivantes :

- ▷ Étant donné que son test est positif, quelle est la probabilité qu'un individu ne soit pas porteur du virus ?
- ▷ Étant donné que son test est négatif, quelle est la probabilité qu'un individu soit porteur du virus ?

La formule de Bayes est parfaitement adaptée à ce type de calculs. Afin de pouvoir l'appliquer, il nous faut toutefois une information supplémentaire : dans la population totale, la fraction de porteurs est approximativement de $1/1000$.

Formalisons tout cela. On introduit les événements suivants :

$$T=\{\text{le test est positif}\},\qquad V=\{\text{l'individu est contaminé}\}.$$

On a donc les informations suivantes :

$$\mathbb{P}(T\,|\,V^{\mathrm{c}})=\tfrac{1}{100},\qquad\mathbb{P}(T^{\mathrm{c}}\,|\,V)=\tfrac{2}{100},\qquad\mathbb{P}(V)=\tfrac{1}{1\,000},$$

et on veut calculer $\mathbb{P}(V^{\mathrm{c}}\,|\,T)$ et $\mathbb{P}(V\,|\,T^{\mathrm{c}})$. La formule de Bayes nous dit que

$$\mathbb{P}(V^{\mathrm{c}}\,|\,T)=\frac{\mathbb{P}(T\,|\,V^{\mathrm{c}})\,\mathbb{P}(V^{\mathrm{c}})}{\mathbb{P}(T\,|\,V^{\mathrm{c}})\,\mathbb{P}(V^{\mathrm{c}})+\mathbb{P}(T\,|\,V)\,\mathbb{P}(V)}.$$

Toutes les valeurs correspondant aux quantités du membre de droite sont connues (observez que $\mathbb{P}(T \mid V) = 1 - \mathbb{P}(T^{\mathrm{c}} \mid V) = 98/100$) et l'on obtient $\mathbb{P}(V^{\mathrm{c}} \mid T) \cong 0{,}91$. En d'autres termes, même si son test est positif, un individu a plus de 90% de chances de ne pas être porteur du virus !

Un calcul similaire montre par contre que $\mathbb{P}(V \mid T^{\mathrm{c}}) \cong 0{,}000\,02$, ce qui montre que c'est bien là que se trouve l'utilité de ce test, puisque la probabilité de déclarer non porteur un individu contaminé est de l'ordre de $2/100\,000$.

Observez que le calcul ci-dessus ne s'applique qu'à un individu « normal ». Dans le cas d'un individu appartenant à une population à risque, la probabilité à priori d'être porteur, $\mathbb{P}(V)$, peut devenir proche de 1, plutôt que très petite comme précédemment. Cela change complètement les conclusions : dans ce cas, la probabilité d'être non porteur alors que le test est positif est faible, tandis que la probabilité d'être porteur alors que le test est négatif est importante. ⌟

L'usage des probabilités conditionnelles peut se révéler délicat et il est conseillé d'appliquer consciencieusement les règles plutôt que de se reposer uniquement sur son intuition.

Exemple 1.24. (Paradoxe du prisonnier) Trois hommes se sont fait arrêter dans une sombre dictature. Ils apprennent de leur garde que le dictateur a décidé arbitrairement que l'un d'entre eux va être libéré et les deux autres exécutés ; le garde n'est pas autorisé à annoncer à un prisonnier quel sera son sort. Le prisonnier A sait donc que la probabilité qu'il soit épargné est de $1/3$. Afin d'obtenir davantage d'informations, il décide d'interroger le garde. Il lui demande de lui donner en secret le nom d'un de ses camarades qui sera exécuté (après tout, il sait déjà que c'est le cas d'au moins l'un d'entre eux, le garde ne révèle donc pas grand chose). Le garde nomme le prisonnier B. Le prisonnier A sait à présent qu'entre lui-même et C, l'un va être libéré, et l'autre exécuté et en conclut que sa probabilité de survie est à présent de $1/2$. Pourquoi A se trompe-t-il ? Quelle est à présent la probabilité que C soit libéré ? ⌟

Remarque 1.25. *Dans l'exemple précédent, le problème est partiellement mal posé, car la stratégie employée par le garde pour choisir quel prisonnier désigner, lorsque les prisonniers B et C vont tous deux être exécutés, n'est pas indiquée. Dans une telle situation, on suppose généralement qu'il prend sa décision de façon uniforme (après tout, sans aucune information sur le sujet, tout autre choix serait difficile à justifier). Plus généralement, on pourrait modéliser la procédure de décision du garde en introduisant un paramètre supplémentaire.*

Si l'exemple précédent est très artificiel et se règle facilement en appliquant avec soin les règles de la théorie des probabilités, l'exemple suivant montre que des difficultés réelles, subtiles et difficiles à traiter apparaissent également dans des applications pratiques tout à fait concrètes.

Exemple 1.26. (**Paradoxe de Simpson**[8]) Un scientifique a effectué des expériences cliniques afin de déterminer les efficacités relatives de deux traitements. Il a obtenu les résultats suivants :

	Traitement A	Traitement B
Succès	219	1 010
Échec	1 801	1 190

Le traitement A ayant été administré à 2 020 personnes, et 219 d'entre elles ayant été guéries, son taux de succès est donc de $219/2\,020$, ce qui est très inférieur au taux correspondant pour le traitement B qui est de $1\,010/2\,200$. Par conséquent, le traitement B est plus efficace que le traitement A.

Après qu'il a annoncé sa conclusion, il reçoit la visite de l'un de ses assistants, qui est en désaccord avec l'interprétation des résultats. Il lui présente le tableau suivant, dans lequel les résultats précédents sont donnés en tenant compte du sexe des patients :

8. Edward Hugh Simpson (1922 – 2019), statisticien britannique. Ce paradoxe, discuté par ce dernier en 1951, l'avait déjà été en 1899 par Karl Pearson et ses coauteurs, puis en 1903 par George Udny Yule.

	Femmes		Hommes	
	Traitement A	Traitement B	Traitement A	Traitement B
Succès	200	10	19	1 000
Échec	1 800	190	1	1 000

Chez les femmes, les taux de succès des traitements sont de $1/10$ et $1/20$ respectivement, et chez les hommes de $19/20$ et $1/2$. Le traitement A est donc plus efficace que le traitement B dans les 2 cas. L'assistant a bien raison : quel que soit le sexe du patient, ses chances de guérir sont supérieures avec le traitement A.

Dans l'exemple précédent, le paradoxe résulte de deux causes. D'une part, les traitements ont été distribués de façon très inhomogène selon le sexe des patients : 2 000 femmes ont reçu le traitement A, alors que seules 200 ont reçu le traitement B ; 2 000 hommes ont reçu le traitement B, alors que seuls 20 ont reçu le traitement A. D'autre part, la maladie est beaucoup plus difficile à soigner chez les femmes (environ 11% de guérison) que chez les hommes (un peu plus de 50% de guérison).

Ce paradoxe apparaît régulièrement dans des études statistiques. Observez aussi la difficulté suivante : si l'on n'avait pas relevé le sexe des patients, on aurait été obligé de baser notre analyse sur le premier raisonnement, et on serait arrivé à une conclusion erronée. En particulier, comment être certain qu'il n'existe pas d'autres facteurs que le sexe (l'âge, le poids, …) dont on n'aurait pas tenu compte et qui modifierait une fois de plus la conclusion ? Techniquement, on parle de **facteur de confusion**.

Un cas réel célèbre s'est produit en 1973 lorsque l'université de Berkeley a constaté que les chiffres des admissions en *graduate school* (ce qui correspond au niveau master et doctorat ici) montraient que les hommes ayant déposé leur candidature avaient plus de chance d'être admis que les femmes, et la différence était si importante qu'elle ne pouvait raisonnablement être attribuée au hasard (44% contre 35%). Craignant d'être poursuivie pour discrimination sexuelle, elle a mandaté un statisticien afin d'examiner la situation. Une analyse séparée des différents départements a alors montré qu'aucun d'entre eux n'était significativement biaisé en faveur des hommes ; en fait, la plupart montraient un petit (et pas très significatif) biais en faveur des femmes ! L'explication se trouve être que les femmes avaient tendance à porter leur choix sur des départements dont les taux d'admission sont faibles, tandis que les hommes avaient tendance à postuler dans des départements avec forts taux d'admission (les départements de sciences humaines ayant tendance à disposer de beaucoup moins de fonds de recherche que les départements de sciences et d'ingénierie, ce qui résulte en un plus faible nombre de places). ⌟

1.3 Indépendance

En général, apprendre qu'un événement B est réalisé modifie la probabilité qu'un autre événement A soit réalisé : la probabilité à priori de A, $\mathbb{P}(A)$, est remplacée par la probabilité à posteriori, $\mathbb{P}(A \mid B)$, en général différente. Lorsque l'information que B est réalisé ne modifie pas la probabilité d'occurrence de A, c'est-à-dire lorsque $\mathbb{P}(A \mid B) = \mathbb{P}(A)$, on dit que les événements A et B sont indépendants. Il y a cependant au moins deux bonnes raisons pour ne pas utiliser cette propriété comme définition de l'indépendance : d'une part, elle n'a de sens que lorsque $\mathbb{P}(B) > 0$, et d'autre part, les deux événements ne jouent pas un rôle symétrique. La notion de probabilité conditionnelle conduit donc à la définition suivante.

Définition 1.27. *Deux événements A et B sont **indépendants** sous $\mathbb{P}$ si*

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

*Plus généralement, une famille d'événements $(A_i)_{i \in I}$, avec I arbitraire, est **indépendante** sous $\mathbb{P}$ si*

$$\mathbb{P}(\bigcap_{i \in J} A_i) = \prod_{i \in J} \mathbb{P}(A_i),$$

pour tout $J \subset I$ fini.

Exemple 1.28. Lorsque I est fini, il ne suffit pas, en général, de vérifier que $\mathbb{P}(\bigcap_{i\in I} A_i) = \prod_{i\in I} \mathbb{P}(A_i)$: il est essentiel de vérifier que la factorisation a lieu pour toute sous-collection d'événements. Par exemple, si l'on lance successivement deux dés équilibrés, $\Omega = \{(i,j) \mid 1 \leqslant i,j \leqslant 6\}$, et que l'on considère les événements

$$A = \{\text{le } 1^{\text{er}} \text{ dé montre un } 1, \text{ un } 2 \text{ ou un } 3\},\ B = \{\text{le } 1^{\text{er}} \text{ dé montre un } 3, \text{ un } 4 \text{ ou un } 5\},$$
$$C = \{\text{La somme des deux dés est égale à } 9\},$$

alors on observe que $\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{2}$, $\mathbb{P}(C) = \frac{1}{9}$, $\mathbb{P}(A\cap B\cap C) = \frac{1}{36}$. Ainsi,

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

Par contre, $\mathbb{P}(A\cap B) = \frac{1}{6}$, $\mathbb{P}(A\cap C) = \frac{1}{36}$ et $\mathbb{P}(B\cap C) = \frac{1}{12}$. On a donc

$$\mathbb{P}(A\cap B) \neq \mathbb{P}(A)\mathbb{P}(B), \quad \mathbb{P}(A\cap C) \neq \mathbb{P}(A)\mathbb{P}(C), \quad \mathbb{P}(B\cap C) \neq \mathbb{P}(B)\mathbb{P}(C). \qquad \lrcorner$$

Lemme 1.29. *Soit A, B deux événements indépendants. Alors A et B^{c} sont indépendants, et A^{c} et B^{c} sont indépendants.*

Plus généralement, si $A_1, \dots, A_n$ sont indépendants, alors

$$B_1, \dots, B_n,$$

où $B_i \in \{A_i, A_i^{\mathrm{c}}\}$, sont aussi indépendants.

Démonstration. Laissée en exercice. □

Remarque 1.30. *Si une famille d'événements $(A_i)_{i\in I}$ satisfait $\mathbb{P}(A_i\cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$, pour toute paire $i \neq j$, on dit que la famille est **deux à deux indépendante**. L'indépendance deux à deux* n'implique pas l'indépendance*, comme le montre l'exemple suivant.*

Exemple 1.31. On place dans une boîte 4 billets sur lesquels sont respectivement inscrits les 4 nombres suivants : $112, 121, 211$ et 222. On tire au hasard un des 4 billets (uniformément) et on considère les événements suivants :

$$A_1 = \{\text{Le premier chiffre est un « 1 »}\},$$
$$A_2 = \{\text{Le deuxième chiffre est un « 1 »}\},$$
$$A_3 = \{\text{Le troisième chiffre est un « 1 »}\}.$$

Comme $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) = \frac{1}{2}$ et $\mathbb{P}(A_1\cap A_2) = \mathbb{P}(A_1\cap A_3) = \mathbb{P}(A_2\cap A_3) = \frac{1}{4}$, les événements A_1, A_2 et A_3 sont deux à deux indépendants. Ces trois événements ne sont toutefois pas indépendants, puisque $\mathbb{P}(A_1\cap A_2\cap A_3) = 0$ et $\mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3) = \frac{1}{8}$. $\lrcorner$

Exemple 1.32. Retournons au graphe aléatoire $\mathcal{G}(n,m)$; on suppose $n \geqslant 3$ et $m \geqslant 2$. La probabilité que deux sommets distincts donnés i et j soient reliés par une arête (ce que l'on notera $i \sim j$) est donnée par (rappelez-vous que $N := \binom{n}{2}$)

$$\mathbb{P}(i\sim j) = \frac{\binom{N-1}{m-1}}{\binom{N}{m}} = \frac{m}{N}.$$

En effet, le numérateur correspond au nombre total de façons de choisir les $m-1$ arêtes restantes parmi les $N-1$ arêtes du graphe complet encore disponibles.

D'autre part, soit i,j,k,ℓ quatre sommets tels que $\{i,j\} \neq \{k,\ell\}$. La probabilité qu'on ait à la fois $i\sim j$ et $k\sim\ell$ est donnée par

$$\mathbb{P}(i\sim j, k\sim\ell) = \frac{\binom{N-2}{m-2}}{\binom{N}{m}} = \frac{m(m-1)}{N(N-1)}.$$

On voit donc que les événements $i\sim j$ et $k\sim\ell$ ne sont pas indépendants. $\lrcorner$

Il convient d'être attentif lorsque l'on utilise la notion d'indépendance. En particulier, l'idée intuitive d'indépendance peut être parfois mise en défaut, comme le montrent les deux exemples suivants.

Exemple 1.33. Un événement peut être indépendant de lui-même. En effet, ceci a lieu si et seulement s'il a probabilité 0 ou 1, puisque, dans ce cas, on a bien

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A) \iff \mathbb{P}(A) \in \{0, 1\}.$$ ⌟

Exemple 1.34. Considérons des familles avec 3 enfants et intéressons-nous au sexe des enfants ; on suppose que chacune des 8 possibilités a la même probabilité $1/8$. Soit A l'événement « la famille a des enfants des deux sexes », et B l'événement « la famille a au plus une fille ». On a

$$\mathbb{P}(A) = \tfrac{3}{4}, \qquad \mathbb{P}(B) = \tfrac{1}{2}, \qquad \mathbb{P}(A \cap B) = \tfrac{3}{8},$$

et donc A et B sont indépendants.

Faisons la même chose avec des familles de 4 enfants. Dans ce cas,

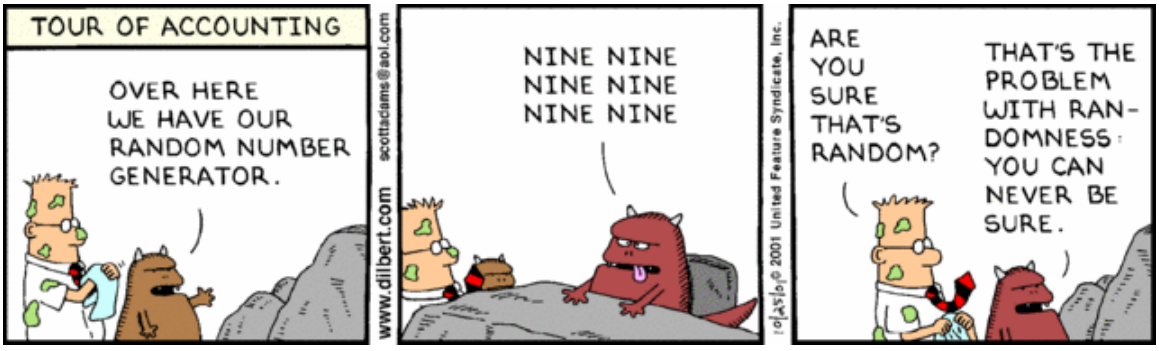
$$\mathbb{P}(A) = \tfrac{7}{8}, \qquad \mathbb{P}(B) = \tfrac{5}{16}, \qquad \mathbb{P}(A \cap B) = \tfrac{1}{4},$$

et donc A et B *ne sont pas* indépendants. ⌟

Définition 1.35. *Deux collections $\mathcal{G}_1, \mathcal{G}_2$ de parties de $\mathcal{F}$ sont **indépendantes** si*

$$\forall A_1 \in \mathcal{G}_1, A_2 \in \mathcal{G}_2, \qquad \mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2).$$

Définition 1.36. *Soit C un événement avec $\mathbb{P}(C) > 0$. Une famille d'événements $(A_i)_{i \in I}$ est **indépendante conditionnellement à C** sous $\mathbb{P}$ si elle est indépendante sous $\mathbb{P}(\cdot \mid C)$.*



©Scott Adams

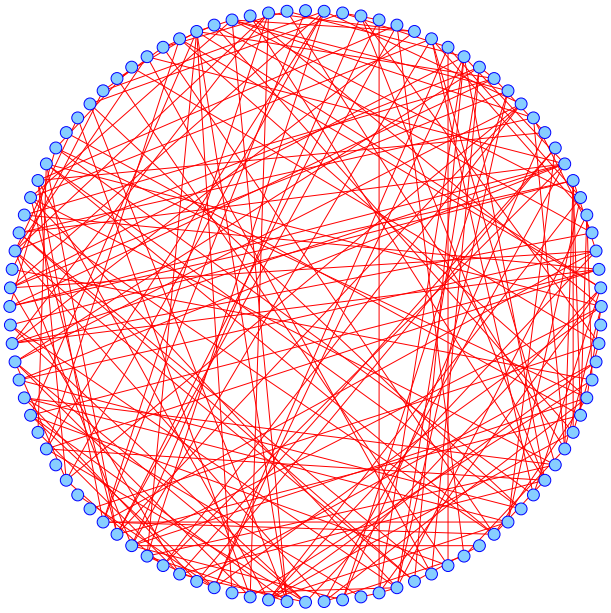
2 Variables aléatoires

Random variables are much like Guinea pigs, neither a pig, nor from Guinea.

dtldarek (sur math.stackexchange)

2.1 Variables aléatoires

Il est souvent plus pratique d'associer une (ou plusieurs) valeur numérique au résultat d'une expérience aléatoire, plutôt que de travailler directement avec une réalisation. Par exemple, lorsque n et m sont grands, une réalisation du graphe aléatoire $\mathcal{G}(n, m)$ de l'Exemple 1.9 est un objet trop complexe pour être directement intéressant ; à titre d'illustration, voici une réalisation du graphe aléatoire $\mathcal{G}(100, 200)$:



Dans un tel cas, il est en général plus utile de se concentrer sur certaines propriétés numériques de cette réalisation, comme, par exemple, le nombre d'arêtes incidentes en un sommet, le nombre de composantes connexes, ou la taille de la plus grande composante connexe.

Mathématiquement, étant donné un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, de telles valeurs numériques sont des fonctions $X : \Omega \to \mathbb{R}$ associant à un résultat de l'expérience une valeur dans $\mathbb{R}$. On s'intéresse à des

événements du type $\{\omega \in \Omega \,|\, X(\omega) \in A\}$, où $A \in \mathcal{B}(\mathbb{R})$ [1]. Pour que cela ait du sens, il faut que cet événement appartiennent à $\mathcal{F}$, c'est-à-dire il faut que l'on ait $X^{-1}(A) \in \mathcal{F}$ pour tout $A \in \mathcal{B}(\mathbb{R})$; en d'autres termes, on veut que l'application $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ soit mesurable.

Évidemment, il n'y a aucune raison de se limiter à des fonctions numériques. On peut considérer des fonctions prenant valeurs dans un espace mesurable $(E, \mathcal{E})$ arbitraire.

Définition 2.1. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. Une **variable aléatoire** est une application mesurable $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$.*

Lorsque $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, on parlera de **variable aléatoire réelle**, et lorsque $(E, \mathcal{E}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, on parlera de **vecteur aléatoire réel**.

Exemple 2.2. On considère le graphe aléatoire $\mathcal{G}(n, m)$. Pour chaque $k \in \mathbb{N}$, la fonction N_k donnant le nombre de sommets ayant k arêtes incidentes est une variable aléatoire réelle. Dans la réalisation de $\mathcal{G}(8, 6)$ de l'Exemple 1.9, on a $N_0 = 1$, $N_1 = 3$, $N_2 = 3$, $N_3 = 1$ et $N_k = 0$ pour les autres valeurs de k. ⌟

Exemple 2.3. L'expérience consiste à lancer un dé une infinité de fois et on s'intéresse à la variable aléatoire donnant le numéro du premier lancer lors duquel un 6 est sorti. Dans ce cas, $\Omega := \{\omega = (\omega_k)_{k \geqslant 1} \in \{1, 2, 3, 4, 5, 6\}^{\mathbb{N}^*}\}$, $\mathcal{F} := \sigma(\mathcal{C})$, où $\mathcal{C}$ est l'ensemble des cylindres, à savoir

$$\mathcal{C} := \bigcup_{n \geqslant 1} \bigcup_{s_1 \in \{1,2,3,4,5,6\}} \cdots \bigcup_{s_n \in \{1,2,3,4,5,6\}} \{\omega \in \Omega \,|\, \forall 1 \leqslant i \leqslant n,\, \omega_i = s_i\}.$$

(En d'autres termes, $\mathcal{C}$ est l'ensemble des événements ne dépendant que d'un nombre fini de lancers.)

La variable aléatoire $X : \Omega \to \mathbb{R} \cup \{+\infty\}$ donnant le premier lancer lors duquel un 6 est sorti est

$$X(\omega) := \inf\{n \in \mathbb{N}^* \,|\, \omega_n = 6\},$$

avec la convention que $\inf \varnothing = +\infty$. Il s'agit bien d'une variable aléatoire, puisque, pour tout $k \in \mathbb{N}^*$

$$X^{-1}(\{k\}) = \{\omega \in \Omega \,|\, \omega_1 \neq 6, \ldots, \omega_{k-1} \neq 6, \omega_k = 6\} \in \mathcal{C} \subset \mathcal{F}.$$

et

$$X^{-1}(\{+\infty\}) = \{\omega \in \Omega \,|\, \forall k \in \mathbb{N}^*,\, \omega_k \neq 6\} = \bigcap_{n \in \mathbb{N}^*} \{\omega \in \Omega \,|\, \forall k \in \{1, \ldots, n\},\, \omega_k \neq 6\} \in \mathcal{F}. \qquad ⌟$$

2.2 Loi d'une variable aléatoire

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité, $(E, \mathcal{E})$ un espace mesurable et $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ une variable aléatoire. Cette dernière induit naturellement [2] une mesure de probabilité sur $(E, \mathcal{E})$ via la mesure image

$$\forall A \in \mathcal{E}, \qquad X_*\mathbb{P}(A) := \mathbb{P}(X^{-1}(A)) \equiv \mathbb{P}(X \in A).$$

Définition 2.4. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité, $(E, \mathcal{E})$ un espace mesurable et $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ une variable aléatoire. La mesure de probabilité $\mathbb{P}_X$ sur $(E, \mathcal{E})$ définie par*

$$\forall A \in \mathcal{E}, \qquad \mathbb{P}_X(A) := \mathbb{P}(X \in A)$$

*est appelée la **loi** (ou la **distribution**) de X.*

Définition 2.5. *Deux variables aléatoires prenant valeurs dans le même espace sont **égales en loi** ou **identiquement distribuées** si $\mathbb{P}_X = \mathbb{P}_Y$. On écrit alors $X \overset{\text{loi}}{=} Y$.*

1. On veut certainement pouvoir considérer des événements de ce type lorsque A est un intervalle. Mais on veut également pouvoir considérer des intersections ou unions de tels événements, ce qui conduit naturellement à considérer la plus petite tribu contenant les intervalles, à savoir $\mathcal{B}(\mathbb{R})$.

2. La vérification est laissée en exercice.

Remarque 2.6. *Observez que l'égalité en loi de X et de Y n'implique pas que les variables aléatoires X et Y coïncident : il n'est même pas nécessaire qu'elles soient définies sur un même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$!*

Exemple 2.7. Considérons le lancer de deux dés non pipés, et notons X la variable aléatoire correspondant à la somme des valeurs obtenues. Alors, la probabilité que la somme appartienne à l'intervalle $[\sqrt{5}, \pi + 1]$ est donnée par

$$\mathbb{P}_X([\sqrt{5}, \pi + 1]) = \mathbb{P}(X \in \{3, 4\}) = \mathbb{P}(\{(1,2), (2,1), (1,3), (3,1), (2,2)\}) = \frac{5}{36}. \qquad \lrcorner$$

Exemple 2.8. Soit $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ une variable aléatoire. S'il existe un sous-ensemble $E \subset S$ fini ou dénombrable tel que $\mathbb{P}(X \in E) = 1$, on dit que X est une **variable aléatoire discrète**. Dans ce cas, la loi de X peut s'écrire

$$\mathbb{P}_X = \sum_{x \in E} p_x \delta_x,$$

où $p_x \coloneqq \mathbb{P}(X = x)$ et δ_x est la mesure de Dirac en x. En effet, pour tout $A \in \mathcal{S}$,

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \mathbb{P}(X \in A \cap E) = \mathbb{P}\Big(\bigcup_{x \in A \cap E} \{X = x\} \Big) = \sum_{x \in A \cap E} \mathbb{P}(X = x) = \sum_{x \in E} p_x \delta_x(A). \qquad \lrcorner$$

Exemple 2.9. Lorsque $X : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ satisfait $\mathbb{P}(X = x) = 0$ pour tout $x \in \mathbb{R}^n$, on dit que X est une **variable aléatoire continue**. Si, de plus, la loi de X est absolument continue par rapport à la mesure de Lebesgue, on dit que X est une **variable aléatoire à densité**. Par le théorème de Radon–Nikodym, il existe alors une fonction intégrable $f_X : \mathbb{R}^n \to \mathbb{R}_+$ telle que

$$\forall A \in \mathcal{B}(\mathbb{R}^n), \qquad \mathbb{P}_X(A) = \int_A f_X \mathrm{d}\lambda_n.$$

La fonction f_X est appelée la **densité (de probabilité)** de X. Comme discuté dans la Remarque 1.12, deux densités distinctes de la variable aléatoire X coïncident presque partout. $\lrcorner$

Exemple 2.10. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. On a parfois envie de parler d'un **élément aléatoire** X tiré au hasard dans Ω selon la mesure de probabilité $\mathbb{P}$. Un tel élément aléatoire peut être vu comme une variable aléatoire $X : (\Omega, \mathcal{F}) \to (\Omega, \mathcal{F})$, donnée par $X(\omega) \coloneqq \omega$. Évidemment, dans ce cas, on $\mathbb{P}_X = \mathbb{P}$. $\lrcorner$

Remarque 2.11. *La mesure de probabilité $\mathbb{P}_X$ contient toute l'information nécessaire pour étudier les propriétés statistiques de la variable aléatoire X (c'est-à-dire, la façon dont sont distribués les résultats de mesures successives de X, effectuées indépendamment et dans les mêmes conditions). En particulier, si l'on n'est intéressé que par cette variable aléatoire, l'espace de probabilité de départ $(\Omega, \mathcal{F}, \mathbb{P})$ peut être complétement ignoré, et n'est souvent même pas spécifié, l'espace de probabilité pertinent étant $(E, \mathcal{E}, \mathbb{P}_X)$. Insistons également sur le fait que la fonction $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ seule ne suffit pas à caractériser une variable aléatoire : la connaissance de la mesure $\mathbb{P}$ sur $(\Omega, \mathcal{F})$ ou de la loi $\mathbb{P}_X$ sur $(E, \mathcal{E})$ est indispensable.*

2.2.1 Quelques exemples de lois discrètes

On présente ici quelques-unes des lois discrètes les plus importantes.

Loi de Bernoulli

Soit $p \in [0, 1]$, $E \coloneqq \{0, 1\}$ et $\mathcal{E} \coloneqq \mathscr{P}(E)$. La variable aléatoire $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ suit une **loi de Bernoulli** de paramètre $p \in [0, 1]$ si

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

Dans ce cas, on écrira $X \sim \mathsf{bernoulli}(p)$.

On parle souvent d'**épreuve de Bernoulli**, et les événements $\{X = 1\}$ et $\{X = 0\}$ sont respectivement appelés **succès** et **échec**.

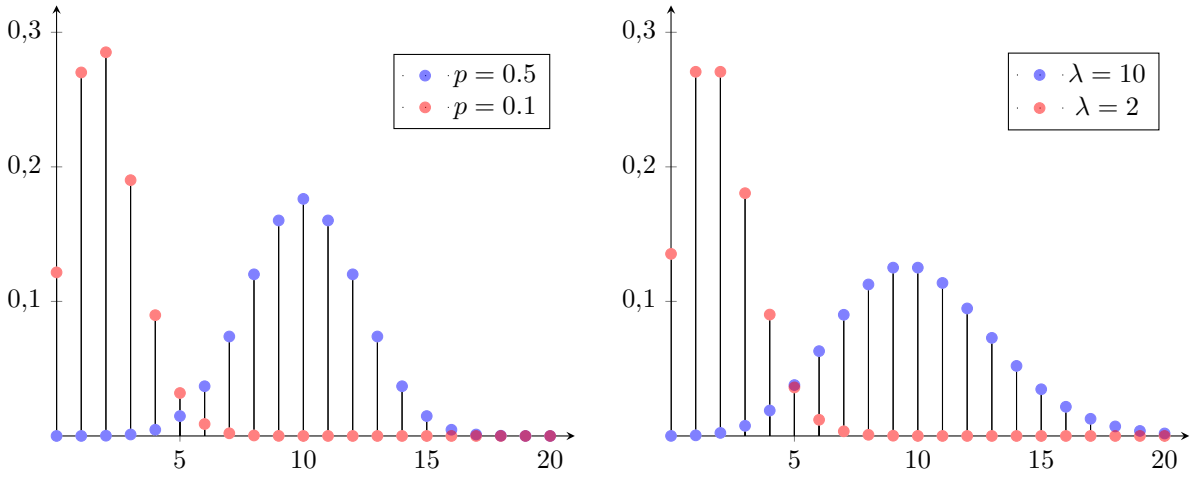


FIGURE 2.1: Gauche : loi $\mathsf{binom}(20, p)$. Droite : loi $\mathsf{poisson}(\lambda)$.

Exemple 2.12.

1. Un lancer à pile ou face est une épreuve de Bernoulli (avec, par exemple, $X(\mathsf{P}) = 1$ et $X(\mathsf{F}) = 0$).
2. Pour tout $A \in \mathcal{F}$, la **fonction indicatrice** de A, $\mathbf{1}_A : \Omega \to \{0, 1\}$, définie par

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{si } \omega \in A, \\ 0 & \text{si } \omega \notin A, \end{cases}$$

est une variable aléatoire discrète suivant une loi de Bernoulli de paramètre $\mathbb{P}(A)$. ⌟

Loi binomiale

Soit $n \in \mathbb{N}^*$, $p \in [0, 1]$, $E \coloneqq \{0, 1, \dots, n\}$ et $\mathcal{E} \coloneqq \mathscr{P}(E)$. La variable aléatoire $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ suit une **loi binomiale** de paramètres n et p si

$$\forall k \in \{0, \dots, n\}, \qquad \mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

Dans ce cas, on écrira $X \sim \mathsf{binom}(n, p)$.

Exemple 2.13. Répétons n fois de manière indépendante une épreuve de Bernoulli de paramètre p, et notons X la variable aléatoire représentant le nombre de succès obtenus à l'issue des n épreuves. Puisqu'il y a $\binom{n}{k}$ façons d'obtenir k succès sur n épreuves, on voit que $X \sim \mathsf{binom}(n, p)$. ⌟

Loi de Poisson

Soit $\lambda \in \mathbb{R}_+^*$, $E \coloneqq \mathbb{N}$ et $\mathcal{E} \coloneqq \mathscr{P}(E)$. La variable aléatoire $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ suit une **loi de Poisson**[3] de paramètre λ si

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

Dans ce cas, on écrira $X \sim \mathsf{poisson}(\lambda)$.

Exemple 2.14. Considérons une variable aléatoire $X \sim \mathsf{binom}(n, p)$ avec n très grand et p très petit (modélisant par exemple la transmission d'un gros fichier via internet : n est la taille en bits du fichier, et p la probabilité qu'un bit donné soit modifié au cours de la transmission). Alors X suit approximativement une loi

3. Siméon Denis Poisson (1781, Pithiviers – 1840, Sceaux), mathématicien, géomètre et physicien français.

de Poisson de paramètre $\lambda = np$ (c'est ce qu'on appelle parfois la **loi des petits nombres**). Plus précisément, considérons à présent une suite $(p_n)_{n\in\mathbb{N}^*} \subset [0,1]$ telle que $\lim_{n\to\infty} np_n = \lambda > 0$. Pour chaque $n \in \mathbb{N}^*$, soit $X_n \sim \mathsf{binom}(n, p_n)$, et soit $X \sim \mathsf{poisson}(\lambda)$. Alors, pour tout $k \in \mathbb{N}$ et tout $n \geqslant k$,

$$\mathbb{P}(X_n = k) = \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{1}{k!}\frac{n}{n}\frac{n-1}{n}\frac{n-2}{n}\cdots\frac{n-k+1}{n}(np_n)^k(1-p_n)^{n-k}.$$

En prenant, à k fixé, la limite $n \to \infty$, on voit que chacun des quotients $(n-i)/n$ converge vers 1, que $(np_n)^k$ converge vers λ^k, que $(1-p_n)^n$ converge vers $e^{-\lambda}$ et que $(1-p_n)^{-k}$ tend vers 1. Par conséquent,

$$\forall k \in \mathbb{N}, \qquad \lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k). \qquad \lrcorner$$

Loi géométrique

Soit $p \in [0,1]$, $E \coloneqq \mathbb{N}^*$ et $\mathcal{E} \coloneqq \mathscr{P}(E)$. La variable aléatoire $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ suit une **loi géométrique** de paramètre p si[4]

$$\forall k \in \mathbb{N}^*, \qquad \mathbb{P}(X = k) = p(1-p)^{k-1}.$$

Dans ce cas, on écrira $X \sim \mathsf{géom}(p)$.

Exemple 2.15. Répétons de façon indépendante une épreuve de Bernoulli de paramètre p jusqu'à ce que le premier succès ait lieu. La variable aléatoire X correspondant au numéro du lancer auquel a lieu ce premier succès suit alors clairement une loi géométrique de paramètre p. $\lrcorner$

Une propriété remarquable de la loi géométrique est sa **perte de mémoire**.

Lemme 2.16. *Soit $X \sim \mathsf{géom}(p)$. Alors, pour tout $k \geqslant 1$,*

$$\forall n \in \mathbb{N}^*, \qquad \mathbb{P}(X = n + k \,|\, X > n) = \mathbb{P}(X = k).$$

Démonstration. On a

$$\mathbb{P}(X = n+k \,|\, X > n) = \frac{\mathbb{P}(X = n+k)}{\mathbb{P}(X > n)} = \frac{p(1-p)^{n+k-1}}{\sum_{m>n} p(1-p)^{m-1}},$$

et le dénominateur est égal à $(1-p)^n \sum_{m>0} p(1-p)^{m-1} = (1-p)^n$. $\square$

Cette propriété affirme par exemple que même si le numéro 53 (sur 90 numéros possibles) n'est pas sorti pendant 182 tirages consécutifs à la loterie, cela ne rend pas sa prochaine apparition plus probable[5].

Loi hypergéométrique

Soit $N \in \mathbb{N}^*$, $K \in \{0, 1, \ldots, N\}$, $n \in \{0, 1, \ldots, N\}$, $E \coloneqq \{(n+K-N)\vee 0, \ldots, n \wedge K\}$[6] et $\mathcal{E} \coloneqq \mathscr{P}(E)$. La variable aléatoire $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ suit une **loi hypergéométrique** de paramètres N, K et n si

$$\forall \ell \in E, \qquad \mathbb{P}(X = \ell) = \frac{\binom{K}{\ell}\binom{N-K}{n-\ell}}{\binom{N}{n}}.$$

Dans ce cas, on écrira $X \sim \mathsf{hypergeom}(N, K, n)$.

4. Attention, il existe malheureusement deux conventions distinctes pour la loi géométrique : celle donnée ci-dessus et celle définissant cette loi sur $E \coloneqq \mathbb{N}$ via $\mathbb{P}(X = k) = p(1-p)^k$ pour tout $k \in \mathbb{N}$.

5. Cela s'est produit en 2005 en Italie. Le numéro 53 n'est pas sorti pendant 182 tirages consécutifs à la loterie de Venise, ce qui a conduit de très, très nombreux Italiens à miser de grosses sommes, certains tout ce qu'ils possédaient. Le total des mises s'est élevé à plus de 4 milliards d'euros, et cette histoire s'est terminée par de nombreuses ruines et même des suicides...

6. On utilise les notations usuelles : $a \vee b = \max(a, b)$ et $a \wedge b = \min(a, b)$.

Exemple 2.17. Une urne contient N boules, dont b sont bleues et $r = N - b$ sont rouges. Un échantillon de $n \leqslant N$ boules est tiré de l'urne, sans remise. On vérifie facilement que le nombre B de boules bleues dans l'échantillon suit la loi hypergéométrique de paramètres N, b et n, $B \sim \mathsf{hypergeom}(N, b, n)$. ⌟

Lemme 2.18. *Soit $n \in \mathbb{N}$ et soit $(K_N)_{N\in\mathbb{N}^*}$ tel que $0 \leqslant K_N \leqslant N$ pour tout $N \geqslant 1$ et $\lim_{N\to\infty} K_N/N = p$. Pour chaque $N \in \mathbb{N}^*$ tel que $N \geqslant n$, soit $X_N \sim \mathsf{hypergeom}(N, K_N, n)$, et soit $X \sim \mathsf{binom}(n, p)$. Alors,*

$$\forall \ell \in \{0, \dots, n\}, \qquad \lim_{N\to\infty} \mathbb{P}(X_N = \ell) = \mathbb{P}(X = \ell).$$

Démonstration. Exercice. □

Ce lemme montre qu'il est possible de remplacer la loi hypergéométrique de paramètres N, K et n par une loi binomiale de paramètres n et $p := K/N$ dès que la taille n de l'échantillon est suffisamment petite par rapport à la taille N de la population. Ceci est intuitif, puisque si l'on effectue un tirage avec remise d'un petit échantillon à partir d'une grande population, il y a très peu de chances de tirer le même individu deux fois... Dans la pratique, on remplace la loi hypergéométrique dès que $10n < N$. Un exemple classique concerne les sondages. On considère fréquemment le sondage de n personnes comme n sondages indépendants alors qu'en réalité le sondage est exhaustif (on n'interroge jamais deux fois la même personne). Comme n (nombre de personnes interrogées) est beaucoup plus petite que N(population sondée), cette approximation est légitime.

2.2.2 Quelques exemples de lois à densité

On présente ici quelques-unes des lois lois à densité les plus importantes.

Loi exponentielle

Soit $\lambda \in \mathbb{R}_+^*$. La variable aléatoire $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ suit la **loi exponentielle** de paramètre λ si sa loi est absolument continue par rapport à Lebesgue, avec densité de probabilité

$$\lambda\, e^{-\lambda x}\, \mathbf{1}_{[0,\infty)}(x).$$

Dans ce cas, on écrira $X \sim \mathsf{exp}(\lambda)$.

Cette loi joue un rôle central dans la théorie des processus markoviens à temps continu.

Exemple 2.19. La loi exponentielle peut être vue comme limite de la distribution géométrique, et apparaît dans la pratique pour la description du temps d'attente entre deux événements imprédictibles (appels téléphoniques, tremblements de terre, émission de particules par désintégration radioactive, etc.). Considérons une suite d'épreuves de Bernoulli effectuées aux temps $\delta, 2\delta, 3\delta, \dots$ et soit W le temps du premier succès. Alors, pour tout $k \in \mathbb{N}$,

$$\mathbb{P}(W > k\delta) = (1 - p)^k.$$

Fixons à présent un temps $t > 0$. Jusqu'au temps t, il y aura eu $k = \lfloor t/\delta \rfloor$ épreuves. On veut laisser δ tendre vers 0. Afin d'obtenir un résultat qui ne soit pas trivial, on considère p de la forme $p = \lambda\delta$ pour une constante $\lambda > 0$. Dans ce cas,

$$\mathbb{P}(W > t) = \mathbb{P}(W > \lfloor t/\delta \rfloor \delta) = (1 - \lambda\delta)^{\lfloor t/\delta \rfloor} \to e^{-\lambda t}.$$ ⌟

Il est aussi aisé de vérifier (exercice) que la loi exponentielle possède la même propriété de perte de mémoire que la loi géométrique (Lemme 2.16) : $\mathbb{P}(X > s + t \,|\, X > s) = \mathbb{P}(X > t)$ pour tout, $s, t \geqslant 0$.

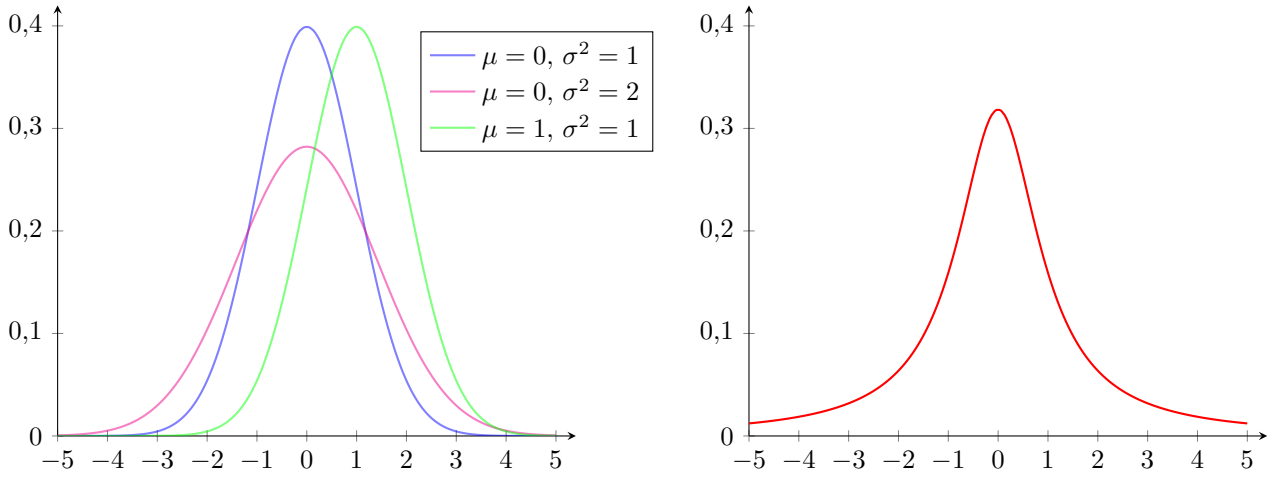


FIGURE 2.2: Gauche : densité de probabilité de la loi normale. Droite : densité de probabilité de la loi de Cauchy.

Loi normale

Soit $\mu \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}_+^*$. La variable aléatoire $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ suit la **loi normale** de paramètres μ et σ^2 si sa loi est absolument continue par rapport à Lebesgue, avec densité de probabilité

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big).$$

Dans ce cas, on écrira $X \sim \mathcal{N}(\mu, \sigma^2)$.

Il s'agit sans doute de la loi la plus importante, de par son ubiquité (à cause du théorème central limite, que l'on étudiera plus tard). Lorsque $\mu = 0$ et $\sigma^2 = 1$, on parle de **loi normale standard** ou de loi normale **centrée réduite**.

Loi de Cauchy

La variable aléatoire $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ suit la **loi de Cauchy**[7] si sa loi est absolument continue par rapport à Lebesgue, avec densité de probabilité

$$\frac{1}{\pi(1+x^2)}.$$

Dans ce cas, on écrira $X \sim \mathsf{cauchy}$.

2.3 Parenthèse technique

Définition 2.20. *Une collection $\mathcal{C}$ de parties d'un ensemble Ω est un **π-système** si elle est stable par intersection, c'est-à-dire si $A \cap B \in \mathcal{C}$ pour tout $A, B \in \mathcal{C}$.*

*Une collection $\mathcal{M}$ de parties d'un ensemble Ω est une **classe monotone** si elle possède les propriétés suivantes :*

- $\Omega \in \mathcal{M}$;
- $B \setminus A \in \mathcal{M}$, *pour tout* $A, B \in \mathcal{M}$ *tels que* $A \subset B$;
- $\bigcup_{i \geqslant 1} A_i \in \mathcal{M}$, *pour toute suite* $A_1, A_2, \ldots \in \mathcal{M}$ *d'ensembles 2 à 2 disjoints.*

7. Augustin Louis, baron Cauchy (1789, Paris – 1857, Sceaux), mathématicien français.

Étant donné $\mathcal{C} \subset \mathscr{P}(\Omega)$, on peut considérer la collection $M(\mathcal{C})$ de toutes les classes monotones contenant $\mathcal{C}$; cette collection est non-vide puisqu'elle contient $\mathscr{P}(\Omega)$. Comme l'intersection de classes monotones est encore une classe monotone, on peut définir la **classe monotone $\mathcal{M}(\mathcal{C})$ engendrée par $\mathcal{C}$** par

$$\mathcal{M}(\mathcal{C}) \coloneqq \bigcap_{\mathcal{M} \in M(\mathcal{C})} \mathcal{M}.$$

Manifestement, toute tribu est une classe monotone. En particulier, $\sigma(\mathcal{C})$ est une classe monotone contenant $\mathcal{C}$. Par conséquent, $\sigma(\mathcal{C}) \supset \mathcal{M}(\mathcal{C})$. Le résultat suivant montre que si $\mathcal{C}$ est un π-système, alors il y a en fait égalité.

Lemme 2.21 (Lemme des classes monotones)**.** *Soit $\mathcal{C}$ un π-système. Alors, $\sigma(\mathcal{C}) = \mathcal{M}(\mathcal{C})$.*

Démonstration. Il suffit de montrer que $\mathcal{M}(\mathcal{C})$ est une tribu, puisque cela impliquera alors que $\sigma(\mathcal{C}) \subset \mathcal{M}(\mathcal{C})$.

La collection $\mathcal{D}_1 \coloneqq \{A \subset \Omega \mid A \cap B \in \mathcal{M}(\mathcal{C}),\ \forall B \in \mathcal{C}\}$ est manifestement une classe monotone. De plus, $\mathcal{C}$ étant un π-système, $\mathcal{D}_1 \supset \mathcal{C}$. Par conséquent, $\mathcal{D}_1 \supset \mathcal{M}(\mathcal{C})$, par minimalité de $\mathcal{M}(\mathcal{C})$. Ceci signifie que $A \cap B \in \mathcal{M}(\mathcal{C})$ pour tout $A \in \mathcal{M}(\mathcal{C})$ et $B \in \mathcal{C}$.

De façon similaire, la collection $\mathcal{D}_2 \coloneqq \{A \subset \Omega \mid A \cap B \in \mathcal{M}(\mathcal{C}),\ \forall B \in \mathcal{M}(\mathcal{C})\}$ est également une classe monotone et, par l'observation précédente, $\mathcal{D}_2 \supset \mathcal{C}$. Par conséquent, $\mathcal{D}_2 \supset \mathcal{M}(\mathcal{C})$, c'est-à-dire $A \cap B \in \mathcal{M}(\mathcal{C})$ pour tout $A, B \in \mathcal{M}(\mathcal{C})$. $\mathcal{M}(\mathcal{C})$ est donc un π-système.

Montrons à présent que $\mathcal{M}(\mathcal{C})$ est en fait une tribu. Soit $A_1, A_2 \ldots \in \mathcal{M}(\mathcal{C})$. $\mathcal{M}(\mathcal{C})$ étant un π-système, il suit que les ensembles

$$B_i \coloneqq A_i \setminus \bigcup_{j<i} A_j = A_i \cap \bigcap_{j<i} (\Omega \setminus A_j)$$

appartiennent également à $\mathcal{M}(\mathcal{C})$ et sont disjoints deux à deux. On en conclut que $\bigcup_{i \geqslant 1} A_i = \bigcup_{i \geqslant 1} B_i \in \mathcal{M}(\mathcal{C})$, ce qui termine la preuve. □

Le lemme précédent permet de démontrer le résultat suivant, fort utile.

Théorème 2.22 (Théorème d'unicité)**.** *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $\mathcal{G} \subset \mathscr{P}(\Omega)$ un π-système tel que $\mathcal{F} = \sigma(\mathcal{G})$. Alors, $\mathbb{P}$ est complètement déterminée par sa restriction $\mathbb{P}_{|\mathcal{G}}$ à $\mathcal{G}$.*

Démonstration. Soit $\mathbb{Q}$ une mesure de probabilité sur $(\Omega, \mathcal{F})$ telle que $\mathbb{Q}_{|\mathcal{G}} = \mathbb{P}_{|\mathcal{G}}$. On vérifie aisément que la collection $\mathcal{D} \coloneqq \{A \in \mathcal{F} \mid \mathbb{P}(A) = \mathbb{Q}(A)\}$ est une classe monotone. De plus, $\mathcal{D} \supset \mathcal{G}$ par hypothèse. Il suit donc de la minimalité de $\mathcal{M}(\mathcal{G})$ que $\mathcal{D} \supset \mathcal{M}(\mathcal{G})$.

$\mathcal{G}$ étant un π-système, le Lemme 2.21 implique que $\mathcal{F} = \sigma(\mathcal{G}) = \mathcal{M}(\mathcal{G}) \subset \mathcal{D}$, et donc que $\mathcal{D} = \mathcal{F}$. On en conclut que $\mathbb{P} = \mathbb{Q}$. □

2.4 Fonction de répartition

L'ensemble $\mathcal{C} \coloneqq \{(-\infty, x] \mid x \in \mathbb{R}\}$ forme un π-système tel que $\sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R})$. Il suit donc du Théorème d'unicité 2.22 qu'une mesure de probabilité $\mathbb{P}$ sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ est entièrement déterminée par la fonction $F_{\mathbb{P}}(x) = \mathbb{P}\big((-\infty, x]\big)$. En particulier, la loi d'une variable aléatoire réelle X est entièrement déterminée par la fonction $F_X(x) = \mathbb{P}(X \leqslant x)$.

Définition 2.23. *Soit X une variable aléatoire réelle. La fonction $F_X : \mathbb{R} \to [0, 1]$ définie par*

$$F_X(x) \coloneqq \mathbb{P}(X \leqslant x)$$

*est appelée la **fonction de répartition** de X.*

Remarque 2.24. *Il est traditionnel de noter Φ la fonction de répartition associée à la loi $\mathcal{N}(0, 1)$ et nous le ferons par la suite.*

Lemme 2.25. *Soit X une variable aléatoire réelle. Alors, sa fonction de répartition F_X possède les propriétés suivantes :*

1. *F_X est croissante ;*
2. *$\lim_{x\to-\infty} F_X(x) = 0$;*
3. *$\lim_{x\to+\infty} F_X(x) = 1$;*
4. *F_X est continue à droite.*

Démonstration. *1.* Cette affirmation suit de l'inclusion $\{X \leqslant x\} \subset \{X \leqslant y\}$ lorsque $x \leqslant y$.
2. Soit $(A_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ définie par $A_n := \{X \leqslant -n\}$. Alors $(A_n)_{n\in\mathbb{N}}$ est une suite décroissante et $\bigcap_{n\in\mathbb{N}} A_n = \varnothing$. Le Lemme 1.4 implique donc que

$$\lim_{n\to\infty} F_X(-n) = \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(\varnothing) = 0.$$

La conclusion suit de la monotonicité de F_X.
3. Soit $(B_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ définie par $B_n := \{X \leqslant n\}$. Alors $(B_n)_{n\in\mathbb{N}}$ est une suite croissante et $\bigcup_{n\in\mathbb{N}} B_n = \Omega$. Le Lemme 1.4 implique donc que

$$\lim_{n\to\infty} F_X(n) = \lim_{n\to\infty} \mathbb{P}(B_n) = \mathbb{P}(\Omega) = 1.$$

La conclusion suit de la monotonicité de F_X.
4. Soit $(x_n)_{n\in\mathbb{N}}$ une suite décroissante convergeant vers x et $C_n := \{X \leqslant x_n\}$. La suite $(C_n)_{n\in\mathbb{N}}$ est décroissante et $\lim_{n\to\infty} C_n = \{X \leqslant x\}$. Il suit donc du Lemme 1.4 que

$$\lim_{n\to\infty} F_X(x_n) = \lim_{n\to\infty} \mathbb{P}(C_n) = \mathbb{P}\big(\lim_{n\to\infty} C_n\big) = \mathbb{P}\big(X \leqslant x\big) = F_X(x). \qquad \square$$

Le lemme suivant montre que ces quatre propriétés caractérisent les fonctions de répartition. En particulier, l'ensemble des mesures de probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ est en bijection avec l'ensemble des fonctions de ce type.

Lemme 2.26. *Soit $F : \mathbb{R} \to \mathbb{R}$ une fonction possédant les quatre propriétés du Lemme 2.25. Alors, il existe une variable aléatoire réelle X sur $((0,1), \mathcal{B}((0,1)), \lambda)$ telle que $F_X = F$. Cette variable aléatoire est explicitement donnée par*

$$\forall u \in (0,1), \qquad X(u) = \inf\{c \in \mathbb{R} \,|\, F(c) \geqslant u\}.$$

Démonstration. Les propriétés 2 et 3 impliquent que $X(u) \in (-\infty, \infty)$ pour tout $u \in (0,1)$. De plus, l'infimum est en fait un minimum, puisque F est continue à droite. Ceci implique que $X(u) \leqslant c$ si et seulement si $u \leqslant F(c)$. En particulier, $\{u \in (0,1) \,|\, X(u) \leqslant c\} = (0, F(c)] \cap (0,1) \in \mathcal{B}((0,1))$. On en conclut que X est bien une variable aléatoire réelle et $F_X(c) = \lambda\big((0, F(c)] \cap (0,1)\big) = F(c)$ pour tout $c \in \mathbb{R}$. $\square$

Le résultat suivant énonce quelques propriétés élémentaires de la variable aléatoire X pouvant être lue directement de sa fonction de répartition. En particulier, il montre que les discontinuités de F_X correspondent précisément aux **atomes** de la loi $\mathbb{P}_X$, c'est-à-dire aux singletons de probabilité strictement positive (et la hauteur du saut est égale à la probabilité de l'atome correspondant).

Lemme 2.27. *Soit X une variable aléatoire réelle de fonction de répartition F_X. Alors,*

1. *$\mathbb{P}(X > x) = 1 - F_X(x)$, pour tout $x \in \mathbb{R}$.*
2. *$\mathbb{P}(x < X \leqslant y) = F_X(y) - F_X(x)$, pour tout $x < y$ dans $\mathbb{R}$.*
3. *$\mathbb{P}(X = x) = F_X(x) - \lim_{y\uparrow x} F_X(y)$ pour tout $x \in \mathbb{R}$.*

Démonstration. Les deux premières affirmations sont immédiates. Pour la troisième, soit $(x_n)_{n\geqslant 1}$ une suit croissante convergeant vers x et $A_n = \{x_n < X \leqslant x\}$. Puisque $\lim_{n\to\infty} A_n = \{X = x\}$, il suit du Lemme 1.4 que

$$\mathbb{P}(X = x) = \lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \big(F_X(x) - F_X(x_n)\big),$$

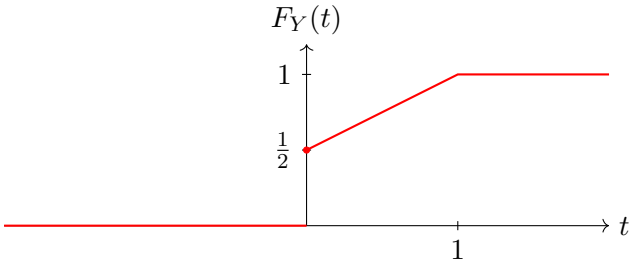
par le point *2.* (la limite existe par monotonicité de F_X). $\square$

Il suit du lemme précédent que la fonction de répartition associée à une variable aléatoire réelle X continue est continue, puisque $\mathbb{P}(X = x) = 0$ pour tout $x \in \mathbb{R}$. Si X est absolument continue, de densité f_X, alors sa fonction de répartition satisfait

$$F_X(x) = \mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f_X(s)\,\mathrm{d}\lambda(s).$$

Remarque 2.28. *On peut démontrer (c'est le Théorème de différentiation de Lebesgue) que, dans ce cas, F_X est différentiable presque partout. Ceci permet d'associer à X une densité de probabilité de façon canonique : on prendra $f_X(x) = F_X'(x)$ en tout point où F_X est différentiable et $f_X(x) = 0$ ailleurs.*

Exemple 2.29. Soit $X \sim \mathrm{Bern}(\frac{1}{2})$ et $U \sim \mathsf{U}(0,1)$ deux variables aléatoires telles que les événements $\{X = a\}$ et $\{U \leqslant u\}$ sont indépendants pour tout $a \in \{0,1\}$ et $u \in \mathbb{R}$. La fonction de répartition de $Y \coloneqq XU$ est



⌟

2.5 Tribu engendrée par une variable aléatoire

Définition 2.30. *Soit X une variable aléatoire à valeurs dans $(E, \mathcal{E})$. La **tribu engendrée par X** est la tribu*[8]

$$\sigma(X) \coloneqq \{X^{-1}(A) \mid A \in \mathcal{E}\}.$$

Le résultat suivant montre qu'une variable aléatoire réelle Y est $\sigma(X)$-mesurable précisément si Y est une fonction de X.

Lemme 2.31. *Soit $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ et $Y : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ deux variables aléatoires. Les deux assertions suivantes sont équivalentes :*

1. *Y est $\sigma(X)$-mesurable.*
2. *Il existe $\varphi : (E, \mathcal{E}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mesurable telle que $Y = \varphi(X)$.*

Démonstration. L'implication *2* $\Rightarrow$ *1* étant immédiate, on ne démontre que *1* $\Rightarrow$ *2*. Supposons tout d'abord que Y est une fonction étagée : $Y = \sum_{i=1}^{n} y_i \mathbf{1}_{C_i}$ avec $y_i \in \mathbb{R}$ (tous distincts) et $C_i \in \sigma(X)$ pour tout $1 \leqslant i \leqslant n$. Alors, pour chaque i, il existe $A_i \in \mathcal{E}$ tel que $C_i = X^{-1}(A_i)$. On a donc

$$Y = \sum_{i=1}^{n} y_i \mathbf{1}_{C_i} = \sum_{i=1}^{n} y_i \mathbf{1}_{X^{-1}(A_i)} = \sum_{i=1}^{n} y_i \mathbf{1}_{A_i} \circ X = \varphi \circ X,$$

où l'on a posé $\varphi \coloneqq \sum_{i=1}^{n} y_i \mathbf{1}_{A_i}$. Lorsque Y est positive, il existe une suite croissante de de fonctions Y_n étagées et $\sigma(X)$-mesurables. On sait qu'il existe alors des fonctions mesurables $\varphi_n : E \to \mathbb{R}$ mesurable telle que $Y_n = \varphi_n(X)$. Posons $\varphi(x) \coloneqq \limsup_{n\to\infty} \varphi_n(x)$. Par construction, φ est mesurable. De plus, pour tout $\omega \in \Omega$, la limite $\lim_{n\to\infty} \varphi_n(X(\omega))$ existe, puisque $\varphi_n(X(\omega)) = Y_n(\omega) \to Y(\omega)$; on a donc

$$\varphi(X(\omega)) = \lim_{n\to\infty} \varphi_n(X(\omega)) = \lim_{n\to\infty} Y_n(\omega) = Y(\omega).$$

Le cas où Y n'est pas nécessairement positive s'en déduit en posant $Y = Y_+ - Y_-$. □

8. La vérification qu'il s'agit bien d'une tribu est laissée en exercice.

2.6 Loi conjointe

Soit X et Y deux variables aléatoires. Les lois $\mathbb{P}_X$ et $\mathbb{P}_Y$ contiennent toute l'information nécessaire à une étude statistique de chacune de ces variables. Par contre, elles ne fournissent aucune information sur leurs propriétés relativement l'une à l'autre. Par exemple, l'événement $\{X = Y\}$ ne fait à priori aucun sens, les deux variables aléatoires n'étant même pas forcément définies sur un même espace de probabilité ! Et même lorsqu'elles sont définies sur un même espace $(\Omega, \mathcal{F}, \mathbb{P})$, le calcul de cette probabilité requiert de pouvoir déterminer $\mathbb{P}(X = x \text{ et } Y = x)$, alors qu'on n'a d'information que sur $\mathbb{P}(X = x)$ et $\mathbb{P}(Y = x)$ *séparément.*

Exemple 2.32. On demande à deux élèves de faire deux jets à pile ou face chacun (à l'aide d'une pièce équilibrée), et de relever les résultats. L'élève appliqué jette deux fois la pièce, obtenant une paire (X_1, X_2). L'élève paresseux ne jette la pièce qu'une fois et écrit le résultat deux fois, obtenant une paire (Y_1, Y_2) avec $Y_1 = Y_2$. Il est clair que X_1, X_2, Y_1, Y_2 sont toutes des variables aléatoires de même loi. Or ces couples ont des propriétés statistiques très différentes : $\mathbb{P}(X_1 = X_2) = \frac{1}{2}$, $\mathbb{P}(Y_1 = Y_2) = 1$. ⌟

Une façon de résoudre ce problème est de considérer X et Y non pas comme deux variables aléatoires, mais comme les composantes d'un vecteur aléatoire (X, Y). La loi de ce vecteur contient, elle, toute l'information sur le comportement statistique de la paire (X, Y).

Définition 2.33. *Soit $X_1, \dots, X_n$ des variables aléatoires définies sur un même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. La loi du vecteur $(X_1, \dots, X_n)$ est appelée la **loi conjointe** des variables aléatoires $X_1, \dots, X_n$.*

La loi conjointe contient en particulier toute l'information sur la loi de chacune des variables aléatoires impliquées.

Lemme 2.34. *Soit $X_1, \dots, X_n$ des variables aléatoires $X_i : (\Omega, \mathcal{F}) \to (E_i, \mathcal{E}_i)$ définies sur un même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. Alors, pour tout $k \in \{1, \dots, n\}$ et tout $A \in \mathcal{E}_k$,*

$$\mathbb{P}_{X_k}(A) = \mathbb{P}_{(X_1,\dots,X_n)}(E_1 \times \cdots \times E_{k-1} \times A \times E_{k+1} \times \cdots \times E_n).$$

La loi d'une des composantes (ou la loi conjointe d'un ensemble de composantes) d'un vecteur aléatoire $(X_1, \dots, X_n)$ est appelée une loi **marginale** de $\mathbb{P}_{(X_1,\dots,X_n)}$.

Démonstration. Il suffit d'observer que

$$\begin{aligned}\mathbb{P}_{(X_1,\dots,X_n)}&(E_1 \times \cdots \times E_{k-1} \times A \times E_{k+1} \times \cdots \times E_n) \\ &= \mathbb{P}(X_1 \in E_1, \dots, X_{k-1} \in E_{k-1}, X_k \in A, X_{k+1} \in E_{k+1}, \dots, X_n \in E_n). \quad \square\end{aligned}$$

Définition 2.35. *Soit $X_1, \dots, X_n$ des variables aléatoires réelles définies sur un même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. Si la loi conjointe $\mathbb{P}_{(X_1,\dots,X_n)}$ est absolument continue par rapport à λ_n, on appelle la densité de probabilité $f_{(X_1,\dots,X_n)}$ qui lui est associée la **densité de probabilité conjointe** de $X_1, \dots, X_n$.*

2.7 Indépendance de variables aléatoires

Rappelons que deux événements A et B sont indépendants si l'occurrence de A ne modifie pas la probabilité de réalisation de B ; mathématiquement, nous avons traduit cela par la propriété $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Nous aimerions à présent définir une notion similaire d'indépendance entre deux variables aléatoires, correspondant à l'idée intuitive que la connaissance de la valeur prise par une variable aléatoire n'a pas d'influence sur la distribution de l'autre variable aléatoire.

Définition 2.36. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. Soit I un ensemble arbitraire et, pour chaque $i \in I$, soit $X_i : (\Omega, \mathcal{F}) \to (E_i, \mathcal{E}_i)$ une variable aléatoire. Les variables aléatoires $(X_i)_{i \in I}$ sont **indépendantes** si, pour tout $J \subset I$ fini et toute collection $A_i \in \mathcal{E}_i$, $i \in J$,*

$$\mathbb{P}(\forall i \in J,\ X_i \in A_i) = \prod_{i \in J} \mathbb{P}(X_i \in A_i).$$

Intuitivement, si l'information procurée par une variable aléatoire X ne nous renseigne pas sur une autre variable aléatoire Y, il devrait en être de même pour des fonctions de X et de Y. C'est ce que montre le lemme suivant.

Lemme 2.37. *Soit $(X_i)_{i\in I}$ une famille de variables aléatoires indépendantes $X_i : (\Omega, \mathcal{F}) \to (E_i, \mathcal{E}_i)$, et soit $(\varphi_i)_{i\in I}$ une famille de fonctions mesurables $\varphi_i : (E_i, \mathcal{E}_i) \to (E_i', \mathcal{E}_i')$. Alors la famille*

$$(\varphi_i(X_i))_{i\in I}$$

est également indépendante.

Démonstration. Il suit de l'indépendance de la famille $(X_i)_{i\in I}$ que, pour tout $J \subset I$ fini,

$$\mathbb{P}(\forall i \in J,\, \varphi_i(X_i) \in A_i) = \mathbb{P}(\forall i \in J,\, X_i \in \varphi_i^{-1}(A_i)) = \prod_{i\in J} \mathbb{P}(X_i \in \varphi_i^{-1}(A_i)) = \prod_{i\in J} \mathbb{P}(\varphi_i(X_i) \in A_i),$$

pour toute collection $A_i \in \mathcal{E}_i'$, $i \in J$. □

Théorème 2.38. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $X_1, \dots, X_n$ des variables aléatoires $X_i : (\Omega, \mathcal{F}) \to (E_i, \mathcal{E}_i)$. Alors, les variables aléatoires $X_1, \dots, X_n$ sont indépendantes si et seulement si leur loi conjointe est le produit de leurs lois respectives :*

$$\mathbb{P}_{(X_1,\dots,X_n)} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}.$$

Démonstration. Soit $A_i \in \mathcal{E}_i$, $i = 1, \dots, n$. Par définition,

$$\mathbb{P}_{(X_1,\dots,X_n)}(A_1 \times \cdots \times A_n) = \mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n)$$
$$\mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}(A_1 \times \cdots \times A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).$$

Par conséquent, $X_1, \dots, X_n$ sont indépendantes si et seulement si les mesures $\mathbb{P}_{(X_1,\dots,X_n)}$ et $\mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$ coïncident sur tous les pavés $A_1 \times \cdots \times A_n$. La conclusion suit donc du Théorème 2.22, l'ensemble des pavés formant un π-système générant la tribu produit. □

En particulier, si $X_1, \dots, X_n$ sont des variables aléatoires réelles de densité de probabilité $f_{X_1}, \dots, f_{X_n}$ respectivement, alors les variables aléatoires $X_1, \dots, X_n$ sont indépendantes si et seulement si la densité de probabilité conjointe de $X_1, \dots, X_n$ se factorise :

$$f_{(X_1,\dots,X_n)}(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

Définition 2.39. *Une famille de variables aléatoires $(X_i)_{i\in I}$ est dite **i.i.d.** ($\equiv$ **i**ndépendantes et **i**dentiquement **d**istribuées) si elles sont indépendantes et ont toutes la même loi.*

Nous utiliserons également occasionnellement la terminologie suivante.

Définition 2.40. *Une variable aléatoire X est indépendante d'une collection $\mathcal{G}$ de parties de $\mathcal{F}$ si $\sigma(X)$ et $\mathcal{G}$ sont indépendantes.*

2.8 Espérance, variance, covariance et moments

2.8.1 Espérance

Définition 2.41. *Soit X une variable aléatoire réelle. On dit que X **admet une espérance** si*

$$X \geqslant 0 \qquad \text{ou} \qquad \int_\Omega |X(\omega)| \,\mathrm{d}\mathbb{P}(\omega) < \infty.$$

*Dans les deux cas, on définit l'**espérance** de X par*

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

On pose $\mathscr{L}^1 \equiv \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P}) \coloneqq \{X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})) \text{ mesurable} \,|\, \mathbb{E}[|X|] < \infty\}$.

Remarque 2.42. *Afin d'avoir un espace vectoriel normé (et pas juste muni d'une semi-norme), les analystes travaillent plutôt avec l'espace $L_1(\Omega, \mathcal{F}, \mathbb{P})$, obtenu en quotientant $\mathscr{L}^1$ par la relation d'égalité presque partout (presque sûre dans le langage probabiliste). Pour de bonnes raisons sur lesquelles nous ne nous attarderons pas ici, les probabilistes préfèrent travailler avec $\mathscr{L}^1$, quitte à mettre des "presque sûrement" un peu partout.*

Définition 2.43. *Soit $X_1, \dots, X_n$ des variables aléatoires réelles définies sur un même espace de probabilité et $\mathbf{X} = (X_1, \dots, X_n)$. Si $X_1, \dots, X_n$ possèdent une espérance, l'espérance de $\mathbf{X}$ est définie par*

$$\mathbb{E}[\mathbf{X}] \coloneqq (\mathbb{E}[X_1], \dots, \mathbb{E}[X_n]).$$

Soit X_1, X_2 deux variables aléatoires réelles définies sur un même espace de probabilité et $X \coloneqq X_1 + \mathrm{i}X_2$. Si X_1 et X_2 possèdent une espérance, l'espérance de X est définie par

$$\mathbb{E}[X] \coloneqq \mathbb{E}[X_1] + \mathrm{i}\mathbb{E}[X_2].$$

La **formule de transfert** suivant facilite grandement l'évaluation des espérances.

Lemme 2.44. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ une variable aléatoire.*

1. *Si $\varphi : (E, \mathcal{E}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ est une fonction mesurable positive, on a*

$$\mathbb{E}[\varphi(X)] = \int_E \varphi(x) \,\mathrm{d}\mathbb{P}_X(x).$$

2. *Soit $\varphi : E \to \mathbb{R}^d$ ou $\varphi : E \to \mathbb{C}$ une fonction mesurable. Alors, $\varphi(X) \in \mathscr{L}^1$ si et seulement si la fonction φ est $\mathbb{P}_X$-intégrable et, dans ce cas,*

$$\mathbb{E}[\varphi(X)] = \int_E \varphi(x) \,\mathrm{d}\mathbb{P}_X(x).$$

Démonstration. *1.* Lorsque $\varphi = \mathbf{1}_A$, avec $A \in \mathcal{E}$, l'affirmation se réduit à $\mathbb{P}(X \in A) = \mathbb{P}_X(A)$, ce qui est vrai par définition de la loi de X. Par linéarité de l'intégrale, ceci s'étend à toute fonction φ étagée positive. Si φ est mesurable et positive, alors φ est la limite croissante de fonctions étagées positives φ_n et $\varphi \circ X$ est la limite croissante des fonctions $\varphi_n \circ X$. L'affirmation suit donc du théorème de convergence monotone.
2. Considérons à présent une fonction mesurable $\varphi : E \to \mathbb{R}$ pas nécessairement positive. En appliquant la première partie à la fonction $|\varphi|$, on en déduit l'équivalence annoncée. Ainsi, si φ est $\mathbb{P}_X$-intégrable, la conclusion suit en décomposant $\varphi = \varphi_+ - \varphi_-$ et en appliquant la première partie.

Finalement, le cas d'une fonction à valeurs dans $\mathbb{R}^d$ ou $\mathbb{C}$ suit du cas réel en l'appliquant à chaque composante. □

Remarque 2.45. *Lorsque X est une variable aléatoire réelle, le résultat précédent implique (avec $\varphi(x) \coloneqq x$) que*

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \, \mathrm{d}\mathbb{P}_X(x).$$

En particulier, l'espérance de X ne dépend que de la loi de X (et pas, par exemple, de la forme précise de X ou même de l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ sous-jacent).

Exemple 2.46. Soit X une variable aléatoire discrète telle que $\mathbb{P}(X \in E) = 1$ pour un ensemble E fini ou dénombrable E et $\varphi : E \to \mathbb{R}$. On a vu que, dans ce cas, $\mathbb{P}_X = \sum_{x \in E} \mathbb{P}(X = x)\, \delta_x$. On obtient donc

$$\mathbb{E}[\varphi(X)] = \sum_{x \in E} \mathbb{P}(X = x) \int_E \varphi(y) \, \mathrm{d}\delta_x(y) = \sum_{x \in E} \varphi(x)\, \mathbb{P}(X = x). \qquad \lrcorner$$

Remarque 2.47. *On répète N fois une expérience consistant à mesurer la valeur prise par une variable aléatoire réelle discrète X, obtenant ainsi les résultats numériques $x_1, \dots, x_N$. La moyenne de ces résultats est donnée par*

$$m = \frac{1}{N} \sum_{i=1}^{N} x_i = \sum_{x \in E} \frac{N(x)}{N}\, x,$$

où E est tel que $\mathbb{P}(X \in E) = 1$ et $N(x)$ le nombre d'expériences lors desquelles la valeur x est observée. Dans l'interprétation fréquentiste, on s'attend alors à ce que, pour chaque valeur $x \in E$, la fraction $N(x)/N$ converge, lorsque $N \to \infty$, vers la probabilité $\mathbb{P}(X = x)$ d'observer la valeur x. Par conséquent, l'espérance $\sum_{x \in E} x\mathbb{P}(X = x)$ devrait fournir une approximation asymptotiquement correcte de m. Les lois des grands nombres confirmeront cette heuristique, pour des variables aléatoires arbitraires dans $\mathscr{L}^1$.

Exemple 2.48. Un cas particulier élémentaire, mais très utile, de l'exemple précédent est celui des indicatrices. Soit $A, B \in \mathcal{F}$. Alors,

$$\mathbb{E}(\mathbf{1}_A) = \mathbb{P}(A) \qquad \text{et} \qquad \mathbb{E}(\mathbf{1}_A \mathbf{1}_B) = \mathbb{P}(A \cap B). \qquad \lrcorner$$

Exemple 2.49. Soit X une variable aléatoire réelle avec densité f_X. Alors, pour toute fonction $\mathbb{P}_X$-intégrable $\varphi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}} \varphi(x) f_X(x) \, \mathrm{d}\lambda(x). \qquad \lrcorner$$

Remarque 2.50. *On utilise souvent l'espérance pour déterminer si un jeu est équitable : si X représente le gain à la fin du jeu (donc une perte s'il est négatif), alors l'espérance donne le gain moyen.*

Par exemple, considérons le jeu suivant : on lance un dé (équilibré) et on reçoit n francs si le dé indique n. Dans ce cas, le joueur va recevoir en moyenne $3{,}5$ francs. Le jeu lui sera donc favorable si sa mise initiale est inférieure à ce montant et défavorable si elle lui est supérieure.

On pourrait être tenté de dire plus généralement qu'un jeu vaut la peine d'être joué si $\mathbb{E}[X] > 0$ puisqu'en moyenne on gagne plus qu'on ne perd. Il faut cependant se méfier de cette intuition.

Considérons le jeu suivant (très discuté au début du XVIII$^{\text{e}}$ siècle) : on jette une pièce de monnaie jusqu'à l'apparition du premier « face » ; si cela a lieu au T-ième lancer, le joueur reçoit 2^T francs. Quelle serait une mise équitable ? On vérifie facilement que l'espérance de la somme reçue est infinie, et que, par conséquent, le jeu est favorable au joueur quelle que soit sa mise initiale ! C'est le célèbre ***paradoxe de Saint-Pétersbourg****.*

De par sa définition, l'espérance d'une variable aléatoire réelle X hérite immédiatement des propriétés élémentaires de l'intégrale de Lebesgue.

Lemme 2.51. *Soit $X, Y \in \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Alors,*

1. *(Positivité) $X \geqslant 0 \implies \mathbb{E}[X] \geqslant 0$.*
2. *(Inégalité du triangle) $\mathbb{E}[|X|] \geqslant |\mathbb{E}[X]|$.*
3. *(Linéarité) Pour tout $\alpha, \beta \in \mathbb{R}$, $\alpha X + \beta Y \in \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ et $\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$.*

Exemple 2.52. On désire trouver le nombre $a \in \mathbb{R}$ qui approxime le mieux une variable aléatoire réelle X dans le sens qu'il rend la quantité $\mathbb{E}[(X-a)^2]$ minimale. On a

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

En dérivant, on voit que le minimum est atteint en $a = \mathbb{E}[X]$. ⌟

Exemple 2.53. On appelle triangle d'un graphe, un triplet de sommets x, y, z tels que $x \sim y$, $y \sim z$ et $z \sim x$. Quelle est l'espérance du nombre de triangles $K_\triangle$ dans le graphe aléatoire $\mathcal{G}(n,m)$ avec $n \geqslant 3$ et $m \geqslant 3$? Il suit de la linéarité et de l'Exemple 2.48 que

$$\mathbb{E}[K_\triangle] = \mathbb{E}\Big[\sum_{1\leqslant x<y<z\leqslant n} \mathbf{1}_{\{x\sim y, y\sim z, z\sim x\}}\Big] = \sum_{1\leqslant x<y<z\leqslant n} \mathbb{P}(x\sim y, y\sim z, z\sim x).$$

Comme $\mathbb{P}(x\sim y, y\sim z, z\sim x) = \binom{N-3}{m-3}/\binom{N}{m}$ et comme la somme contient $\binom{n}{3}$ termes, on en conclut que

$$\mathbb{E}[K_\triangle] = \binom{n}{3}\frac{m(m-1)(m-2)}{N(N-1)(N-2)}.$$ ⌟

L'espérance hérite également des bonnes propriétés de convergence de l'intégrale de Lebesgue.

Théorème 2.54. *Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires réelles sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$.*

1. *(Convergence monotone) Supposons que, $\forall n$, $0 \leqslant X_n \leqslant X_{n+1}$ et que $X_n \to X$ ponctuellement. Alors,*
$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$
2. *(Convergence dominée) Supposons que $X_n \to X$ ponctuellement et qu'il existe une variable aléatoire $Y \in \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ telle que $|X_n| \leqslant Y$ pour tout n. Alors, $X \in \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ et*
$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$
3. *Supposons que $X_n \geqslant 0$ pour tout n. Alors,*
$$\mathbb{E}\Big[\sum_{n\geqslant 1} X_n\Big] = \sum_{n\geqslant 1} \mathbb{E}[X_n].$$
4. *(Lemme de Fatou) Supposons que $X_n \geqslant 0$ pour tout n. Alors,*
$$\mathbb{E}\big[\liminf_{n\to\infty} X_n\big] \leqslant \liminf_{n\to\infty} \mathbb{E}[X_n].$$

Donnons à présent l'espérance pour les lois introduites plus tôt dans ce chapitre. Comme déjà noté dans la Remarque 2.45, l'espérance $\mathbb{E}[X]$ ne dépend que de la loi de la variable aléatoire X ; on peut donc parler sans ambiguïté de l'espérance d'une loi.

Lemme 2.55. *La table 2.1 donne la valeur de l'espérance pour diverses lois, en fonction de leurs paramètres.*

Démonstration.

1. *Loi de Bernoulli.* L'espérance d'une variable aléatoire $X \sim \mathsf{bernoulli}(p)$ est immédiate à calculer :
$$\mathbb{E}[X] = 1\cdot p + 0\cdot(1-p) = p.$$
2. *Loi binomiale.* La façon la plus simple de calculer l'espérance d'une variable aléatoire $X \sim \mathsf{binom}(n,p)$ est d'utiliser la linéarité de l'espérance. X ayant la même loi que le nombre total de succès après n

Loi	Espérance	Variance
bernoulli(p)	p	$p(1-p)$
binom(n,p)	np	$np(1-p)$
poisson(λ)	λ	λ
géom(p)	$1/p$	$(1-p)/p^2$
hypergeom$(N,K,n))$	Kn/N	$nK(N-K)(N-n)/(N^3-N^2)$
U(a,b)	$(a+b)/2$	$(b-a)^2/12$
exp$(\lambda))$	$1/\lambda$	$1/\lambda^2$
$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2
cauchy	n'existe pas	n'existe pas

TABLE 2.1: L'espérance et la variance de quelques lois importantes, en fonction de leurs paramètres.

épreuves de Bernoulli, on a $X \overset{\text{loi}}{=} Y_1 + \cdots + Y_n$, où les Y_i sont des variables de Bernoulli de paramètre p indépendantes. Par conséquent,

$$\mathbb{E}[X] = \mathbb{E}\Big[\sum_{i=1}^{n} Y_i\Big] = \sum_{i=1}^{n} \mathbb{E}[Y_i] = np.$$

3. *Loi de Poisson.* L'espérance d'une variable aléatoire X suivant une loi de Poisson est donnée par

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda}\lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda.$$

4. *Loi géométrique.* L'espérance d'une variable aléatoire X de loi géométrique est donnée par la série

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} kp(1-p)^{k-1}.$$

Pour en calculer la somme, introduisons la fonction

$$G(x) = \sum_{k=1}^{\infty} x^k = \frac{x}{1-x}.$$

Cette série converge absolument lorsque $|x| < 1$ et, dans ce cas, il est donc possible d'interchanger sommation et dérivation. Par conséquent,

$$G'(x) = \frac{1}{(1-x)^2} = \sum_{k=1}^{\infty} kx^{k-1}.$$

On a donc

$$\mathbb{E}[X] = p\,G'(1-p) = p\frac{1}{p^2} = \frac{1}{p}.$$

5. *Loi hypergéométrique.* L'espérance d'une variable hypergéométrique sera traitée plus tard, dans l'Exemple 4.6.
6. *Loi uniforme sur* $[a,b]$. Soit $X \sim \mathsf{U}(a,b)$.

$$\mathbb{E}[X] = \tfrac{1}{b-a}\int_a^b s\,\mathrm{d}s = \tfrac{a+b}{2}.$$

7. *Loi exponentielle.* Soit $X \sim \mathsf{exp}(\lambda)$.
$$\mathbb{E}[X] = \lambda \int_0^\infty s e^{-\lambda s}\,\mathrm{d}s = \int_0^\infty e^{-\lambda s}\,\mathrm{d}s = \lambda^{-1}.$$
8. *Loi normale.* Soit $X \sim \mathcal{N}(\mu, \sigma^2)$.
$$\mathbb{E}[X] = \tfrac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^\infty s e^{-(s-\mu)^2/2\sigma^2}\,\mathrm{d}s = \mu + \tfrac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^\infty s e^{-s^2/2\sigma^2}\,\mathrm{d}s = \mu.$$
9. *Loi de Cauchy.* Soit $X \sim \mathsf{cauchy}$. Il suffit d'observer que $\mathbb{E}[|X|] = \infty$. □

Exemple 2.56.
1. On vous propose le jeu suivant : on vous tend deux enveloppes en vous informant que le montant contenu dans l'une est le double du montant contenu dans l'autre, et vous devez en choisir une. Expliquez en quoi le raisonnement suivant est faux : soit X le montant contenu dans l'enveloppe que vous avez décidé de tirer ; l'espérance de vos gains si vous changez d'avis est de $\frac{1}{2}\cdot X/2 + \frac{1}{2}\cdot 2X = \frac{5}{4}X > X$, et donc vous feriez mieux de choisir l'autre enveloppe (et bien sûr, on peut alors répéter cet argument une fois que vous avez choisi l'autre enveloppe).
2. On vous présente deux enveloppes. Chacune contient un papier sur lequel est inscrit un nombre entier (positif ou négatif) ; les deux nombres sont arbitraires, mais distincts. Vous gagnez si vous parvenez à tirer le nombre le plus grand. Vous pouvez choisir une des enveloppes et l'ouvrir. Vous devez ensuite décider si vous préférez garder l'enveloppe choisie, ou plutôt prendre l'autre. Montrez qu'il existe un algorithme de décision (changer ou non d'enveloppe en fonction du nombre découvert) qui vous permet de choisir le plus grand nombre strictement plus d'une fois sur deux (dans le sens que si une infinité de personnes appliquaient toutes cette stratégie pour la même paire de nombres, alors la fraction de bonnes réponses serait strictement supérieure à $1/2$). ⌟

Le résultat élémentaire suivant se révèle régulièrement utile.

Lemme 2.57. *1. Soit X est une variable aléatoire à valeurs dans $\mathbb{N}$. Alors,*
$$\mathbb{E}[X] = \sum_{n\geqslant 0} \mathbb{P}(X > n).$$
2. Soit X une variable aléatoire réelle positive. Alors,
$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)\,\mathrm{d}x.$$

Démonstration. Les deux affirmations suivent du théorème de Tonelli. Dans le premier cas,
$$\mathbb{E}[X] = \sum_{m\geqslant 1} m\mathbb{P}(X = m) = \sum_{m\geqslant 1}\sum_{n=0}^{m-1} \mathbb{P}(X = m) = \sum_{n\geqslant 0}\sum_{m=n+1}^{\infty} \mathbb{P}(X = m) = \sum_{n\geqslant 0} \mathbb{P}(X > n).$$
Dans le second cas,
$$\mathbb{E}[X] = \int_{\mathbb{R}} x\,\mathrm{d}\mathbb{P}_X(x) = \int_{\mathbb{R}} \left[\int_0^\infty \mathbf{1}_{\{y<x\}}\,\mathrm{d}y\right] \mathrm{d}\mathbb{P}_X(x) = \int_0^\infty \left[\int_{\mathbb{R}} \mathbf{1}_{\{y<x\}}\,\mathrm{d}\mathbb{P}_X(x)\right] \mathrm{d}y = \int_0^\infty \mathbb{P}(X > y)\,\mathrm{d}y.$$
□

Théorème 2.58 (Inégalité de Jensen[9]). *Soit X une variable aléatoire réelle admettant une espérance et φ une fonction convexe[10]. Alors,*
$$\mathbb{E}[\varphi(X)] \geqslant \varphi(\mathbb{E}[X]).$$
De plus, lorsque φ est strictement convexe, il y a égalité si et seulement si X est une variable aléatoire constante.

Démonstration. Il suit de la définition de la convexité de φ, avec $x = \mathbb{E}[X]$, qu'il existe $a \in \mathbb{R}$ tel que, pour tout $y \in \mathbb{R}$,

$$\varphi(y) \geqslant \varphi(\mathbb{E}[X]) + a(y - \mathbb{E}[X]).$$

On a donc

$$\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}} \varphi(y)\, \mathrm{d}\mathbb{P}_X(y) \geqslant \varphi(\mathbb{E}[X]) + a(\mathbb{E}[X] - \mathbb{E}[X]) = \varphi(\mathbb{E}[X]). \qquad \square$$

2.8.2 Variance, moments d'ordres supérieurs

Définition 2.59. *On appelle $\mathbb{E}[X^n]$ le **moment d'ordre n** de la variable aléatoire réelle X, pourvu que cette espérance soit bien définie.*

Remarque 2.60. *Pour tout $p \geqslant 1$, on définit*

$$\mathscr{L}^p \equiv \mathscr{L}^p(\Omega, \mathcal{F}, \mathbb{P}) := \{X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})) \text{ mesurable} \,|\, \|X\|_p < \infty\},$$

où l'on a introduit la semi-norme $\|X\|_p := \mathbb{E}[|X|^p]^{1/p}$. Nous référons au cours d'analyse fonctionnelle pour plus d'informations sur les propriétés de ces espaces (ou, plus précisément, sur les espaces $L^p \equiv L^p(\Omega, \mathcal{F}, \mathbb{P})$, obtenus en quotientant $\mathscr{L}^p$ par la relation d'égalité presque partout).

Remarque 2.61. *Soit $p \geqslant q \geqslant 1$ et $X \in \mathscr{L}^p$. Il suit de l'inégalité de Jensen[9] que*

$$\|X\|_p = \mathbb{E}[|X|^p]^{1/p} = \mathbb{E}[(|X|^q)^{p/q}]^{1/p} \geqslant \mathbb{E}[(|X|^q)]^{1/q} = \|X\|_q,$$

puisque $x \mapsto |x|^r$ est convexe si $r \geqslant 1$[10]. En particulier, $X \in \mathscr{L}^q$.

Remarque 2.62. *En général, même la donnée de tous les moments d'une variable aléatoire ne suffit pas pour déterminer sa loi. C'est le cas si cette variable aléatoire possède certaines bonnes propriétés que nous ne discuterons pas ici. Mentionnons simplement la condition suffisante suivante : si X et Y sont deux variables aléatoires réelles satisfaisant $\mathbb{E}(e^{\lambda X}) < \infty$ et $\mathbb{E}(e^{\lambda Y}) < \infty$ pour tout $\lambda \in (-\epsilon, \epsilon)$, pour un $\epsilon > 0$, et telles que $\mathbb{E}[X^n] = \mathbb{E}[Y^n]$ pour tout $n \in \mathbb{N}$, alors $X \overset{\text{loi}}{=} Y$.*

Une quantité particulièrement importante est la variance. Si l'espérance donne la valeur moyenne de la variable aléatoire, la variance (ou plutôt sa racine carrée, l'écart-type) mesure sa dispersion[11].

Définition 2.63. *Soit $X \in \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. On appelle **variance** de X la quantité*

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

*(qui peut être égale à $+\infty$). On appelle **écart-type** de X la quantité $\sigma(X) := \sqrt{\mathrm{Var}[X]}$.*

Lemme 2.64.
1. $\mathrm{Var}[X] \geqslant 0$, *et* $\mathrm{Var}[X] = 0$ *si et seulement si* $\mathbb{P}(X = \mathbb{E}[X]) = 1$.
2. $\mathrm{Var}[X] < \infty$ *si et seulement si* $X \in \mathscr{L}^2$.
3. *Pour tout* $X \in \mathscr{L}^2$, $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.
4. *Pour tout* $a, b \in \mathbb{R}$, $\mathrm{Var}[a + bX] = b^2 \mathrm{Var}[X]$.
5. *Pour tout* $X, Y \in \mathscr{L}^2$, $\mathrm{Var}[X + Y] < \infty$.

9. Johan Ludwig William Valdemar Jensen (1859, Naksov – 1925, Copenhague), mathématicien et ingénieur danois.

10. Soit $-\infty \leqslant a < b \leqslant +\infty$. $\varphi : (a, b) \to \mathbb{R}$ est convexe si et seulement si $\forall x \in (a, b), \exists a \in \mathbb{R}, \forall y \in (a, b), \varphi(y) \geqslant \varphi(x) + a(y - x)$. Si l'inégalité est toujours stricte lorsque $y \neq x$, on dit que φ est strictement convexe.

11. On peut trouver bizarre cette façon de mesurer la dispersion ; par exemple, $\mathbb{E}[|X - \mathbb{E}[X]|]$ pourrait sembler un choix plus naturel. Outre le fait que la variance est un objet beaucoup plus agréable à manipuler mathématiquement, on verra plus tard (Remarque 6.17) qu'il s'agit réellement de la quantité naturelle à considérer.

Démonstration. Nous ne démontrerons que deux des affirmations, les autres étant immédiates.
Preuve de 2. Soit $Z \in \mathscr{L}^2$. Alors, pour tout $a \in \mathbb{R}$,

$$\mathbb{E}\big[(Z-a)^2\big] = \mathbb{E}\big[(Z-a)^2 \mathbf{1}_{\{|Z|<2|a|\}}\big] + \mathbb{E}\big[(Z-a)^2 \mathbf{1}_{\{|Z|\geqslant 2|a|\}}\big] \leqslant 9a^2 + \frac{9}{4}\mathbb{E}[Z^2] < \infty.$$

En prenant $Z = X$ et $a = \mathbb{E}[X]$, on obtient que $\mathbb{E}[X^2] < \infty \implies \mathrm{Var}[X] < \infty$.
En prenant $Z = X - \mathbb{E}[X]$ et $a = -\mathbb{E}[X]$, on obtient que $\mathrm{Var}[X] < \infty \implies \mathbb{E}[X^2] < \infty$.
Preuve de 5. Soit $\check{X} := X - \mathbb{E}[X]$ et $\check{Y} := Y - \mathbb{E}[Y]$. Comme $(a+b)^2 \leqslant 2(a^2+b^2)$, pour tout $a, b \in \mathbb{R}$, on peut écrire

$$\mathrm{Var}[X+Y] = \mathbb{E}\big[(\check{X}+\check{Y})^2\big] \leqslant 2\mathbb{E}[\check{X}^2] + 2\mathbb{E}[\check{Y}^2] = 2\mathrm{Var}[X] + 2\mathrm{Var}[Y] < \infty. \qquad \square$$

Le résultat suivant, très utile et dont nous verrons des extensions plus tard, montre un sens dans lequel la variance contrôle les fluctuations d'une variable aléatoire autour de son espérance.

Lemme 2.65 (Inégalité de Bienaymé[12]–Tchebychev[13])**.** *Soit $X \in \mathscr{L}^2$. Alors,*

$$\forall a > 0, \qquad \mathbb{P}\big(\big|X - \mathbb{E}[X]\big| \geqslant a\big) \leqslant \frac{\mathrm{Var}[X]}{a^2}.$$

En particulier, si $a \gg \sigma(X)$, alors $\mathbb{P}\big(\big|X - \mathbb{E}[X]\big| \geqslant a\big) \ll 1$.

Démonstration. Notons $\check{X} := X - \mathbb{E}[X]$. Il suffit d'observer que

$$\mathrm{Var}[X] = \mathbb{E}\big(\check{X}^2\big) \geqslant \mathbb{E}\big(\check{X}^2 \mathbf{1}_{\{\check{X}^2 \geqslant a^2\}}\big) \geqslant a^2\, \mathbb{P}(\check{X}^2 \geqslant a^2) = a^2\, \mathbb{P}(|X - \mathbb{E}[X]| \geqslant a). \qquad \square$$

Il n'est pas difficile de déterminer la variance des lois introduites plus haut.

Lemme 2.66. *La table 2.1 donne les variances des principales lois introduites précédemment.*

Démonstration.

1. *Loi de Bernoulli.* Soit $X \sim \mathsf{bernoulli}(p)$.
$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 1 \cdot p + 0 \cdot (1-p) - p^2 = p(1-p).$$
2. *Loi binomiale.* Voir l'Exemple 2.79.
3. *Loi de Poisson.* Soit $X \sim \mathsf{poisson}(\lambda)$.
$$\mathbb{E}\big[X(X-1)\big] = \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda}\lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = \lambda^2.$$
Par conséquent, $\mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}\big[X(X-1)\big] - \mathbb{E}[X]^2 + \mathbb{E}[X] = \lambda$.
4. *Loi géométrique.* Soit $X \sim \mathsf{géom}(p)$.
$$\mathbb{E}[X(X-1)] = \sum_{k=1}^{\infty} k(k-1)p(1-p)^{k-1}.$$
Pour calculer la somme, on procède comme pour l'espérance, en introduisant la fonction
$$G(x) = \sum_{k=1}^{\infty} x^k = \frac{x}{1-x},$$
et en utilisant le fait que $G''(x) = \frac{2}{(1-x)^3} = \sum_{k=1}^{\infty} k(k-1)\, x^{k-2}$. Par conséquent,
$$\mathrm{Var}[X] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X^2] = p(1-p)G''(1-p) + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

12. Irénée-Jules Bienaymé (1796, Paris - 1878, Paris), probabiliste et statisticien français.
13. Pafnouti Lvovitch Tchebychev (1821, Okatovo - 1894, Saint-Petersbourg), mathématicien russe. Son nom est aussi translittéré comme Chebyshov, Chebyshev, ou Tschebyscheff.

5. *Loi hypergéométrique.* Voir l'Exemple 4.6.
6. *Loi uniforme sur* $[a,b]$. Soit $X \sim \mathsf{U}(a,b)$.
$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \tfrac{1}{b-a}\int_a^b s^2\,\mathrm{d}s - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$
7. *Loi exponentielle.* Soit $X \sim \mathsf{exp}(\lambda)$.
$$V(X) = \lambda\int_0^\infty s^2 e^{-\lambda s}\,\mathrm{d}s - \lambda^{-2} = \lambda^{-2}.$$
8. *Loi normale.* Soit $X \sim \mathcal{N}(\mu,\sigma^2)$.
$$\mathrm{Var}[X] = \tfrac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^\infty (s-\mu)^2 e^{-(s-\mu)^2/2\sigma^2}\,\mathrm{d}s = \sigma^2\tfrac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^\infty e^{-s^2/2\sigma^2}\,\mathrm{d}s = \sigma^2.$$
9. *Loi de Cauchy.* Soit $X \sim \mathsf{cauchy}$. La variance n'existe pas : X n'appartient même pas à $\mathscr{L}^1$. □

2.8.3 Covariance et corrélation

En général, $\mathrm{Var}[X+Y] \neq \mathrm{Var}[X] + \mathrm{Var}[Y]$: en effet, un bref calcul montre que
$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big].$$
Ceci motive la définition suivante.

Définition 2.67. *On appelle* ***covariance*** *de deux variables aléatoires réelles* $X, Y \in \mathscr{L}^2$ *la quantité*
$$\mathrm{Cov}[X,Y] := \mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$
Deux variables aléatoires X *et* Y *sont* ***non corrélées*** *si* $\mathrm{Cov}[X,Y] = 0$.

En particulier,
$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}[X,Y],$$
et donc, lorsque X, Y sont non corrélées,
$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

Remarque 2.68. *Faites attention : la variance n'est pas un opérateur linéaire, même restreinte aux variables aléatoires non corrélées, puisque* $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$.

Lemme 2.69. *Soit* $X, X_1, X_2, Y \in \mathscr{L}^2$.

1. *La covariance est une forme bilinéaire symétrique :*
$$\begin{aligned} \mathrm{Cov}[X,Y] &= \mathrm{Cov}[Y,X], \\ \forall a,b\in\mathbb{R},\qquad \mathrm{Cov}[aX,bY] &= a\,b\,\mathrm{Cov}[X,Y], \\ \mathrm{Cov}[X_1+X_2,Y] &= \mathrm{Cov}[X_1,Y] + \mathrm{Cov}[X_2,Y]. \end{aligned}$$
2. *Pour des variables aléatoires* $X_1,\dots,X_n \in \mathscr{L}^2$, *on a*
$$\mathrm{Var}\Big[\sum_{i=1}^n X_i\Big] = \sum_{i=1}^n \mathrm{Var}[X_i] + \sum_{i\neq j}\mathrm{Cov}[X_i,X_j].$$

Démonstration. Laissée en exercice. □

En statistiques, une autre quantité est souvent utilisée pour mesurer la corrélation entre deux variables aléatoires, ayant l'avantage de ne pas changer si les variables aléatoires X et Y sont multipliées par des coefficients positifs (en particulier, si on change d'unités).

Définition 2.70. *On appelle* ***coefficient de corrélation*** *de deux variables aléatoires $X, Y \in \mathscr{L}^2$ de variances non-nulles la quantité*

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

Lemme 2.71 (Inégalité de Cauchy–Schwarz[14])**.** *Soit $X, Y \in \mathscr{L}^2$.*

$$\mathbb{E}[XY]^2 \leqslant \mathbb{E}[X^2]\mathbb{E}[Y^2],$$

avec égalité si et seulement si $\mathbb{P}(aX = bY) = 1$ pour des réels a et b dont au moins un est non nul.

Démonstration. On peut supposer que $\mathbb{E}[X^2] \neq 0$ et $\mathbb{E}[Y^2] \neq 0$ (sinon la variable aléatoire correspondante est égale à 0 avec probabilité 1 et l'affirmation est triviale). Fixons $b \in \mathbb{R}^*$. Dans ce cas, on a,

$$\forall a \in \mathbb{R}, \qquad a^2\mathbb{E}[X^2] - 2ab\mathbb{E}[XY] + b^2\mathbb{E}[Y^2] = \mathbb{E}\big[(aX - bY)^2\big] \geqslant 0.$$

Par conséquent, le membre de gauche est une fonction quadratique de la variable a s'annulant en au plus un point. Ceci implique que son discriminant doit être négatif ou nul, c'est-à-dire

$$\mathbb{E}[XY]^2 - \mathbb{E}[X^2]\mathbb{E}[Y^2] \leqslant 0.$$

Le discriminant est nul si et seulement s'il y a un unique zéro, ce qui ne peut avoir lieu que si

$$\exists a \in \mathbb{R}, \qquad \mathbb{E}\big[(aX - bY)^2\big] = 0.$$ □

Il suit de l'inégalité de Cauchy–Schwarz que la valeur absolue du coefficient de corrélation $\rho(X,Y)$ est égal à 1 si et seulement s'il existe une relation affine entre les variables aléatoires X et Y.

Corollaire 2.72. *Soit $X, Y \in \mathscr{L}^2$.*

$$|\rho(X,Y)| \leqslant 1,$$

avec égalité si et seulement si $\mathbb{P}(Y = aX + b) = 1$ pour des réels a et b.

Démonstration. Il suffit d'appliquer l'inégalité de Cauchy–Schwarz aux variables aléatoires $\check{X} \coloneqq X - \mathbb{E}[X]$ et $\check{Y} \coloneqq Y - \mathbb{E}[Y]$. □

Considérons deux quantités aléatoires (par exemple des résultats de mesures), et supposons que l'on cherche à résumer la relation qui existe entre ces dernières à l'aide d'une droite. On parle alors d'ajustement linéaire. Comment calculer les caractéristiques de cette droite ? En faisant en sorte que l'erreur que l'on commet en représentant la liaison entre nos variables par une droite soit la plus petite possible. Le critère formel le plus souvent utilisé, mais pas le seul possible, est de minimiser la somme de toutes les erreurs effectivement commises au carré. On parle alors d'**ajustement selon la méthode des moindres carrés**. La droite résultant de cet ajustement s'appelle une **droite de régression**. Le résultat suivant montre que le coefficient de corrélation mesure la qualité de la représentation de la relation entre nos variables par cette droite.

14. Hermann Amandus Schwarz (1843, Hermsdorf – 1921, Berlin), mathématicien allemand.

Lemme 2.73. *Pour toute paire de variables aléatoires réelles $X, Y \in \mathscr{L}^2$, on a*

$$\min_{a,b\in\mathbb{R}} \mathbb{E}\big[(Y - aX - b)^2\big] = (1 - \rho(X,Y)^2)\,\mathrm{Var}[Y].$$

De plus, le minimum est atteint pour $a = \mathrm{Cov}[X,Y]/\mathrm{Var}[X]$ et $b = \mathbb{E}[Y - aX]$.

Démonstration. En écrivant, comme d'habitude, $\check{X} \coloneqq X - \mathbb{E}[X]$ et $\check{Y} \coloneqq Y - \mathbb{E}[Y]$, on a

$$\mathbb{E}\big[(Y - aX - b)^2\big] = \mathbb{E}\big[(\check{Y} - a\check{X} - \check{b})^2\big],$$

où on a posé $\check{b} \coloneqq b + a\mathbb{E}[X] - \mathbb{E}[Y]$. On vérifie alors aisément que

$$\mathbb{E}\big[(Y - aX - b)^2\big] = a^2\mathbb{E}[\check{X}^2] - 2a\mathbb{E}[\check{X}\check{Y}] + \mathbb{E}[\check{Y}^2] + \check{b}^2 = a^2\mathrm{Var}[X] - 2a\mathrm{Cov}[X,Y] + \mathrm{Var}[Y] + \check{b}^2,$$

et le membre de droite est minimum lorsque $\check{b} = 0$, c'est-à-dire $b = \mathbb{E}[Y] - a\mathbb{E}[X]$, et

$$a = \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}.$$

□

Exemple 2.74. En physiologie, la loi de Kleiber[15] affirme que le métabolisme M d'un animal et son poids P satisfont la relation

$$M \propto P^\alpha,$$

avec α souvent proche de $3/4$ (alors que des arguments simples de dimensionnalité suggéreraient plutôt $2/3$). Afin de vérifier qu'une telle relation est valide pour une population donnée, on peut procéder comme suit : puisque

$$M \approx aP^\alpha \iff \log M \approx \log a + \alpha \log P,$$

on se ramène, en posant $X \coloneqq \log M$ et $Y \coloneqq \log P$, à vérifier qu'il y a une relation affine entre X et Y ; *cf.* Figure 2.3. Concrètement, on estime, à partir d'un échantillon, les paramètres a et α, ainsi que le coefficient de corrélation $\rho(X,Y)$. Ce dernier permet alors de mesurer la qualité de l'approximation linéaire ainsi obtenue. (Comment estimer ces paramètres à partir d'un échantillon relève de la Statistique ; nous étudierons ce type de problèmes dans le Chapitre 12.) ⌟

Définition 2.75. *Soit $X_1, \dots, X_n$ des variables aléatoires réelles et $\mathbf{X} = (X_1, \dots, X_n)$. La **matrice de covariance** du vecteur aléatoire X est la matrice $\mathrm{Cov}[\mathbf{X}] \in M_{n,n}(\mathbb{R})$ dont l'élément i,j est donné par $\mathrm{Cov}[X_i, X_j]$.*

2.8.4 Absence de corrélation et indépendance

Dans cette section, nous allons nous intéresser au lien entre indépendance et absence de corrélation. Commençons par une caractérisation équivalente de l'indépendance.

Lemme 2.76. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $X_i : (\Omega, \mathcal{F}) \to (E_i, \mathcal{E}_i)$, $i = 1, \dots, n$, une famille de variables aléatoires. Les propositions suivantes sont équivalentes :*

1. *Les variables aléatoires $X_1, \dots, X_n$ sont indépendantes.*
2. *Pour toute collection de fonctions φ_i à valeurs dans $\mathbb{R}$, $1 \leqslant i \leqslant n$, telles que $\varphi_i(X_i) \in \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P})$,*

$$\mathbb{E}\Big[\prod_{i=1}^n \varphi_i(X_i)\Big] = \prod_{i=1}^n \mathbb{E}[\varphi_i(X_i)].$$

15. Max Kleiber (1893, Zürich – 1976, Davis), biologiste suisse.

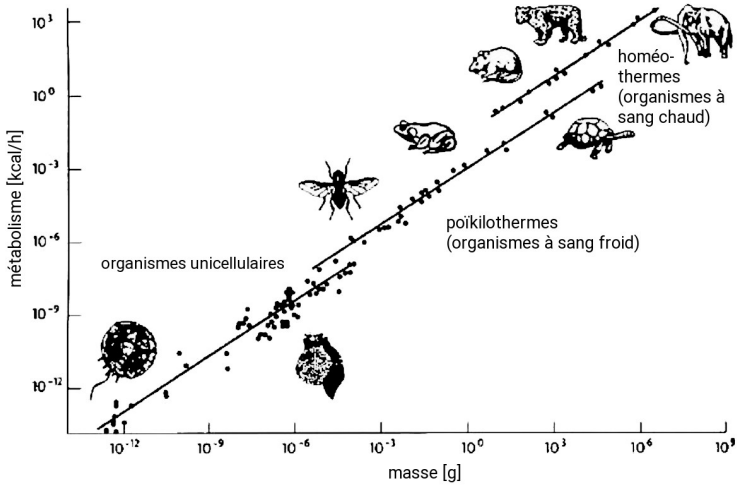


FIGURE 2.3: Métabolisme en fonction de la masse pour trois groupes d'organismes : homéothermes, poïkilothermes et unicellulaires. Les droites correspondent à la loi de Kleiber. *[Source : Scaling in Biology, edité par J. H. Brown et G. B. West, Oxford University Press, 2000]*

Démonstration. *1.* $\implies$ *2.* Cela suit immédiatement du Lemme 2.44 et de la factorisation de $\mathbb{P}_{(X_1,\dots,X_n)}$ (Théorème 2.38) :

$$\begin{aligned}\mathbb{E}\big[\prod_{i=1}^n \varphi_i(X_i)\big] &= \int_{E_1\times\cdots\times E_n} \varphi_1(x_1)\cdots\varphi_n(x_n)\,\mathrm{d}\mathbb{P}_{(X_1,\dots,X_n)}(x_1,\dots,x_n)\\ &= \int_{E_1\times\cdots\times E_n} \varphi_1(x_1)\cdots\varphi_n(x_n)\,\mathrm{d}\mathbb{P}_{X_1}(x_1)\cdots\mathrm{d}\mathbb{P}_{X_n}(x_n) = \prod_{i=1}^n \mathbb{E}[\varphi_i(X_i)].\end{aligned}$$

2. $\implies$ *1.* Soit $A_i \in \mathcal{E}_i$, $1 \leqslant i \leqslant n$. En appliquant *2.* avec les fonctions $\varphi_i(y) \coloneqq \mathbf{1}_{\{y\in A_i\}}$, on obtient

$$\mathbb{P}(\forall i \in \{1,\dots,n\},\ X_i \in A_i) = \mathbb{E}\Big[\prod_{i=1}^n \mathbf{1}_{\{X_i\in A_i\}}\Big] = \prod_{i=1}^n \mathbb{E}[\mathbf{1}_{\{X_i\in A_i\}}] = \prod_{i=1}^n \mathbb{P}(X_i \in A_i). \qquad \square$$

En particulier, l'indépendance implique l'absence de corrélation.

Lemme 2.77. *Soit $X, Y \in \mathscr{L}^1$ deux variables aléatoires indépendantes. Alors,* $\mathrm{Cov}[X,Y] = 0$.

Démonstration. C'est une conséquence immédiate du Lemme 2.76. $\square$

Remarque 2.78. *En particulier, si $X_1,\dots,X_n \in \mathscr{L}^2$ sont deux à deux indépendantes, alors*

$$\mathrm{Var}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \mathrm{Var}[X_i].$$

Exemple 2.79. Soit $X \sim \mathsf{binom}(n,p)$. On a vu que, dans ce cas, $X \overset{\text{loi}}{=} Y_1 + \cdots + Y_n$, où les variables aléatoires Y_i sont i.i.d., $Y_i \sim \mathsf{bernoulli}(p)$. On obtient donc immédiatement que $\mathrm{Var}[X] = np(1-p)$. ⌟

Remarque 2.80. *Le Lemme 2.77 montre que deux variables aléatoires réelles indépendantes sont toujours non corrélées. La réciproque est fausse en général, comme le montre l'exemple suivant.*

Exemple 2.81. Considérons $\Omega \coloneqq \{-1,0,1\}$ avec la distribution uniforme. Soit $X(\omega) \coloneqq \omega$ et $Y(\omega) \coloneqq |\omega|$ deux variables aléatoires. Alors, $\mathbb{E}[X] = 0$, $\mathbb{E}[Y] = 2/3$ et $\mathbb{E}[XY] = 0$. Par conséquent X et Y sont non corrélées. Elles ne sont par contre manifestement pas indépendantes. ⌟

2.8.5 Une première version de la loi (faible) des grands nombres

Nous avons motivé heuristiquement le fait que l'espérance $\mathbb{E}[X]$ devrait correspondre à la limite des moyennes des résultats obtenus en mesurant X lors d'une suite d'expériences aléatoires, lorsque le nombre d'expériences tend vers l'infini. Nous allons à présent rendre cela un peu plus précis.

Définition 2.82. *Soit $X_1, X_2, \dots, X_n$ une famille de variables aléatoires réelles. Leur **moyenne empirique** est la variable aléatoire*

$$\overline{X}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Nous pouvons à présent démontrer une première version de la loi des grands nombres.

Théorème 2.83 (Loi faible des grands nombres pour des variables non corrélées)**.** *Soit $X_1, \dots, X_n$ des variables aléatoires réelles non corrélées, de même espérance μ et de même variance $\sigma^2 < \infty$. Alors,*

$$\forall \epsilon > 0, \qquad \mathbb{P}(|\overline{X}_n - \mu| \geqslant \epsilon) \leqslant \frac{\sigma^2}{\epsilon^2 n}.$$

En particulier, $\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| \geqslant \epsilon) = 0$, pour tout $\epsilon > 0$.

Démonstration. Les variables aléatoires X_i étant non corrélées, il est facile de déterminer la variance de $\overline{X}_n$:

$$\mathrm{Var}[\overline{X}_n] = \mathrm{Var}\Big[\frac{1}{n}\sum_{i=1}^{n} X_i\Big] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[X_i] = \frac{\sigma^2}{n}.$$

Comme $\mathbb{E}(\overline{X}_n) = \mu$, le résultat suit de l'inégalité de Bienaymé–Tchebychev (Lemme 2.65) :

$$\mathbb{P}(|\overline{X}_n - \mu| \geqslant \epsilon) \leqslant \frac{\mathrm{Var}[\overline{X}_n]}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2 n}.$$ □

Exemple 2.84. On effectue 10 000 lancers d'une pièce de monnaie équilibrée. Afin de travailler avec des variables centrées, on encode le résultat du k-ième jet par une variable X_k telle que $\mathbb{P}(X_k = 1) = \mathbb{P}(X_k = -1) = \frac{1}{2}$ (au lieu de 0 et 1). La loi faible des grands nombres énoncée ci-dessus affirme que $\overline{X}_n \in [-\epsilon, \epsilon]$ avec grande probabilité lorsque n est suffisamment grand. L'estimée du théorème nous donne

$$\mathbb{P}\big(|\overline{X}_n| \geqslant \epsilon\big) \leqslant \frac{1}{n\epsilon^2}.$$

Par exemple, pour 10 000 lancers et $\epsilon = 0{,}1$, on a

$$\mathbb{P}(|\overline{X}_{10\,000}| \geqslant 0{,}1) \leqslant \frac{1}{100}.$$

Notez que ce n'est qu'une borne supérieure sur cette probabilité. On verra plus tard qu'elle est très pessimiste dans le cas présent (on montrera en effet que $\mathbb{P}(|\overline{X}_{10\,000}| \geqslant 0{,}1) \leqslant 3{,}5 \cdot 10^{-22}$). ⌟

Remarque 2.85 (Lien avec l'approche fréquentiste)**.** *Ce qu'affirme la loi faible des grands nombres, c'est que pour une précision ϵ donnée, la probabilité que l'espérance et la moyenne empirique diffèrent de plus de ϵ peut être rendue aussi petite qu'on le souhaite en considérant un échantillon suffisamment grand. En ce sens, elle justifie à posteriori l'axiomatique de la théorie de probabilités, en faisant le lien avec la notion intuitive de fréquence de réalisation d'un événement. En effet, considérons une expérience aléatoire, décrite par un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, que l'on répète N fois, de façon indépendante, obtenant une suite de résultats $(\omega_1, \omega_2, \dots, \omega_N)$. Alors, pour tout événement $A \in \mathcal{F}$, les variables aléatoires $Y_k(\omega_1, \dots, \omega_N) \coloneqq \mathbf{1}_A(\omega_k)$ sont i.i.d., avec $\mathbb{E}[Y_k] = \mathbb{P}(A)$.*

Par conséquent, si l'on note $N(A) \coloneqq \#\{1 \leqslant k \leqslant N \,|\, \omega_k \in A\} = \sum_{k=1}^{N} Y_k$ *le nombre d'expériences lors desquelles l'événement* A *est réalisé, on a, pour tout* $\epsilon > 0$*,*

$$\lim_{N\to\infty} \mathbb{P}\Big(\Big|\frac{N(A)}{N} - \mathbb{P}(A)\Big| \geqslant \epsilon\Big) = \lim_{N\to\infty} \mathbb{P}\Big(\Big|\frac{1}{N}\sum_{k=1}^{N} Y_k - \mathbb{E}[Y_1]\Big| \geqslant \epsilon\Big) = 0,$$

ce qui est parfaitement en accord avec l'interprétation fréquentiste des probabilités.

Nous reviendrons sur la loi des grands nombres, ainsi que sur des résultats plus précis concernant le comportement asymptotique de la moyenne empirique, au chapitre 6.

2.9 Suites infinies de variables aléatoires réelles indépendantes

Étant donné des mesures de probabilité $\mathbb{P}_1, \dots, \mathbb{P}_n$ sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, il est aisé de construire un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et des variables aléatoires réelles $X_1, \dots, X_n$ sur cet espace de sorte à ce que $X_1, \dots, X_n$ soient indépendantes et de lois $\mathbb{P}_1, \dots, \mathbb{P}_n$ respectivement. Il suffit de prendre $\Omega \coloneqq \mathbb{R}^n$, $\mathcal{F} \coloneqq \mathcal{B}(\mathbb{R}^n)$, $\mathbb{P} \coloneqq \mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n$ et de considérer les variables aléatoires réelles $X_1, \dots, X_n$ définies par $X_i(\omega) \coloneqq \omega_i$ pour tout $\omega = (\omega_1, \dots, \omega_n) \in \Omega$. Le Théorème 2.38 implique que $X_1, \dots, X_n$ sont indépendantes et, par définition de la mesure produit,

$$\mathbb{P}_{X_i}(A) = \mathbb{P}(X_i \in A) = \mathbb{P}(\{\omega \in \mathbb{R}^n \,|\, \omega_i \in A\}) = \mathbb{P}(\mathbb{R}^{i-1} \times A \times \mathbb{R}^{n-i}) = \mathbb{P}_i(A),$$

pour tout $A \in \mathcal{B}(\mathbb{R})$.

Dans la suite, nous considérerons souvent des suites infinies $(X_k)_{k\geqslant 1}$ de variables aléatoires réelles indépendantes de lois $(\mathbb{P}_k)_{k\geqslant 1}$. On souhaite donc étendre la construction précédente. On commence par poser $\Omega \coloneqq \mathbb{R}^{\mathbb{N}^*}$ et $X_i(\omega) \coloneqq \omega_i$, où $\omega = (\omega_k)_{k\geqslant 1} \in \Omega$. Il reste à définir une tribu appropriée $\mathcal{F}$ sur Ω et la mesure de probabilité $\mathbb{P}$...

La tribu $\mathcal{F}$ est celle qui est engendrée par les ensembles de la forme $\{\omega \in \Omega \,|\, \omega_1 \in A_1, \dots, \omega_n \in A_n\}$, pour tout $n \in \mathbb{N}^*$ et toute collection $A_1, \dots, A_n \in \mathcal{B}(\mathbb{R})$; on notera également $\mathcal{B}(\mathbb{R}^{\mathbb{N}^*})$ cette tribu. En d'autres termes, $\mathcal{B}(\mathbb{R}^{\mathbb{N}^*})$ est la tribu engendrée par les événements ne dépendant que d'un nombre fini de coordonnées. Pour ces événements, on sait comment on désire définir $\mathbb{P}$:

$$\mathbb{P}(\{\omega \in \Omega \,|\, \omega_1 \in A_1, \dots, \omega_n \in A_n\}) \coloneqq \mathbb{P}_1(A_1) \cdots \mathbb{P}_n(A_n).$$

Il se trouve que cela suffit pour définir de manière unique une mesure de probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{F})$.

Théorème 2.86 (Théorème d'extension de Kolmogorov)**.** *On suppose donnée une famille de mesures de probabilité* μ_n *sur* $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ *qui sont* ***consistantes*** *:*

$$\forall n \in \mathbb{N}^*,\ \forall A_1, \dots, A_n \in \mathcal{B}(\mathbb{R}), \qquad \mu_{n+1}(A_1 \times \cdots \times A_n \times \mathbb{R}) = \mu_n(A_1 \times \cdots \times A_n).$$

Alors, il existe une unique mesure de probabilité $\mathbb{P}$ *sur* $(\mathbb{R}^{\mathbb{N}^*}, \mathcal{B}(\mathbb{R}^{\mathbb{N}^*}))$ *telle que*

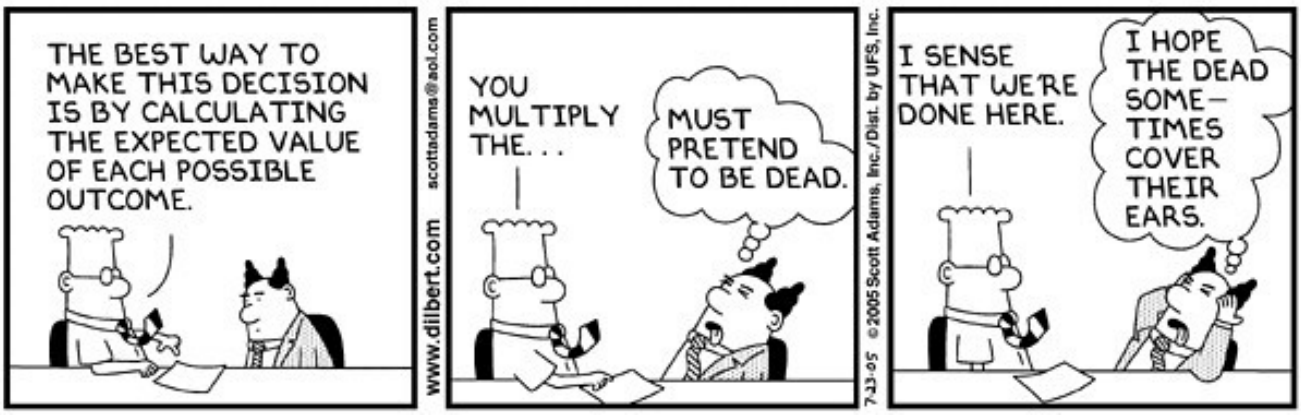
$$\forall n \in \mathbb{N}^*,\ \forall A_1, \dots, A_n \in \mathcal{B}(\mathbb{R}), \qquad \mathbb{P}(\{\omega \in \mathbb{R}^{\mathbb{N}^*} \,|\, \omega_1 \in A_1, \dots, \omega_n \in A_n\}) = \mu_n(A_1 \times \cdots \times A_n).$$

Il suffit alors d'appliquer ce théorème aux mesures

$$\mu_n \coloneqq \mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n,$$

qui sont manifestement consistantes.

Remarque 2.87. *On ne démontrera pas ce théorème. Sa preuve, sous la forme énoncée ci-dessus, est plutôt simple, mais relève de la théorie de la mesure et n'enseigne pas grand chose sur la théorie des probabilités. Il existe des versions de ce résultat pour des espaces beaucoup plus généraux, mais la version donnée ci-dessus suffira pour nos besoins dans ce cours.*


THE BEST WAY TO MAKE THIS DECISION IS BY CALCULATING THE EXPECTED VALUE OF EACH POSSIBLE OUTCOME.
www.dilbert.com scottadams@aol.com
YOU MULTIPLY THE. . .
MUST PRETEND TO BE DEAD.
7-13-05 ©2005 Scott Adams, Inc./Dist. by UFS, Inc.
I SENSE THAT WE'RE DONE HERE.
I HOPE THE DEAD SOME-TIMES COVER THEIR EARS.

©Scott Adams

3 Marche aléatoire simple sur $\mathbb{Z}$

The concept of a random walk is simple but rich for its many applications, not only in finance but also in physics and the description of natural phenomena. It is arguably one of the most founding concepts in modern physics as well as in finance [...]

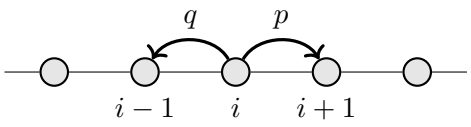
Didier Sornette

Les marches aléatoires forment une classe très importante de processus stochastiques (c'est-à-dire une suite de variables aléatoires, en général dépendantes, indexées par un paramètre que l'on identifie au temps), ayant de multiples connexions avec d'autres sujets en théorie des probabilités, mais également en analyse, en algèbre, etc. Dans ce chapitre, nous discuterons quelques propriétés élémentaires, mais parfois surprenantes, des marches aléatoires simples sur $\mathbb{Z}$.

3.1 Description du processus

Commençons par donner une description informelle de ce processus. Celui-ci modélise l'évolution (aléatoire) d'une particule sur les sommets du graphe $\mathbb{Z}$, selon les règles suivantes :

- ▷ la particule part (au temps $n = 0$) du sommet $a \in \mathbb{Z}$;
- ▷ la particule se déplace aux temps $n = 1, 2, \ldots$;
- ▷ les déplacements se font du sommet $i \in \mathbb{Z}$ occupé à cet instant vers l'un de ses voisins, $i - 1$ et $i + 1$;
- ▷ la probabilité de se déplacer vers le sommet $i + 1$ est p;
- ▷ les sauts sont indépendants.



On notera S_n, $n \in \mathbb{N}$, la position de la particule au temps n. Comme dans la Section 2.9, on considère un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ sur lequel est définie une suite infinie $(X_k)_{k \geqslant 1}$ de variables aléatoires réelles i.i.d. telles que

$$\mathbb{P}(X_k = 1) = p \qquad \text{et} \qquad \mathbb{P}(X_k = -1) = q \coloneqq 1 - p.$$

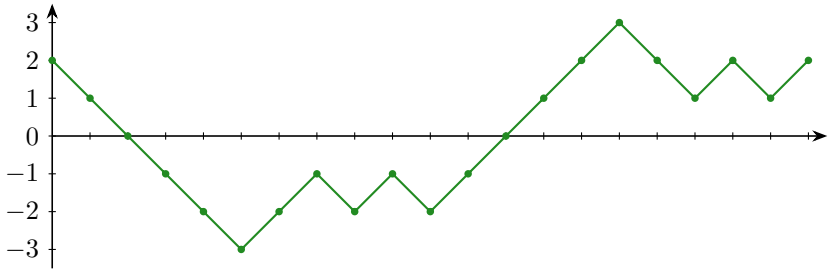


FIGURE 3.1: Les 20 premiers pas d'une trajectoire de la marche aléatoire simple partant de $a = 2$.

Le processus $(S_n)_{n\in\mathbb{N}}$ est alors défini de la façon suivante :

$$\forall n \in \mathbb{N}, \qquad S_n \coloneqq a + \sum_{k=1}^{n} X_k.$$

La loi du processus $(S_n)_{n\in\mathbb{N}}$ ainsi défini est notée $\mathbb{P}_{p,a}$. On appelle $(S_n)_{n\in\mathbb{N}}$ la **marche aléatoire simple**[1] sur $\mathbb{Z}$ partant de a et de paramètre p. Dans le cas où $p = 1/2$, on dit que la marche est **symétrique**.

Il suit de la définition de $(S_n)_{n\in\mathbb{N}}$ que, pour tout $n \in \mathbb{N}^*$ et toute trajectoire $(s_0, s_1, \ldots, s_n) \in \mathbb{Z}^{n+1}$, on a

$$\mathbb{P}_{p,a}(S_0 = s_0, S_1 = s_1, \ldots, S_n = s_n) = \delta_{s_0,a} \prod_{i=1}^{n} \mathbb{P}(X_i = s_i - s_{i-1}). \tag{3.1}$$

Note : Dans la suite de ce chapitre, afin d'alléger l'écriture, on écrira simplement $\mathbb{P}_a$ au lieu de $\mathbb{P}_{p,a}$. Les résultats seront toujours soit valides pour tout $p \in [0, 1]$, soit uniquement dans le cas symétrique, $p = \frac{1}{2}$, auquel cas cela sera spécifié explicitement.

3.2 Quelques propriétés importantes

Notons $\mathcal{F}_n \coloneqq \sigma(X_1, \ldots, X_n)$ la tribu des événements ne dépendant que des n premiers pas de la marche.

Lemme 3.1. *1. **(Homogénéité spatiale)** Pour tout $a \in \mathbb{Z}$,*

$$\mathbb{P}_a(S_0 = s_0, \ldots, S_n = s_n) = \mathbb{P}_0(S_0 = s_0 - a, \ldots, S_n = s_n - a).$$

*2. **(Propriété de Markov**[2]**)** Pour tout $B \in \mathcal{F}_n$ et tout $s \in \mathbb{Z}$ tels que $\mathbb{P}_a(S_n = s, B) > 0$, on a*

$$\mathbb{P}_a\big((S_n, S_{n+1}, \ldots) \in A \;\big|\; S_n = s, B\big) = \mathbb{P}_s\big((S_0, S_1, \ldots) \in A\big),$$

pour tout ensemble de trajectoires A mesurable.

Il est important de comprendre intuitivement ce qu'affirme la propriété de Markov : conditionnellement à $S_n = s$, ce qui a pu arriver à la marche jusqu'au temps n n'a pas d'influence sur son comportement après le temps n.

1. Le qualificatif *simple* fait référence au fait que la marche ne peut se déplacer que d'un point de $\mathbb{Z}$ vers l'un des deux points voisins.

2. Andreï Andreïevitch Markov (1856, Riazan – 1922, Saint-Pétersbourg), mathématicien russe, considéré comme le fondateur de la théorie des processus stochastiques.

Démonstration. La première affirmation est immédiate, puisque (3.1) implique que

$$\mathbb{P}_a(S_0 = s_0, \dots, S_n = s_n) = \delta_{s_0,a} \prod_{i=1}^{n} \mathbb{P}(X_i = s_i - s_{i-1}),$$
$$\mathbb{P}_0(S_0 = s_0 - a, \dots, S_n = s_n - a) = \delta_{s_0-a,0} \prod_{i=1}^{n} \mathbb{P}(X_i = s_i - s_{i-1}).$$

Pour la seconde propriété, on observe que

$$\begin{aligned}\mathbb{P}_a((S_n, S_{n+1}, S_{n+2}, \dots) \in A \,|\, S_n = s, B) &= \mathbb{P}_a((s, s + X_{n+1}, s + X_{n+1} + X_{n+2}, \dots) \in A \,|\, S_n = s, B) \\ &= \mathbb{P}((s, s + X_{n+1}, s + X_{n+1} + X_{n+2}, \dots) \in A) \\ &= \mathbb{P}_s\big((S_0, S_1, S_2, \dots) \in A\big),\end{aligned}$$

où l'on a utilisé le fait que $\{S_n = s\} \cap B \in \mathcal{F}_n$ est indépendant de tout événement appartenant à $\sigma((X_{n+k})_{k\geqslant 1})$ pour la deuxième égalité, et le fait que $(s, s + X_{n+1}, s + X_{n+1} + X_{n+2}, \dots) \overset{\text{loi}}{=} (s, s + X_1, s + X_1 + X_2, \dots)$, puisque les variables aléatoires $(X_i)_{i\geqslant 1}$ sont i.i.d.. □

Un peu de combinatoire élémentaire permet de déterminer aisément la loi de S_n.

Lemme 3.2. *Pour tout $n \geqslant 1$,*

$$\mathbb{P}_a(S_n = s) = \binom{n}{\frac{n+s-a}{2}} p^{(n+s-a)/2} q^{(n-s+a)/2},$$

si $s - a \in \{-n + 2k \,|\, 0 \leqslant k \leqslant n\}$, et $\mathbb{P}_a(S_n = s) = 0$ sinon.

En particulier, il suit de la formule de Stirling (Lemme A.19) que, lorsque $p = q = \frac{1}{2}$ et $n \to \infty$,

$$\mathbb{P}_0(S_{2n} = 0) = \frac{1 + \mathsf{o}(1)}{\sqrt{\pi}} n^{-1/2}. \tag{3.2}$$

Démonstration. Par homogénéité spatiale, il suffit de considérer le cas $a = 0$.

Soit $N_\pm := \#\{1 \leqslant i \leqslant n \,|\, X_i = \pm 1\}$. On a manifestement $n = N_+ + N_-$ et $S_n = N_+ - N_-$. Par conséquent, $S_n = s$ si et seulement si $N_+ = \frac{n+s}{2}$ et $N_- = \frac{n-s}{2}$. Évidemment, ceci n'est possible que si $n + s$ est pair, et si $|s| \leqslant n$. On a donc, puisque $N_+ \sim \mathsf{binom}(n, p)$,

$$\mathbb{P}_0(S_n = s) = \mathbb{P}(N_+ = (n + s)/2) = \binom{n}{\frac{n+s}{2}} p^{\frac{n+s}{2}} (1 - p)^{\frac{n-s}{2}}.$$ □

3.3 Ruine du joueur

Parmi les nombreuses interprétations de la marche aléatoire simple sur $\mathbb{Z}$, une des plus classiques est la suivante : S_n représente la fortune d'un joueur après la n-ième manche d'un jeu au cours duquel il fait une mise égale à 1 à chaque manche (pourvu que sa fortune soit strictement positive), et la double avec probabilité $0 < p < 1$ (sa fortune augmentant donc d'une unité), ou la perd avec probabilité $q = 1 - p$ (sa fortune diminuant ainsi d'une unité). a représente alors la fortune initiale du joueur.

Dans cette interprétation, le problème suivant est naturel. Le joueur ne peut continuer à jouer qu'aussi longtemps que sa fortune reste strictement positive. Supposons qu'il décide qu'il arrêtera de jouer lorsqu'il aura atteint son objectif de parvenir à une fortune égale $N > a$. Quelle est la probabilité qu'il soit ruiné avant de réaliser son objectif ?

Pour $b \in \mathbb{Z}$, notons $\overline{\tau}_b := \inf\{n \in \mathbb{N} \mid S_n = b\}$ le temps de première visite en b (notez que $\overline{\tau}_b = 0$ si b correspond au point de départ de la marche), avec la convention habituelle que $\inf \varnothing = +\infty$. Observez que $\overline{\tau}_b$ est une variable aléatoire (à valeurs dans $\mathbb{N} \cup \{+\infty\}$). En effet,

$$\{\overline{\tau}_b = k\} = \{S_0 \neq b, S_1 \neq b, \dots, S_{k-1} \neq b, S_k = b\} \in \mathcal{F}_k \subset \mathcal{F},$$
$$\{\overline{\tau}_b = +\infty\} = \bigcap_{k \geqslant 1} \{\overline{\tau}_b > k\} = \bigcap_{k \geqslant 1} \{S_0 \neq b, S_1 \neq b, \dots, S_k \neq b\} \in \mathcal{F}.$$

L'événement « le joueur est ruiné avant d'atteindre son objectif » correspond à $A := \{\overline{\tau}_0 < \overline{\tau}_N\}$. Il suit alors de la loi de la probabilité totale et de la propriété de Markov que, pour tout $0 < a < N$,

$$\begin{aligned}\mathbb{P}_a(A) &= \mathbb{P}_a(A \mid S_1 = a+1)\,\mathbb{P}_a(S_1 = a+1) + \mathbb{P}_a(A \mid S_1 = a-1)\,\mathbb{P}_a(S_1 = a-1) \\ &= p\,\mathbb{P}_{a+1}(A) + q\,\mathbb{P}_{a-1}(A).\end{aligned}$$

On obtient donc, pour tout $0 < a < N$,

$$\mathbb{P}_{a+1}(A) - \mathbb{P}_a(A) = \frac{q}{p}(\mathbb{P}_a(A) - \mathbb{P}_{a-1}(A)) = \frac{q^2}{p^2}(\mathbb{P}_{a-1}(A) - \mathbb{P}_{a-2}(A)) = \cdots = \frac{q^a}{p^a}(\mathbb{P}_1(A) - \mathbb{P}_0(A)).$$

Par conséquent, puisque $\mathbb{P}_0(A) = 1$,

$$\mathbb{P}_a(A) = \mathbb{P}_0(A) + \sum_{k=1}^{a}(\mathbb{P}_k(A) - \mathbb{P}_{k-1}(A)) = 1 + \sum_{k=1}^{a} \frac{q^{k-1}}{p^{k-1}}(\mathbb{P}_1(A) - 1). \tag{3.3}$$

Cas symétrique : $p = \frac{1}{2}$. Dans ce cas, $q/p = 1$ et (3.3) implique que

$$\mathbb{P}_a(A) = 1 + a(\mathbb{P}_1(A) - 1).$$

Comme $\mathbb{P}_N(A) = 0$, on en déduit que $\mathbb{P}_1(A) - 1 = -\frac{1}{N}$ et donc que

$$\mathbb{P}_a(A) = 1 - \frac{a}{N}.$$

Cas non symétrique : $p \neq \frac{1}{2}$. Dans ce cas, il suit de (3.3) que

$$\mathbb{P}_a(A) = 1 + \frac{1 - (q/p)^a}{1 - (q/p)}(\mathbb{P}_1(A) - 1).$$

Cette fois, on tire de $\mathbb{P}_N(A) = 0$ que

$$\mathbb{P}_1(A) - 1 = -\frac{1 - (q/p)}{1 - (q/p)^N}.$$

Par conséquent,

$$\mathbb{P}_a(A) = \frac{(q/p)^a - (q/p)^N}{1 - (q/p)^N}.$$

3.4 Le premier retour au point de départ

Nous allons à présent étudier le temps $\tau_0 := \inf\{n \in \mathbb{N}^* \mid S_n = 0\}$ mis par la marche partant de 0 pour retourner à son point de départ.

Lemme 3.3. *Pour tout $n \geqslant 1$,*

$$\mathbb{P}_0(\tau_0 > n, S_n = b) = \frac{|b|}{n}\mathbb{P}_0(S_n = b).$$

En particulier,

$$\mathbb{P}_0(\tau_0 > n) = \frac{1}{n}\mathbb{E}_0\big[|S_n|\big].$$

Démonstration. La première identité est triviale lorsque b n'est pas une valeur atteignable par S_n (c'est-à-dire, lorsque $n+b$ est impair ou $|b| > n$). Elle est également triviale lorsque $|b| = n$, puisque la trajectoire n'a alors pas le temps de revisiter 0. On peut donc supposer que $|b| \leqslant n-1$ et que $n+b$ est pair.

Observez que l'événement $\{\tau_0 > n, S_n = b\} \in \mathcal{F}_n$; il suffit donc de s'intéresser aux premiers n pas de la marche. Comme toutes les portions de trajectoire $(s_0 = 0, s_1, \ldots, s_{n-1}, s_n = b)$ joignant $(0,0)$ à (n,b) sont équiprobables (de probabilité $p^{(n+b)/2}q^{(n-b)/2}$), il suffit de déterminer combien d'entre elles ne revisitent pas l'origine.

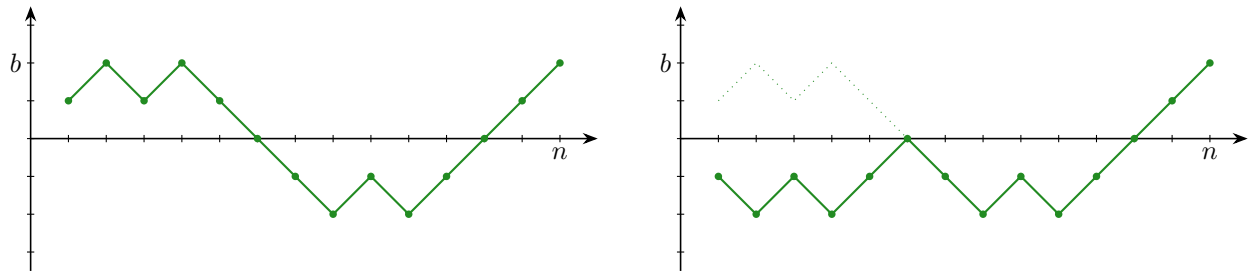
On suppose, sans perte de généralité, que $b > 0$. Dans ce cas, toutes les trajectoires contribuant à l'événement $\{\tau_0 > n, S_n = b\}$ satisfont $S_1 = 1$. Introduisons donc les ensembles suivants :

- ▷ $\mathcal{T}_+[(1,1),(n,b)]$: ensemble de toutes les portions de trajectoires joignant $(1,1)$ à (n,b) sans intersecter l'axe des abscisses.
- ▷ $\mathcal{T}_0[(1,1),(n,b)]$: ensemble de toutes les portions de trajectoires joignant $(1,1)$ à (n,b) intersectant l'axe des abscisses.
- ▷ $\mathcal{T}[(1,1),(n,b)]$: ensemble de toutes les portions de trajectoires joignant $(1,1)$ à (n,b).

Manifestement,

$$\#\mathcal{T}_+[(1,1),(n,b)] = \#\mathcal{T}[(1,1),(n,b)] - \#\mathcal{T}_0[(1,1),(n,b)].$$

On a vu, dans la preuve du Lemme 3.2, que $\#\mathcal{T}[(1,1),(n,b)] = \binom{n-1}{\frac{n+b}{2}-1}$. Il nous faut donc déterminer $\#\mathcal{T}_0[(1,1),(n,b)]$. L'observation essentielle, appelée **principe de réflexion**, est la suivante : l'ensemble $\mathcal{T}_0[(1,1),(n,b)]$ est en bijection avec l'ensemble $\mathcal{T}[(1,-1),(n,b)]$ des portions de trajectoires joignant $(1,-1)$ à (n,b) : il suffit de réfléchir les τ_0 premiers pas de la trajectoire à travers l'axe des abscisses, tout en conservant intacte la seconde partie de la trajectoire.



Or, $\#\mathcal{T}[(1,-1),(n,b)] = \binom{n-1}{\frac{n+b}{2}}$, d'où l'on déduit que

$$\#\mathcal{T}_+[(1,1),(n,b)] = \binom{n-1}{\frac{n+b}{2}-1} - \binom{n-1}{\frac{n+b}{2}} = \frac{b}{n}\binom{n}{\frac{n+b}{2}}. \tag{3.4}$$

Par conséquent,

$$\mathbb{P}_0(\tau_0 > n, S_n = b) = \frac{b}{n}\binom{n}{\frac{n+b}{2}}p^{(n+b)/2}q^{(n-b)/2} = \frac{b}{n}\,\mathbb{P}_0(S_n = b),$$

la dernière égalité résultant du Lemme 3.2. La seconde assertion du lemme suit en sommant sur b. □

On peut facilement déduire du résultat précédent une relation très simple dans le cas symétrique, qui possède une conséquence remarquable.

Théorème 3.4. *Dans le cas symétrique,*

$$\forall n \in \mathbb{N}^*, \qquad \mathbb{P}_0(\tau_0 > 2n) = \mathbb{P}_0(S_{2n} = 0).$$

En particulier, la marche aléatoire simple symétrique sur $\mathbb{Z}$ retourne à son point de départ presque sûrement :

$$\mathbb{P}_0(\tau_0 < \infty) = 1.$$

Démonstration. En appliquant le résultat du lemme précédent et en exploitant la symétrie, on obtient

$$\begin{aligned}
\mathbb{P}_0(\tau_0 > 2n) &= \sum_{k=1}^{n} \frac{2k}{2n} \big(\mathbb{P}_0(S_{2n} = -2k) + \mathbb{P}_0(S_{2n} = 2k)\big) \\
&= 2\sum_{k=1}^{n} \frac{2k}{2n} \mathbb{P}_0(S_{2n} = 2k) \\
&= 2\sum_{k=1}^{n} \frac{2k}{2n} \binom{2n}{n+k} 2^{-2n} \\
&= 2^{-2n+1} + 2^{-2n+1} \sum_{k=1}^{n-1} \left\{ \binom{2n-1}{n+k-1} - \binom{2n-1}{n+k} \right\} \\
&= 2^{-2n+1} + 2^{-2n+1} \left\{ \binom{2n-1}{n} - \binom{2n-1}{2n-1} \right\} \\
&= 2^{-2n+1} \binom{2n-1}{n} = 2^{-2n} \binom{2n}{n} = \mathbb{P}_0(S_{2n} = 0),
\end{aligned}$$

la quatrième ligne suivant de (3.4) et la cinquième d'un télescopage. Combiné à (3.2), le résultat précédent montre que

$$\mathbb{P}_0(\tau_0 > 2n) = \frac{1 + o(1)}{\sqrt{\pi}} n^{-1/2}. \tag{3.5}$$

En particulier, en utilisant le Lemme 1.4, on en conclut que

$$\mathbb{P}_0(\tau_0 = +\infty) = \lim_{n\to\infty} \mathbb{P}_0(\tau_0 > n) = 0. \qquad \square$$

Remarque 3.5. *Bien que presque sûrement fini, le retour peut prendre beaucoup de temps : l'espérance de τ_0 est infinie ! En effet, il suit de (3.5) et du Lemme 2.57 que*

$$\mathbb{E}_0[\tau_0/2] = \sum_{n\geqslant 0} \mathbb{P}_0(\tau_0 > 2n) = \infty.$$

Ceci conduit à des propriétés assez contre-intuitives. Considérons, par exemple, une suite de $n \gg 1$ lancers d'une pièce équilibrée. Notons R_n le nombre d'instants où il y a eu égalité entre le nombre de « pile » et de « face » obtenus (ce qui correspond au nombre de retours de la marche symétrique à son point de départ lors des n premiers pas). On pourrait s'attendre à ce que $\mathbb{E}[R_{2n}] \approx 2\mathbb{E}[R_n]$, c'est-à-dire que si on lance la pièce deux fois plus souvent, alors on observera en moyenne approximativement deux fois plus souvent l'égalité. Ce n'est pas du tout ce qui se passe. En effet,

$$\mathbb{E}_0[R_n] = \mathbb{E}_0\Big[\sum_{k=1}^{n} \mathbf{1}_{\{S_k=0\}}\Big] = \sum_{k=1}^{n} \mathbb{P}_0(S_k = 0) = \mathsf{O}(\sqrt{n}),$$

et la fréquence des retours tend donc vers 0 comme $n^{-1/2}$. Il faudra donc lancer la pièce quatre fois plus souvent pour doubler le nombre moyen d'instants où l'égalité a lieu !

Le résultat suivant fournit une formule explicite pour la loi du temps de premier retour en 0.

Corollaire 3.6. *Pour tout $n \in \mathbb{N}^*$,*

$$\mathbb{P}_0(\tau_0 = n) = \frac{q}{n-1}\mathbb{P}_0(S_{n-1} = 1) + \frac{p}{n-1}\mathbb{P}_0(S_{n-1} = -1).$$

Démonstration. Puisque $\{\tau_0 = n\} = \{\tau_0 > n-1\} \cap \{S_n = 0\}$, on déduit de la propriété de Markov que

$$\begin{aligned}\mathbb{P}_0(\tau_0 = n) &= \mathbb{P}_0(\tau_0 = n, S_{n-1} = 1) + \mathbb{P}_0(\tau_0 = n, S_{n-1} = -1)\\ &= \mathbb{P}_0(S_n = 0 \,|\, \tau_0 > n-1, S_{n-1} = 1)\,\mathbb{P}_0(\tau_0 > n-1, S_{n-1} = 1)\\ &\quad + \mathbb{P}_0(S_n = 0 \,|\, \tau_0 > n-1, S_{n-1} = -1)\,\mathbb{P}_0(\tau_0 > n-1, S_{n-1} = -1)\\ &= q\,\frac{1}{n-1}\mathbb{P}_0(S_{n-1} = 1) + p\,\frac{1}{n-1}\mathbb{P}_0(S_{n-1} = -1),\end{aligned}$$

où l'on a utilisé le résultat du Lemme 3.3. □

Dans le cas de la marche aléatoire simple symétrique, on obtient donc

$$\mathbb{P}_0(\tau_0 = n) = \frac{1}{2n-2}\mathbb{P}_0(|S_{n-1}| = 1) = \frac{1}{n-1}\mathbb{P}_0(S_n = 0),$$

puisque $\mathbb{P}_0(S_n = 0 \,|\, |S_{n-1}| = 1) = \frac{1}{2}$. (On aurait évidemment également pu tirer ce résultat directement du Théorème 3.4.)

En particulier, lorsque $n \to \infty$,

$$\mathbb{P}_0(\tau_0 = 2n) = \frac{1 + o(1)}{2\sqrt{\pi}} n^{-3/2}.$$

3.5 Première visite en un sommet

Nous avons montré que, dans le cas symétrique, τ_0 est presque-sûrement fini. Nous allons à présent montrer qu'il en est de même du temps τ_b de la première visite en $b \neq 0$.

Théorème 3.7. *On considère la marche symétrique sur $\mathbb{Z}$. Alors,*

$$\forall b \in \mathbb{Z}, \qquad \mathbb{P}_0(\tau_b < \infty) = 1.$$

Démonstration. On peut supposer, sans perte de généralité, que $b \geqslant 0$. On a déjà établi que $\mathbb{P}_0(\tau_0 < \infty) = 1$. En conditionnant sur X_1, on voit que la fonction $b \mapsto \mathbb{P}_0(\exists n \geqslant 0 \,:\, S_n = b)$ est solution de

$$\begin{cases} f(b) = \frac{1}{2}(f(b+1) + f(b-1)) & \text{pour tout } b \in \mathbb{N}^*,\\ f(0) = 1. \end{cases}$$

Évidemment, les solutions de cette équation sont données par les fonctions de la forme $f(b) = 1 + \alpha b$, $\alpha \in \mathbb{R}$. Observons que l'on doit avoir $0 \leqslant f(b) \leqslant 1$ pour tout $b \in \mathbb{N}$, puisque l'on cherche une probabilité. Or, l'unique solution bornée est obtenue lorsque $\alpha = 0$, ce qui correspond à $f \equiv 1$. On en conclut donc que

$$\forall b \neq 0, \qquad \mathbb{P}_0(\tau_b < \infty) = \mathbb{P}_0(\exists n \geqslant 0 \,:\, S_n = b) = 1.$$ □

En particulier, la marche aléatoire simple symétrique sur $\mathbb{Z}$ visite presque sûrement tous les sommets de $\mathbb{Z}$:

$$\mathbb{P}_0(\exists b \in \mathbb{Z},\, \tau_b = \infty) = \mathbb{P}_0\Big(\bigcup_{b\in\mathbb{Z}} \{\tau_b = \infty\}\Big) \leqslant \sum_{b\in\mathbb{Z}} \mathbb{P}_0(\tau_b = \infty) = 0.$$

En fait, il est facile de voir qu'elle le fait presque-sûrement infiniment souvent. Notons

$$N_b := \sum_{k=1}^{\infty} \mathbf{1}_{\{S_k = b\}}$$

le nombre de visites en $b \in \mathbb{Z}$. N_b est bien une variable aléatoire à valeurs dans $\overline{\mathbb{N}} := \mathbb{N} \cup \{+\infty\}$, puisque, pour tout $k \in \mathbb{N}$,

$$\{N_b = k\} = \bigcup_{\substack{I \subset \mathbb{N}^* \\ |I|=k}} \Big[\bigcap_{\ell \in I} \{S_\ell = b\} \cap \bigcap_{m \notin I} \{S_m \neq b\} \Big] \in \mathcal{F}.$$

En particulier, $\{N_b < \infty\} = \bigcup_{k \in \mathbb{N}} \{N_b = k\} \in \mathcal{F}$ et on a donc également $\{N_b = +\infty\} = \{N_b < \infty\}^c \in \mathcal{F}$.

Théorème 3.8. *Dans le cas symétrique,*

$$\mathbb{P}_0(N_b = \infty \text{ pour tous les } b \in \mathbb{Z}) = 1.$$

Démonstration. Soit $\tau_b^{(n)}$ le temps de la n-ième visite en b (avec $\tau_b^{(n)} := \infty$ si $N_b < n$, et $\tau_b^{(0)} := 0$). En conditionnant sur le temps auquel se produit le k-ième retour, on obtient, pour tout $k \in \mathbb{N}$,

$$\mathbb{P}_0(N_b = k) = \sum_{\ell \geqslant k} \mathbb{P}_0(N_b = k \,|\, \tau_b^{(k)} = \ell)\, \mathbb{P}_0(\tau_b^{(k)} = \ell).$$

La propriété de Markov implique donc, puisque $\{\tau_b^{(k)} = \ell\}$ ne dépend que des ℓ premiers pas de la trajectoire et implique que $S_\ell = b$,

$$\begin{aligned} \mathbb{P}_0(N_b = k \,|\, \tau_b^{(k)} = \ell) &= \mathbb{P}_0(S_j \neq b, \forall j > \ell \,|\, \tau_b^{(k)} = \ell) \\ &= \mathbb{P}_0(S_j \neq b, \forall j > \ell \,|\, S_\ell = b, \tau_b^{(k)} = \ell) \\ &= \mathbb{P}_0(S_j \neq 0, \forall j > 0) = \mathbb{P}_0(\tau_0 = \infty) = 0. \end{aligned}$$

Il s'ensuit que $\mathbb{P}_0(N_b = k) = 0$, pour tout $k \in \mathbb{N}$. Par conséquent,

$$\mathbb{P}_0(N_b < \infty) = \sum_{k \geqslant 0} \mathbb{P}_0(N_b = k) = 0, \qquad \forall b \in \mathbb{Z}.$$

La conclusion suit, puisque

$$\mathbb{P}_0(\exists b \in \mathbb{Z} :\ N_b < \infty) \leqslant \sum_{b \in \mathbb{Z}} \mathbb{P}_0(N_b < \infty) = 0. \qquad \square$$

La loi de τ_b peut également être déterminée explicitement pour toutes les valeurs de p.

Théorème 3.9. *Pour tout $b \neq 0$,*

$$\mathbb{P}_0(\tau_b = n) = \mathbb{P}_0(\tau_0 > n, S_n = b) = \frac{|b|}{n} \mathbb{P}_0(S_n = b).$$

Démonstration. Cette preuve repose sur la méthode du *renversement du temps*. On associe à une portion de trajectoire

$$(0, S_1, S_2, \ldots, S_n) = (0, X_1, X_1 + X_2, \ldots, X_1 + \cdots + X_n),$$

la portion de trajectoire renversée (voir Fig. 3.2)

$$(0, R_1, R_2, \ldots, R_n) := (0, X_n, X_n + X_{n-1}, \ldots, X_n + \cdots + X_1).$$

Manifestement, $(0, S_1, S_2, \ldots, S_n) \overset{\text{loi}}{=} (0, R_1, R_2, \ldots, R_n)$. Observez à présent que $S_n = b > 0$ et $\tau_0 > n$ si et seulement si $R_n = b$ et $R_n - R_i = X_1 + \cdots + X_{n-i} \neq 0$ pour tout $0 \leqslant i < n$, ce qui signifie que la première visite de la marche renversée au point b a lieu au temps n (ceci est également illustré sur la Figure 3.2). On a donc démontré le résultat suivant :

$$\mathbb{P}_0(S_n = b, \tau_0 > n) = \mathbb{P}_0(R_n = b, \forall 0 \leqslant i < n,\ R_i \neq b) = \mathbb{P}_0(S_n = b, \forall 0 \leqslant i < n,\ S_i \neq b) = \mathbb{P}_0(\tau_b = n),$$

la deuxième égalité suivant de l'égalité en loi. La conclusion suit donc du Lemme 3.3. $\square$

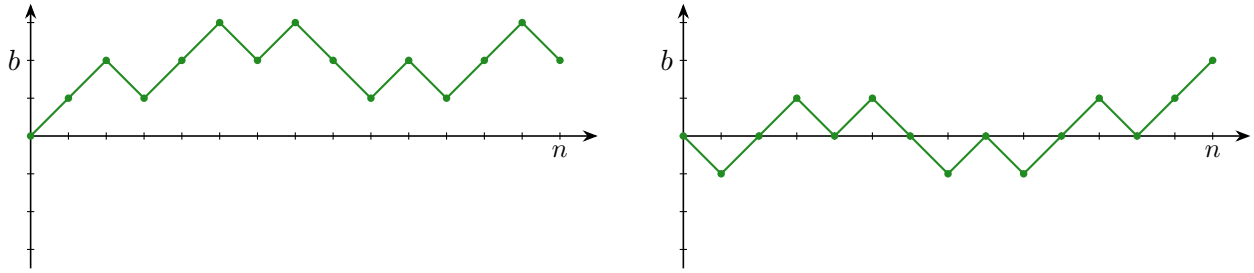


FIGURE 3.2: Une portion de trajectoire $(0, S_1, \dots, S_n)$ et la portion de trajectoire renversée $(0, R_1, \dots, R_n)$.

Ce résultat a une conséquence surprenante.

Corollaire 3.10. *Dans le cas symétrique, le nombre moyen de visites de la marche (partant de* 0*) en un sommet* $b \neq 0$ *quelconque avant de retourner à l'origine est égal à* 1.

Démonstration. Il suit du théorème précédent que le nombre moyen de visites au sommet $b \neq 0$ avant le premier retour en 0 est égal à

$$\mathbb{E}_0\Big[\sum_{n\geqslant 1} \mathbf{1}_{\{\tau_0>n, S_n=b\}}\Big] = \sum_{n\geqslant 1} \mathbb{P}_0(\tau_0 > n, S_n = b) = \sum_{n\geqslant 1} \mathbb{P}_0(\tau_b = n) = \mathbb{P}_0(\tau_b < \infty).$$

La conclusion suit donc du Théorème 3.7. □

On considère le jeu suivant : on jette successivement une pièce bien équilibrée et le joueur gagne un franc à chaque fois que le nombre de « pile » excède le nombre de « face » par exactement m lancers ; le jeu s'interrompt dès que les nombres de « pile » et de « face » obtenus sont égaux. Quelle est la mise initiale équitable pour le joueur ? Le corollaire ci-dessus montre que celle-ci est de 1 franc, *quelle que soit la valeur de* m !

3.6 La loi de l'arc-sinus pour la dernière visite en 0

On peut également s'intéresser au moment de la *dernière* visite en 0 au cours des $2n$ premiers pas :

$$\nu_0(2n) := \max\{0 \leqslant k \leqslant 2n \mid S_k = 0\}.$$

Théorème 3.11 (Loi de l'arcsinus pour la dernière visite en 0)**.** *On suppose que* $p = 1/2$*. Pour tout* $0 \leqslant k \leqslant n$*,*

$$\mathbb{P}_0(\nu_0(2n) = 2k) = \mathbb{P}_0(S_{2k} = 0)\mathbb{P}_0(S_{2n-2k} = 0).$$

En particulier, pour tout $0 < \alpha < \beta < 1$*,*

$$\lim_{n\to\infty} \mathbb{P}_0\Big(\frac{\nu_0(2n)}{2n} \in [\alpha, \beta]\Big) = \frac{2}{\pi}\left(\arcsin\sqrt{\beta} - \arcsin\sqrt{\alpha}\right).$$

Démonstration. La première affirmation suit de l'observation suivante :

$$\begin{aligned}
\mathbb{P}_0(\nu_0(2n) = 2k) &= \mathbb{P}_0(S_{2k} = 0, S_{2k+1} \neq 0, \dots, S_{2n} \neq 0) \\
&= \mathbb{P}_0(S_{2k} = 0)\, \mathbb{P}_0(S_{2k+1} \neq 0, \dots, S_{2n} \neq 0 \mid S_{2k} = 0) \\
&= \mathbb{P}_0(S_{2k} = 0)\, \mathbb{P}_0(S_1 \neq 0, \dots, S_{2n-2k} \neq 0) \\
&= \mathbb{P}_0(S_{2k} = 0)\, \mathbb{P}_0(S_{2n-2k} = 0)\,,
\end{aligned}$$

la dernière identité résultant du Théorème 3.4.

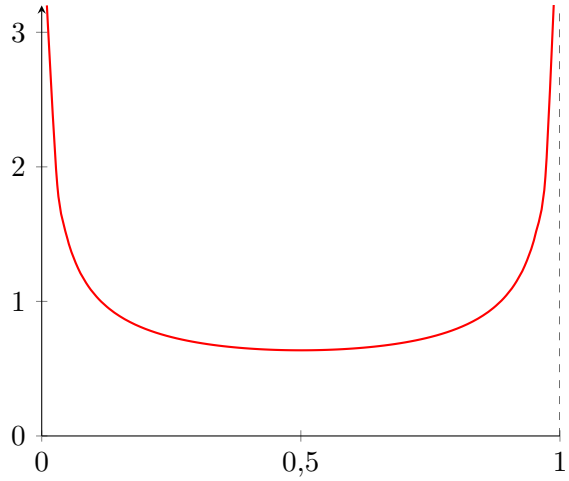


FIGURE 3.3: La densité de probabilité de la loi de l'arcsinus.

Pour la seconde affirmation, observons tout d'abord qu'une application de la formule de Stirling donne

$$\mathbb{P}_0(S_{2k}=0)\mathbb{P}_0(S_{2n-2k}=0)=\binom{2k}{k}\binom{2n-2k}{n-k}2^{-2n}=\frac{1+o(1)}{\pi\sqrt{k(n-k)}},$$

lorsque k et $n-k$ tendent tous deux vers l'infini. On a donc

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}_0\Big(\frac{\nu_0(2n)}{2n}\in[\alpha,\beta]\Big)&=\lim_{n\to\infty}\frac{1}{\pi n}\sum_{k=\lceil\alpha n\rceil}^{\lfloor\beta n\rfloor}\Big(\frac{k}{n}\big(1-\frac{k}{n}\big)\Big)^{-1/2}\\&=\frac{1}{\pi}\int_\alpha^\beta\big(x(1-x)\big)^{-1/2}\,\mathrm{d}x\\&=\frac{2}{\pi}\arcsin\sqrt{\beta}-\frac{2}{\pi}\arcsin\sqrt{\alpha}.\end{aligned}$$

□

Le lemme précédent a des conséquences peut-être assez surprenantes au premier abord : si l'on procède à un grand nombre de lancers à pile ou face, la dernière fois que le nombre de « pile » et le nombre de « face » obtenus ont coïncidé est proche du début ou de la fin de la série avec une probabilité substantielle : on a, par exemple (voir également la Figure 3.3),

$$\begin{aligned}\mathbb{P}_0(\nu(10\,000)\leqslant 100)&\cong\frac{2}{\pi}\arcsin\sqrt{0{,}01}\cong 6{,}4\%\,,\\\mathbb{P}_0(\nu(10\,000)\geqslant 9\,900)&\cong\frac{2}{\pi}\arcsin\sqrt{0{,}01}\cong 6{,}4\%\,,\\\mathbb{P}_0(\nu(10\,000)\leqslant 1\,000)&\cong\frac{2}{\pi}\arcsin\sqrt{0{,}1}\cong 20{,}5\%\,.\end{aligned}$$

3.7 La loi de l'arc-sinus pour les temps de séjour

Nous allons à présent établir un autre résultat classique, également contre-intuitif lorsqu'on le rencontre pour la première fois. Il concerne le temps total passé par la marche au-dessus de 0, pendant les $2n$ premiers pas. Plus précisément, on s'intéresse à la variable aléatoire

$$t_{2n}^+ := |\{0\leqslant i<2n\mid \max(S_i,S_{i+1})>0\}|$$

définissant le temps pendant lequel la marche est positive (*cf.* Fig. 3.4). Observez que t_{2n}^+ est nécessairement pair.

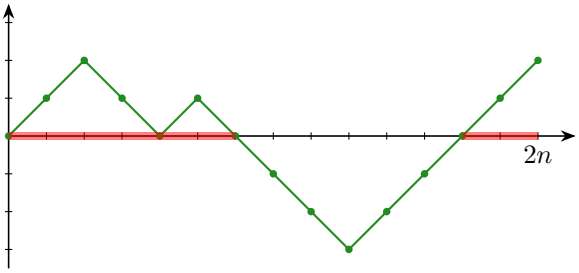


FIGURE 3.4: Le temps passé au-dessus de 0 jusqu'au temps $2n$, t^+_{2n}, correspond à la longueur des intervalles marqués en rouge.

Théorème 3.12 (Loi de l'arcsinus pour les temps de séjour). *Supposons $p = 1/2$. Alors, pour tout $0 \leqslant k \leqslant n$,*

$$\mathbb{P}_0(t^+_{2n} = 2k) = \mathbb{P}_0(S_{2k} = 0)\mathbb{P}_0(S_{2n-2k} = 0).$$

En particulier, pour tout $0 < \alpha < \beta < 1$,

$$\lim_{n\to\infty} \mathbb{P}_0\big(\frac{t^+_{2n}}{2n} \in [\alpha, \beta]\big) = \frac{2}{\pi}\left(\arcsin\sqrt{\beta} - \arcsin\sqrt{\alpha}\right).$$

Démonstration. Pour simplifier les notations, on introduit $f_{2n}(2k) := \mathbb{P}_0(t^+_{2n} = 2k)$, et $g_{2k} := \mathbb{P}_0(S_{2k} = 0)$. La première affirmation prend alors la forme

$$f_{2n}(2k) = g_{2k}g_{2n-2k}. \tag{3.6}$$

La première observation est que, par le Théorème 3.4,

$$\begin{aligned} g_{2n} = \mathbb{P}_0(S_{2n} = 0) = \mathbb{P}_0(\tau_0 > 2n) &= 2\mathbb{P}_0(S_1 = 1, S_2 \geqslant 1, \dots, S_{2n} \geqslant 1) \\ &= \mathbb{P}_0(S_2 \geqslant 1, \dots, S_{2n} \geqslant 1 \,|\, S_1 = 1) \\ &= \mathbb{P}_0(S_1 \geqslant 0, \dots, S_{2n-1} \geqslant 0) \\ &= \mathbb{P}_0(S_1 \geqslant 0, \dots, S_{2n-1} \geqslant 0, S_{2n} \geqslant 0) = f_{2n}(2n). \end{aligned}$$

L'avant-dernière identité suit du fait que, S_{2n-1} étant impair, $S_{2n-1} \geqslant 0$ implique que $S_{2n} \geqslant 0$. Ceci établit (3.6) lorsque $k = n$. L'identité pour $k = 0$ suit alors par symétrie.

Soit $k \in \{1, \dots, n-1\}$. Dans ce cas, lorsque $t^+_{2n} = 2k$, le temps τ_0 du premier retour à l'origine satisfait $\tau_0 = 2r$, avec $1 \leqslant r < n$. Pour $1 \leqslant k < \tau_0$, la marche reste toujours strictement positive ou strictement négative, chacune de ces deux possibilités ayant probabilité $1/2$ par symétrie. Par conséquent,

$$f_{2n}(2k) = \sum_{r=1}^{k} \tfrac{1}{2}\mathbb{P}_0(\tau_0 = 2r) f_{2n-2r}(2k-2r) + \sum_{r=1}^{n-k} \tfrac{1}{2}\mathbb{P}_0(\tau_0 = 2r) f_{2n-2r}(2k),$$

où la première somme prend en compte la contribution des trajectoires restant positives jusqu'au temps τ_0 et la seconde celle des trajectoires négatives jusqu'au temps τ_0.

Pour conclure la preuve, on fait une récurrence. On a déjà vérifié la validité de (3.6) pour tous les $0 \leqslant k \leqslant n$ lorsque $n = 1$. Supposons donc (3.6) vérifiée pour tous les $0 \leqslant k \leqslant n$ lorsque $n < m$. Alors, notant $h_{2r} = \mathbb{P}_0(\tau_0 = 2r)$, il suit de la précédente identité et de l'hypothèse d'induction que

$$f_{2m}(2k) = \tfrac{1}{2}\sum_{r=1}^{k} h_{2r} g_{2k-2r} g_{2m-2k} + \tfrac{1}{2}\sum_{r=1}^{m-k} h_{2r} g_{2k} g_{2m-2r-2k} = g_{2k} g_{2m-2k},$$

ce qui conclut la preuve de (3.6). La dernière identité suit de l'observation que, pour tout $\ell \geqslant 1$,

$$\mathbb{P}_0(S_{2\ell} = 0) = \sum_{r=1}^{\ell} \mathbb{P}_0(S_{2\ell} = 0 \,|\, \tau_0 = 2r)\mathbb{P}_0(\tau_0 = 2r) = \sum_{r=1}^{\ell} \mathbb{P}_0(S_{2\ell-2r} = 0)\mathbb{P}_0(\tau_0 = 2r),$$

c'est-à-dire $g_{2\ell} = \sum_{r=1}^{\ell} g_{2\ell-2r} h_{2r}$.

La seconde affirmation du théorème a déjà été démontrée dans la preuve du Théorème 3.11. □

Discutons à présent quelques conséquences de ce résultat. L'intuition (ainsi qu'une mauvaise compréhension de ce qu'affirme la loi des grands nombres) pourrait laisser à penser qu'après un temps $n \gg 1$, la fraction du temps passé de chaque côté de l'origine devrait être de l'ordre de $1/2$. Or ce n'est pas du tout ce qui a lieu (voir aussi la Figure 3.3) : avec probabilité $1/5$, la marche passera près de $97{,}6\%$ de son temps du même côté de l'origine ; avec probabilité $1/10$, elle le fera pendant $99{,}4\%$ de son temps.

De façon plus imagée, supposons que de nombreuses paires de joueurs jouent à pile ou face, chaque paire lançant la pièce une fois par seconde pendant 365 jours. La loi de l'arcsinus montre alors que dans une partie sur 20, le joueur ayant mené le plus longtemps pendant la partie aura dominé son adversaire *pendant plus de* 364 *jours et* 10 *heures* !

La Figure 3.5 montre cinq trajectoires (de longueur 200) typiques d'une marche aléatoire simple symétrique sur $\mathbb{Z}$. Observez la présence de longues excursions (morceaux de trajectoires compris entre deux retours successifs à l'origine).

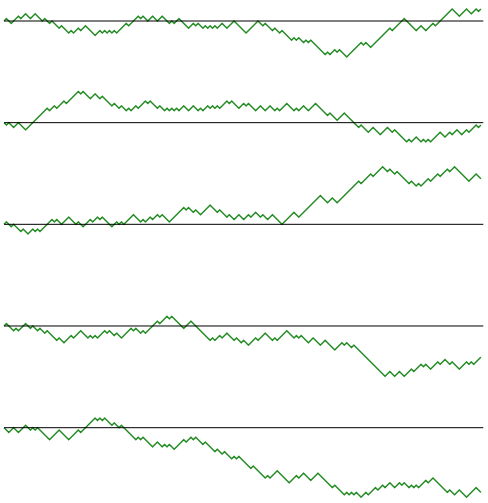
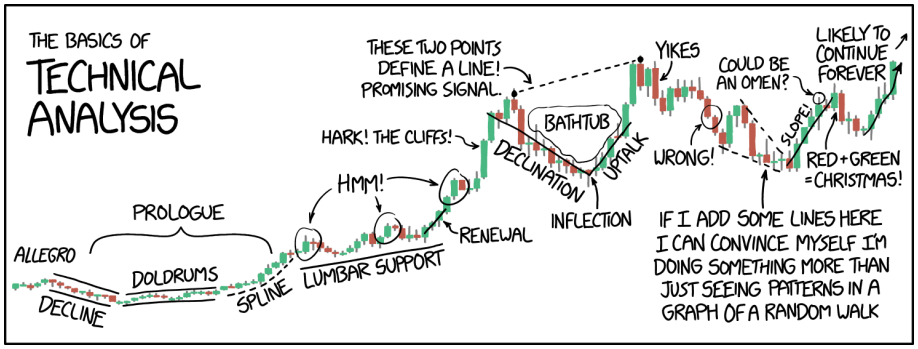


FIGURE 3.5: Cinq trajectoires de la marche aléatoire simple symétrique sur $\mathbb{Z}$ (seuls les 200 premiers pas sont affichés).



https://xkcd.com/2101

4 Fonctions génératrices

A generating function is a clothesline on which we hang up a sequence of numbers for display.

Herbert Wilf

4.1 Définition, propriétés

Soit $a = (a_i)_{i=0}^{\infty}$ une suite de nombres réels. On appelle **fonction génératrice** de la suite a la fonction définie par

$$G_a(s) \coloneqq \sum_{i=0}^{\infty} a_i s^i$$

pour les $s \in \mathbb{C}$ tels que la série converge. Rappelons quelques propriétés de base de ce type de fonctions.

Convergence. Il existe un **rayon de convergence** $R \in [0, +\infty]$ tel que la série converge absolument si $|s| < R$ et diverge si $|s| > R$. La série est uniformément convergente sur les ensembles de la forme $\{s \,|\, |s| \leqslant R'\}$, quel que soit $R' < R$.

Différentiation. $G_a(s)$ peut être dérivée ou intégrée terme à terme un nombre arbitraire de fois, tant que $|s| < R$.

Unicité. S'il existe $0 < R' \leqslant R$ tel que $G_a(s) = G_b(s)$ pour tout s tel que $|s| < R'$, alors $a_n = b_n$ pour tout n. De plus,

$$a_n = \frac{1}{n!} G_a^{(n)}(0).$$

Continuité. (Théorème d'Abel) Si $a_i \geqslant 0$ pour tout i et si $G_a(s)$ est finie pour $|s| < 1$, alors

$$G_a(1^-) \equiv \lim_{\epsilon \downarrow 0} G_a(1-\epsilon) = \sum_{i=0}^{\infty} a_i,$$

que la somme de la série soit finie ou égale à $+\infty$. (Ce résultat est évidemment utile lorsque le rayon de convergence R est égal à 1.)

Étant donnée une variable aléatoire X à valeurs dans $\mathbb{N}$, la fonction de masse de X donne lieu à la suite $(\mathbb{P}(X = k))_{k=0}^{\infty}$; on va s'intéresser à la fonction génératrice qui lui est associée.

Définition 4.1. *Soit X une variable aléatoire à valeurs dans $\mathbb{N}$. On appelle* ***fonction génératrice*** *de X la fonction $G_X : \mathbb{C} \to \mathbb{C}$ donnée par la série entière*

$$G_X(s) := \mathbb{E}[s^X] = \sum_{k=0}^{\infty} \mathbb{P}(X = k) s^k.$$

Remarque 4.2. *Comme $G_X(1) = \mathbb{E}[1] = 1$, le rayon de convergence de G_X est toujours supérieur ou égal à 1.*

Remarque 4.3. *Il suit de l'unicité des coefficients des fonctions génératrices que la loi d'une variable aléatoire est entièrement caractérisée par sa fonction génératrice : $\mathbb{P}(X = k) = \frac{1}{k!} G_X^{(k)}(0)$ pour tout $k \in \mathbb{N}$.*

Exemple 4.4. Calculons la fonction génératrice associée à quelques-unes des lois discrètes introduites au Chapitre 2.

1. *Loi de Bernoulli.* Si $X \sim \mathsf{bernoulli}(p)$,
$$G_X(s) = (1 - p) + ps.$$
2. *Loi binomiale.* Si $X \sim \mathsf{binom}(n, p)$, la formule du binôme implique que
$$G_X(s) = \sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} s^k = ((1 - p) + ps)^n.$$
3. *Loi de Poisson.* Si $X \sim \mathsf{poisson}(\lambda)$,
$$G_X(s) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} s^k = e^{\lambda(s-1)}.$$
4. *Loi géométrique.* Si $X \sim \mathsf{géom}(p)$,
$$\sum_{k=1}^{\infty} p(1 - p)^{k-1} s^k = \frac{ps}{1 - (1 - p)s}.$$ ⌟

Le théorème d'Abel fournit une technique efficace pour calculer les moments de X En effet, si $|s| < 1$,

$$\begin{aligned}
G'_X(s) &= \sum_{k=0}^{\infty} k s^{k-1} \mathbb{P}(X = k) && \implies && G'_X(1^-) = \mathbb{E}[X], \\
G''_X(s) &= \sum_{k=0}^{\infty} k(k-1) s^{k-2} \mathbb{P}(X = k) && \implies && G''_X(1^-) = \mathbb{E}[X(X-1)], \\
G_X^{(\ell)}(s) &= \sum_{k=0}^{\infty} k \cdots (k - \ell + 1) s^{k-\ell} \mathbb{P}(X = k) && \implies && G_X^{(\ell)}(1^-) = \mathbb{E}[X \cdots (X - \ell + 1)].
\end{aligned}$$

On a donc en particulier le résultat suivant.

Lemme 4.5. *Soit X une variable aléatoire à valeurs dans $\mathbb{N}$. Alors,*

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

Exemple 4.6. *Espérance et variance de la loi hypergéométrique.* La formule du binôme montre que la fonction génératrice d'une variable $X \sim \mathsf{hypergeom}(N, K, n)$,

$$G_X(s) = \sum_{k=(n+K-N)\vee 0}^{K \wedge n} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} s^k,$$

est précisément le coefficient de x^n du polynôme

$$Q(x,s) = \frac{1}{\binom{N}{n}}(1+sx)^K(1+x)^{N-K}.$$

Par conséquent, l'espérance de X coïncide avec le coefficient de x^n de

$$\frac{\partial Q}{\partial s}(x,1) = \frac{K}{\binom{N}{n}}x(1+x)^{N-1}$$

et est donc donnée par $G'_X(1^-) = Kn/N$. De façon similaire, on trouve que la variance de X est égale à $nK(N-K)(N-n)/(N^3-N^2)$. ⌟

Remarque 4.7. *En général, si l'on désire calculer les moments d'une variable aléatoire X à valeurs dans $\mathbb{N}$, il est préférable de travailler avec la* ***fonction génératrice des moments*** *de X, qui est définie par*

$$M_X(t) \coloneqq G_X(e^t),$$

pourvu que $e^t < R$, le rayon de convergence de G_X. En effet, on a alors, si $R > 1$ et $t < \log R$,

$$M_X(t) = \sum_{k=0}^{\infty} e^{tk}\,\mathbb{P}(X=k) = \sum_{k=0}^{\infty}\sum_{n=0}^{\infty}\frac{(tk)^n}{n!}\,\mathbb{P}(X=k) = \sum_{n=0}^{\infty}\frac{t^n}{n!}\Big(\sum_{k=0}^{\infty}k^n\,\mathbb{P}(X=k)\Big) = \sum_{n=0}^{\infty}\frac{t^n}{n!}\mathbb{E}[X^n].$$

Les moments de X peuvent donc être aisément obtenus en dérivant $M_X(t)$:

$$\mathbb{E}[X^n] = M_X^{(n)}(0).$$

Les fonctions génératrices se révèlent particulièrement utiles dans l'analyse de sommes de variables aléatoires.

Lemme 4.8. *Soit $X_1,\dots,X_n$ des variables aléatoires indépendantes à valeurs dans $\mathbb{N}$. Alors la fonction génératrice de $S_n \coloneqq X_1+\cdots+X_n$ est donnée par*

$$G_{S_n}(s) = G_{X_1}(s)\cdots G_{X_n}(s).$$

Démonstration. Il suit du Lemme 2.76 que

$$G_{S_n}(s) = \mathbb{E}[s^{X_1+\cdots+X_n}] = \mathbb{E}[s^{X_1}\cdots s^{X_n}] = \mathbb{E}[s^{X_1}]\cdots\mathbb{E}[s^{X_n}] = G_{X_1}(s)\cdots G_{X_n}(s). \qquad \square$$

Exemple 4.9. Soit $X \sim \mathsf{binom}(n,p)$ et $Y \sim \mathsf{binom}(m,p)$ deux variables aléatoires indépendantes. Alors

$$G_{X+Y}(s) = G_X(s)G_Y(s) = ((1-p)+ps)^n\,((1-p)+ps)^m = ((1-p)+ps)^{n+m},$$

ce qui montre que $X+Y \sim \mathsf{binom}(n+m,p)$, par la Remarque 4.3.

De façon similaire, si $X \sim \mathsf{poisson}(\lambda)$ et $Y \sim \mathsf{poisson}(\mu)$ sont deux variables aléatoires indépendantes, alors $X+Y \sim \mathsf{poisson}(\lambda+\mu)$:

$$G_{X+Y}(s) = e^{\lambda(s-1)}e^{\mu(s-1)} = e^{(\lambda+\mu)(s-1)}.$$ ⌟

En fait, on peut même aller plus loin, et considérer la somme d'un nombre *aléatoire* de variables aléatoires. Ceci a de nombreuses applications.

Lemme 4.10. *Soit $(X_k)_{k\geqslant 1}$ une suite de variables aléatoires i.i.d. à valeurs dans $\mathbb{N}$, G_X leur fonction génératrice commune et N une variable aléatoire à valeurs dans $\mathbb{N}$, indépendante de $(X_k)_{k\geqslant 1}$ et dont la fonction génératrice est G_N. Alors la fonction génératrice de $S \coloneqq X_1+\cdots+X_N$ est donnée par*

$$G_S = G_N \circ G_X.$$

(Comme d'habitude, $X_1+\cdots+X_N \coloneqq 0$ si $N=0$.)

Démonstration.

$$\begin{aligned} G_S(s) = \mathbb{E}[s^S] = \mathbb{E}[s^{X_1+\cdots+X_N}\mathbf{1}_{\{N\geqslant 1\}} + \mathbf{1}_{\{N=0\}}] &= \sum_{n\in\mathbb{N}^*} \mathbb{E}[s^{X_1+\cdots+X_n}\mathbf{1}_{\{N=n\}}] + \mathbb{P}(N=0) \\ &= \sum_{n\in\mathbb{N}^*} \mathbb{E}[s^{X_1+\cdots+X_n}]\,\mathbb{P}(N=n) + \mathbb{P}(N=0) \\ &= \sum_{n\in\mathbb{N}^*} \big(G_X(s)\big)^n\,\mathbb{P}(N=n) + \mathbb{P}(N=0) \\ &= \sum_{n\in\mathbb{N}} \big(G_X(s)\big)^n\,\mathbb{P}(N=n) \\ &= G_N\big(G_X(s)\big), \end{aligned}$$

où l'égalité de la seconde ligne suit de l'indépendance de N et des $(X_k)_{k\in\mathbb{N}^*}$. □

Exemple 4.11. On désire modéliser le montant total des achats effectués dans un magasin pendant une période donnée. On suppose que le nombre de clients pendant cette période est donné par une variable aléatoire N, et que les montants dépensés par les clients forment une collection $X_1, X_2, \ldots$ de variables aléatoires i.i.d. de moyenne μ, indépendantes de N. Le montant total des achats est donc donné par la variable aléatoire $S = X_1 + \cdots + X_N$, somme d'un nombre aléatoire de termes. Il suit donc des Lemmes 4.5 et 4.10 que

$$\mathbb{E}[S] = (G_N \circ G_X)'(1^-) = G_N'(G_X(1^-))\,G_X'(1^-) = G_N'(1^-)\,G_X'(1^-) = \mu\mathbb{E}[N].$$

La formule précédente est connue sous le nom de **formule de Wald**[1]. ⌟

Exemple 4.12. Une poule pond N œufs, où N suit une loi de Poisson de paramètre λ. Chaque œuf éclot avec probabilité p indépendamment des autres. Soit K le nombre de poussins. On a $K = X_1 + \cdots + X_N$, où les X_i sont des variables aléatoires de Bernoulli de paramètre p indépendantes. Quelle est la loi de K ? Comme on l'a vu,

$$G_N(s) = \exp\big(\lambda(s-1)\big), \qquad G_X(s) = (1-p) + ps.$$

Par conséquent, le Lemme 4.10 implique que

$$G_K(s) = G_N\big(G_X(s)\big) = \exp\big(\lambda\, p\,(s-1)\big).$$

On en conclut que $K \sim \mathsf{poisson}(\lambda p)$. ⌟

Remarque 4.13. *Dans cette section, on a toujours supposé que les variables aléatoires prenaient valeurs dans $\mathbb{N}$. Il est parfois aussi utile de considérer le cas de variables aléatoires prenant valeurs dans $\overline{\mathbb{N}}$. Pour une telle variable aléatoire X, on voit que $G_X(s) = \mathbb{E}[s^X]$ converge tant que $|s| < 1$ et que*

$$G_X(1^-) = \sum_{k=0}^{\infty} \mathbb{P}(X=k) = 1 - \mathbb{P}(X=\infty).$$

4.2 Application aux processus de branchement

Dans cette section, nous allons illustrer la puissance des fonctions génératrices dans l'étude d'une classe intéressante de processus stochastiques : les **processus de branchement**.

1. Abraham Wald (1902, Cluj-Napoca – 1950, Travancore), mathématicien et statisticien américain d'origine hongroise. On lui doit un des exemples les plus célèbres du biais du survivant. Pendant la seconde guerre mondiale, une de ses tâches consistait à déterminer comment minimiser la perte de bombardiers sous le feu ennemi. Sa recommandation a été de blinder les endroits des appareils qui présentaient *le moins de dommages*. En effet, les endroits endommagés sur les avions étant parvenus à revenir correspondent à des endroits où ceux-ci peuvent apparemment subir des dommages sans que cela n'entrave leur capacité à retourner à la base.

À l'époque victorienne, certaines personnes ont craint la disparition des noms des familles aristocratiques. Sir Francis Galton[2] posa originellement la question de déterminer la probabilité d'un tel événement dans le *Educational Times* de 1873, et le Révérend Henry William Watson[3] répondit avec une solution. Ensemble, ils écrivirent alors, en 1874, un article intitulé « On the probability of extinction of families ». Leur modèle suppose (cela étant considéré comme allant de soi à l'époque de Galton, et étant encore le cas le plus courant dans la plupart des pays) que le nom de famille est transmis à tous les enfants mâles par leur père. Il suppose également que le nombre de fils d'un individu est une variable aléatoire à valeurs dans $\mathbb{N}$, et que le nombre de fils d'hommes différents sont des variables aléatoires indépendantes de même loi.

Plus généralement, supposons qu'une population évolue par générations, et notons Z_n le nombre d'individus de la génération n. Chaque membre de la génération n donne naissance à une famille, éventuellement vide, de la génération suivante ; la taille de la famille est une variable aléatoire. On fait les hypothèses suivantes :

- ▷ les tailles de chaque famille forment une collection de variable aléatoires indépendantes ;
- ▷ les tailles des familles suivent toutes la même loi.

Sous ces hypothèses, le processus est bien défini dès que la taille de la population initiale Z_0 est donnée ; on supposera ici que $Z_0 = 1$; le cas général s'en déduit facilement et est laissé en exercice. Ce modèle peut également représenter la croissance d'une population de cellules, celle de neutrons dans un réacteur, la propagation d'une maladie dans une population, etc.

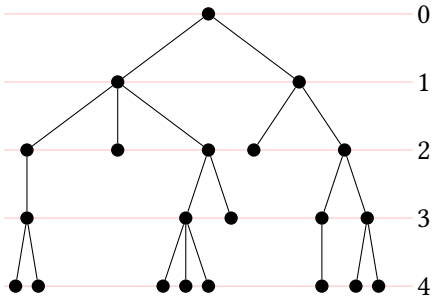


FIGURE 4.1: Les 5 premières générations d'une réalisation d'un processus de branchement de Galton–Watson.

On s'intéresse à la suite aléatoire $Z_0, Z_1, Z_2, \dots$ des tailles des générations successives. Ce processus peut être défini de la façon suivante : on considère une collection[4] $(X_k^n)_{n\geqslant 0, k\geqslant 1}$ de variables aléatoires i.i.d. à valeurs dans $\mathbb{N}$; X_k^n représente le nombre de fils du k-ième individu de la génération n (si celui-ci existe). On note G la fonction génératrice commune à ces variables aléatoires (encodant donc la loi du nombre de fils d'un individu). On peut alors poser $Z_0 \coloneqq 1$ et

$$\forall n \geqslant 1, \qquad Z_n \coloneqq X_1^{n-1} + X_2^{n-1} + \cdots + X_{Z_{n-1}}^{n-1}, \tag{4.1}$$

le nombre d'individus de la génération n étant égal au nombre total de fils des individus de la génération $n-1$. On notera $G_n(s) \coloneqq \mathbb{E}[s^{Z_n}]$ la fonction génératrice de Z_n. Observez qu'en particulier $G_1 = G$, puisque $Z_0 = 1$.

Lemme 4.14. *Pour tout $n \geqslant 1$,*

$$G_n = G^{\circ n} \equiv \underbrace{G \circ G \circ \cdots \circ G}_{n \text{ fois}}.$$

2. Sir Francis Galton (1822, Sparkbrook – 1911, Haslemere), homme de science britannique. L'un des fondateurs de la psychologie différentielle ou comparée. On lui doit le terme anticyclone, ainsi que l'invention du sac de couchage. À partir de 1865, il se consacre à la statistique avec l'objectif de quantifier les caractéristiques physiques, psychiques et comportementales de l'homme, ainsi que leur évolution.

3. Henry William Watson (1827,Londres – 1903, Berkswell), mathématicien britanique.

4. L'ensemble $\mathbb{N} \times \mathbb{N}^*$ étant dénombrable, l'existence d'une telle famille de variables aléatoires est garantie par la discussion de la Section 2.9.

Démonstration. Soit $n \geqslant 1$. Il suit immédiatement de (4.1) et du Lemme 4.10 que

$$G_n = G_{n-1} \circ G,$$

puisque Z_n est la somme de Z_{n-1} variables aléatoires i.i.d. de fonction génératrice G. L'affirmation se démontre alors en itérant l'identité précédente :

$$G_n = G_{n-1} \circ G = (G_{n-2} \circ G) \circ G = \cdots = \underbrace{G \circ G \circ \cdots \circ G}_{n \text{ fois}}. \qquad \square$$

Les moments de la variable aléatoire Z_n peuvent facilement s'exprimer en termes des moments de la variable aléatoire Z_1 décrivant la taille d'une famille typique.

Lemme 4.15. *Soit $\mu := \mathbb{E}[Z_1]$ et $\sigma^2 := \mathrm{Var}[Z_1]$. Alors*

$$\mathbb{E}[Z_n] = \mu^n,$$
$$\mathrm{Var}[Z_n] = \begin{cases} n\sigma^2 & \text{si } \mu = 1 \\ \sigma^2(\mu^n - 1)\mu^{n-1}(\mu - 1)^{-1} & \text{si } \mu \neq 1. \end{cases}$$

Démonstration. Par le Lemme 4.14, $G_n = G^{\circ n} = G \circ G_{n-1} = G_{n-1} \circ G$. Par conséquent, il suit du Lemme 4.5 que

$$\mathbb{E}[Z_n] = G_n'(1^-) = G'(G_{n-1}(1^-))\, G_{n-1}'(1^-) = G'(1^-)\, G_{n-1}'(1^-) = \mu\, \mathbb{E}[Z_{n-1}],$$

ce qui donne bien, après itération, $\mathbb{E}[Z_n] = \mu^n$.

En dérivant deux fois $G \circ G_{n-1}$, on obtient

$$G_n''(1^-) = G''(1^-)\,(G_{n-1}'(1^-))^2 + G'(1^-)\, G_{n-1}''(1^-) = (\sigma^2 - \mu + \mu^2)\mu^{2n-2} + \mu G_{n-1}''(1^-).$$

Lorsque $\mu = 1$, on a donc $Z_n = \sigma^2 + Z_{n-1}$ et l'affirmation suit immédiatement.

Pour traiter le cas $\mu \neq 1$, on calcule $G_n''(1^-)$ d'une autre façon, en dérivant deux fois $G_{n-1} \circ G$:

$$G_n''(1^-) = G_{n-1}''(1^-)\,(G'(1^-))^2 + G_{n-1}'(1^-)\, G''(1^-) = \mu^2 G_{n-1}''(1^-) + \mu^{n-1}(\sigma^2 - \mu + \mu^2).$$

En éliminant $G_{n-1}''(1^-)$ des deux expressions obtenues, on en déduit que

$$G_n''(1^-) = \sigma^2(\mu^n - 1)\mu^{n-1}(\mu - 1)^{-1} - \mu^n + \mu^{2n}.$$

Il suit donc que

$$\mathrm{Var}[Z_n] = \sigma^2(\mu^n - 1)\mu^{n-1}(\mu - 1)^{-1},$$

ce qui démontre la seconde affirmation. $\square$

Une question particulièrement intéressante concerne le destin de la population : va-t-elle s'éteindre après un temps fini ou, au contraire, toutes les générations auront-elles une taille strictement positive ? L'événement correspondant à l'extinction de la population peut s'écrire

$$\{\lim_{n\to\infty} Z_n = 0\} = \{\exists n_0, \forall n \geqslant n_0,\ Z_n = 0\} = \bigcup_{n \geqslant 1} \{Z_n = 0\}.$$

La suite d'événements $(\{Z_n = 0\})_{n \geqslant 1}$ étant croissante, on a

$$\mathbb{P}(\lim_{n\to\infty} Z_n = 0) = \lim_{n\to\infty} \mathbb{P}(Z_n = 0).$$

La question peut donc être reformulée de la façon suivante : la limite $\lim_{n\to\infty} \mathbb{P}(Z_n = 0)$ est-elle égale à 1 (extinction inéluctable) ou strictement inférieure à 1 (survie possible) ? (Observez que s'il est possible qu'un individu n'ait pas de descendance, alors la probabilité d'extinction est toujours strictement positive.)

Le théorème suivant montre que le destin de la population est étroitement lié à la taille moyenne des familles.

Théorème 4.16. *Soit $\mu = \mathbb{E}[Z_1]$, la taille moyenne d'une famille. La probabilité d'extinction*

$$\eta \coloneqq \lim_{n\to\infty} \mathbb{P}(Z_n = 0)$$

est donnée par la plus petite racine positive de l'équation $s = G(s)$. En particulier, $\eta = 1$ si $\mu < 1$ et $\eta < 1$ si $\mu > 1$. Lorsque $\mu = 1$, on a $\eta = 1$ dès que la loi de Z_1 possède une variance positive.

Démonstration. Notons $\eta_n \coloneqq \mathbb{P}(Z_n = 0) = G_n(0)$. L'existence de la limite $\eta \coloneqq \lim_{n\to\infty} \eta_n$ est immédiate puisque $(\eta_n)_{n\geqslant 1}$ est une suite croissante, majorée par 1. On a

$$\forall n \geqslant 1, \qquad \eta_n = G_n(0) = G(G_{n-1}(0)) = G(\eta_{n-1}).$$

Par continuité de G sur $[0,1]$, on a donc bien

$$\eta = \lim_{n\to\infty} \eta_n = \lim_{n\to\infty} G(\eta_{n-1}) = G(\lim_{n\to\infty} \eta_{n-1}) = G(\eta).$$

Vérifions à présent que si $a \leqslant 1$ est une racine positive de cette équation, alors $\eta \leqslant a$ (on peut se restreindre à $a \leqslant 1$, puisque 1 est toujours solution de cette équation). G étant croissante sur $[0,1]$,

$$\eta_1 = G(0) \leqslant G(a) = a.$$

De façon similaire,

$$\eta_2 = G(\eta_1) \leqslant G(a) = a.$$

Il suit, par induction, que $\eta_n \leqslant a$ pour tout n. Par conséquent, $\eta \leqslant a$. On en conclut que η est bien la plus petite racine positive de l'équation $s = G(s)$.

Pour démontrer la seconde affirmation, on observe tout d'abord que

$$G''(s) = \mathbb{E}\big[Z_1(Z_1-1)s^{Z_1-2}\big] = \sum_{k\geqslant 2} k(k-1)s^{k-2}\mathbb{P}(Z_1 = k) \geqslant 0,$$

pour tout $s \in [0,1]$. G est donc convexe (en fait, strictement convexe si $\mathbb{P}(Z_1 \geqslant 2) > 0$) et croissante sur $[0,1]$, avec $G(1) = 1$.

L'équation $s = G(s)$ possède toujours au moins une solution en $s = 1$ et au plus une seconde (par convexité), sauf dans le cas trivial où $\mathbb{P}(Z_1 = 1) = 1$, pour lequel $G(s) = s$ pour tout s et $\eta = 0$.

En excluant ce dernier cas, la question est donc de déterminer si une seconde solution inférieure à 1 existe (lorsqu'elle existe, elle est forcément positive, car $G(0) \geqslant 0$). Un coup d'œil à la Figure 4.2 (et un argument analytique élémentaire laissé en exercice) montre que cela dépend de la valeur de $\mu = G'(1)$. Lorsque $G'(1) = \mu < 1$, la plus petite solution est $\eta = 1$. Lorsque $G'(1) = \mu > 1$, la plus petite solution doit être inférieure à 1, puisque $G(0) \geqslant 0$; on a donc, dans ce cas, $\eta < 1$. Finalement, dans le cas $\mu = 1$, les deux courbes sont tangentes en 1, ce qui implique que $\eta = 1$ dès que Z_1 possède une variance positive (puisque dans ce cas G est strictement convexe sur $[0,1]$). □

4.3 Application à la marche aléatoire simple sur $\mathbb{Z}$

Nous allons à présent donner une illustration de l'utilisation des fonctions génératrices à l'étude de la marche aléatoire simple sur $\mathbb{Z}$. On a déjà vu, au Chapitre 3, que la marche symétrique retournait presque sûrement à son point de départ. Nous allons à présent considérer la même question pour une marche de paramètre p arbitraire.

On note $g_n \coloneqq \mathbb{P}_0(S_n = 0)$ et $h_n \coloneqq \mathbb{P}_0(\tau_0 = n)$. Les fonctions génératrices correspondantes sont notées

$$\mathbb{G}(s) \coloneqq \sum_{n=0}^{\infty} g_n s^n, \qquad \mathbb{H}(s) \coloneqq \sum_{n=1}^{\infty} h_n s^n.$$

Observons que ces deux fonctions génératrices ont un rayon de convergence au moins égal à 1 ($g_n, h_n \in [0,1]$). Notons également que $\mathbb{P}_0(\tau_0 = +\infty)$ peut être strictement positive, auquel cas $\mathbb{H}(1^-) = \mathbb{P}_0(\tau_0 < \infty) < 1$.

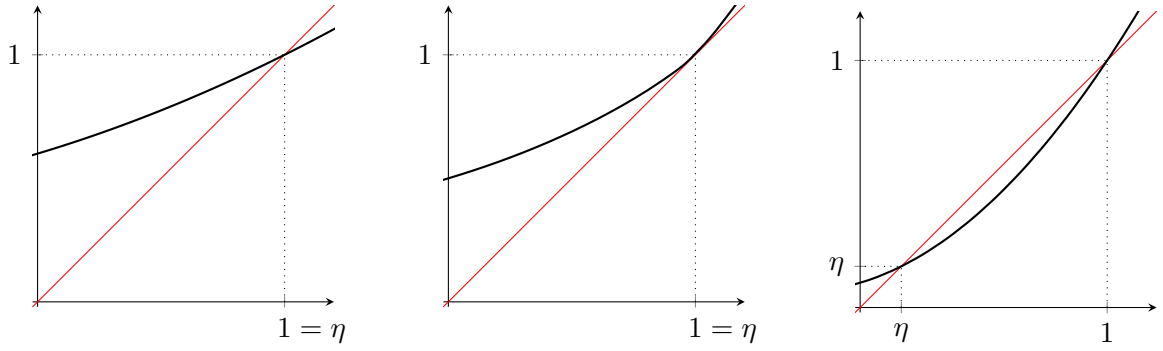


FIGURE 4.2: Solutions de l'équation $G(s) = s$. Gauche : $\mu < 1$. Milieu : $\mu = 1$ et $\mathrm{Var}[Z_1] > 0$. Droite : $\mu > 1$.

Lemme 4.17. *Les assertions suivantes sont vraies pour tout $s \in \mathbb{C}$ tel que $|s| < 1$.*

1. $\mathbb{G}(s) = 1 + \mathbb{G}(s)\mathbb{H}(s)$.
2. $\mathbb{G}(s) = (1 - 4pqs^2)^{-1/2}$.
3. $\mathbb{H}(s) = 1 - (1 - 4pqs^2)^{1/2}$.

Démonstration. *1.* Comme on l'a déjà vu (à la fin de la preuve du Théorème 3.12), on a, pour $n \geqslant 1$,

$$\begin{aligned} g_{2n} = \mathbb{P}_0(S_{2n} = 0) &= \sum_{k=1}^{n} \mathbb{P}_0(\tau_0 = 2k)\mathbb{P}_0(S_{2n} = 0 \,|\, \tau_0 = 2k) \\ &= \sum_{k=1}^{n} \mathbb{P}_0(\tau_0 = 2k)\mathbb{P}_0(S_{2n-2k} = 0) = \sum_{k=1}^{n} h_{2k} g_{2n-2k}. \end{aligned}$$

Par conséquent

$$\mathbb{G}(s) = \sum_{n=0}^{\infty} g_{2n} s^{2n} = 1 + \sum_{n=1}^{\infty} g_{2n} s^{2n} = 1 + \sum_{n=1}^{\infty} \sum_{k=1}^{n} h_{2k} g_{2n-2k} s^{2n}.$$

La conclusion suit donc, puisque

$$\sum_{n=1}^{\infty} \sum_{k=1}^{n} h_{2k} g_{2n-2k} s^{2n} = \sum_{k=1}^{\infty} \sum_{n=k}^{\infty} h_{2k} g_{2n-2k} s^{2n} = \sum_{k=1}^{\infty} h_{2k} s^{2k} \sum_{n=k}^{\infty} g_{2n-2k} s^{2n-2k} = \mathbb{H}(s)\mathbb{G}(s).$$

2. On doit calculer la fonction génératrice associée à la suite

$$g_n = \begin{cases} \binom{n}{n/2} (pq)^{n/2}, & n \text{ pair}, \\ 0 & n \text{ impair}, \end{cases}$$

c'est-à-dire $\mathbb{G}(s) = \sum_{n \geqslant 0} \binom{2n}{n} (pqs^2)^n$. Pour ce faire, on vérifie tout d'abord que

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} = 2^n \frac{(2n-1)!!}{n!} = (-4)^n \frac{(-\frac{1}{2})(-\frac{3}{2}) \cdots (-\frac{2n-1}{2})}{n!} = (-4)^n \binom{-\frac{1}{2}}{n},$$

où l'on a employé les notations standards

$$(2n-1)!! := (2n-1)(2n-3)(2n-5) \cdots 3 = \frac{(2n)!}{(2n)(2n-2)(2n-4) \cdots 2} = \frac{(2n)!}{2^n n!}$$

et, pour $a \in \mathbb{R}$ et $n \in \mathbb{N}$,

$$\binom{a}{n} := \frac{a(a-1)(a-2) \cdots (a-n+1)}{n!}.$$

Par la formule du binôme généralisée (Lemme A.18), pour tout $a \in \mathbb{R}$ et tout x tel que $|x| < 1$,

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

Par conséquent, pour $|4pqs^2| < 1$ (c'est-à-dire $|s| < 1$, puisque $pq \leqslant \frac{1}{4}$),

$$\mathbb{G}(s) = \sum_{n \geqslant 0} \binom{-\frac{1}{2}}{n} (-4pqs^2)^n = (1-4pqs^2)^{-1/2}.$$

3. suit immédiatement de *1* et *2*. □

Théorème 4.18. *La probabilité que la marche retourne au moins une fois à l'origine est égale à*

$$\mathbb{P}_0(\tau_0 < \infty) = \sum_{n=1}^{\infty} h_n = \mathbb{H}(1) = 1 - |p-q|.$$

Dans le cas où cela est certain, c'est-à-dire lorsque $p = q = \frac{1}{2}$, l'espérance du temps de premier retour est infinie,

$$\mathbb{E}_0(\tau_0) = \sum_{n=1}^{\infty} n h_n = \mathbb{H}'(1) = \infty.$$

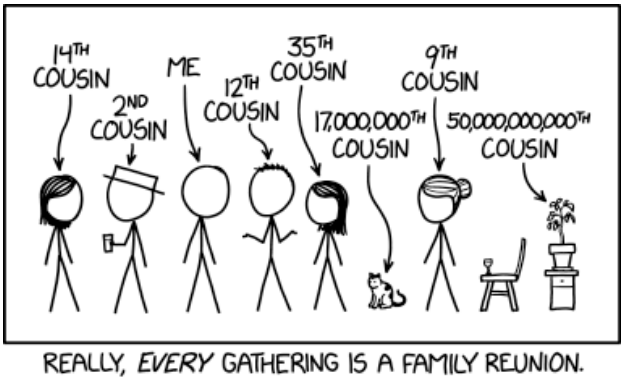
Démonstration. La première affirmation suit après avoir pris la limite $s \uparrow 1$ dans l'expression pour $\mathbb{H}(s)$ donnée dans le Lemme 4.17 (observez que $1 - 4pq = (p-q)^2$).

Dans le cas symétrique $p = \frac{1}{2}$, la fonction génératrice du temps de premier retour devient simplement $\mathbb{H}(s) = 1 - (1-s^2)^{1/2}$. Par conséquent,

$$\mathbb{E}_0(\tau_0) = \lim_{s \uparrow 1} \mathbb{H}'(s) = \infty.$$ □

Définition 4.19. *La marche aléatoire est dite* ***récurrente*** *si le retour à son point de départ est presque certain ; sinon elle est dite* ***transiente****. On dit qu'elle est* ***récurrente-nulle*** *si elle est récurrente et que l'espérance du temps de retour est infinie, et* ***récurrente-positive*** *si cette espérance est finie.*

Le théorème précédent montre que la marche aléatoire simple sur $\mathbb{Z}$ est récurrente-nulle si $p = \frac{1}{2}$ et transiente dans les autres cas.



https://xkcd.com/2608

5 Fonctions caractéristiques

Fourier's theorem is not only one of the most beautiful results of modern analysis, but it may be said to furnish an indispensable instrument in the treatment of nearly every recondite question in modern physics.

Lord Kelvin

Dans ce chapitre, nous allons introduire la notion de fonction caractéristique associée à une variable aléatoire. Celle-ci fournit un outil similaire aux fonctions génératrices, mais applicable à des variables aléatoires réelles arbitraires, ce qui le rend extrêmement puissant.

5.1 Définition et propriétés élémentaires

Définition 5.1. *La **fonction caractéristique** associée à une variable aléatoire réelle X est la fonction $\phi_X : \mathbb{R} \to \mathbb{C}$ définie par*

$$\phi_X(t) \coloneqq \mathbb{E}[e^{\mathrm{i}tX}] = \mathbb{E}[\cos(tX)] + \mathrm{i}\mathbb{E}[\sin(tX)].$$

Exemple 5.2. Si X est une variable aléatoire réelle à valeurs dans $\mathbb{N}$,

$$\phi_X(t) = \mathbb{E}[e^{\mathrm{i}tX}] = \mathbb{E}[(e^{\mathrm{i}t})^X] = G_X(e^{\mathrm{i}t}). \qquad \lrcorner$$

Exemple 5.3. Si X est une variable aléatoire réelle avec densité f_X,

$$\phi_X(t) = \int_{\mathbb{R}} e^{\mathrm{i}tx} f_X(x)\,\mathrm{d}x.$$

La fonction caractéristique coïncide donc, dans ce cas, avec la transformée de Fourier de la densité de X. En fait, la fonction caractéristique ϕ_X, où X est une variable aléatoire quelconque, correspond à la transformée de Fourier de la mesure de probabilité $\mathbb{P}_X$. $\lrcorner$

Exemple 5.4. Soit $X \sim \mathcal{N}(0,1)$. Dans ce cas,

$$\phi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2 + \mathrm{i}tx}\,\mathrm{d}x.$$

En complétant le carré, $x^2 - 2\mathrm{i}tx = (x - \mathrm{i}t)^2 + t^2$, et en déplaçant le chemin d'intégration de la droite réelle à la droite $\{\mathrm{Im}(z) = t\}$ (exercice : justifiez cela), on obtient

$$\phi_X(t) = e^{-\frac{1}{2}t^2}. \qquad \lrcorner$$

Exemple 5.5. Soit $X \sim$ cauchy. Dans ce cas,

$$\phi_X(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{e^{\mathrm{i}tx}}{1+x^2}\,\mathrm{d}x.$$

Pour calculer l'intégrale, on peut utiliser la méthode des résidus. Si $t > 0$, on vérifie facilement (exercice) que

$$\lim_{R\to\infty} \int_{C_R} \frac{e^{\mathrm{i}tx}}{1+x^2}\,\mathrm{d}x = 0,$$

où C_R est le demi-cercle de diamètre $[-R, R]$ dans le demi-plan supérieur. Par conséquent,

$$\phi_X(t) = \frac{1}{\pi}\, 2\mathrm{i}\pi\, \frac{e^{-t}}{2\mathrm{i}} = e^{-t},$$

puisque le résidu en i est égal à $\lim_{x\to\mathrm{i}}(x-\mathrm{i})e^{\mathrm{i}tx}/(1+x^2) = e^{-t}/2\mathrm{i}$. En procédant de façon similaire lorsque $t < 0$ (en considérant cette fois le demi-cercle dans le demi-plan inférieur), on obtient finalement que

$$\phi_X(t) = e^{-|t|},$$

pour tout $t \in \mathbb{R}$. ⌟

Théorème 5.6. *La fonction caractéristique ϕ_X associée à une variable aléatoire réelle X possède les propriétés suivantes :*

1. $\phi_X(0) = 1$ *et* $|\phi_X(t)| \leqslant 1$ *pour tout* $\in \mathbb{R}$.
2. ϕ_X *est uniformément continue sur* $\mathbb{R}$.
3. ϕ_X *est définie positive, c'est-à-dire*
$$\sum_{j,k} \phi(t_j - t_k) z_j \bar{z}_k \geqslant 0,$$
pour tout $t_1, \ldots, t_n \in \mathbb{R}$ *et tout* $z_1, \ldots, z_n \in \mathbb{C}$.

Démonstration. Soit ϕ une fonction caractéristique. Alors $\phi(0) = \mathbb{E}[1] = 1$ et $|\phi(t)| \leqslant \mathbb{E}[|e^{\mathrm{i}tX}|] = 1$.

De plus,

$$|\phi(t+s) - \phi(t)| = |\mathbb{E}[e^{\mathrm{i}(t+s)X} - e^{\mathrm{i}tX}]| \leqslant \mathbb{E}[|e^{\mathrm{i}tX}(e^{\mathrm{i}sX} - 1)|] = \mathbb{E}[|e^{\mathrm{i}sX} - 1|].$$

Soit $Y(s) := |e^{\mathrm{i}sX} - 1|$. Comme $0 \leqslant Y \leqslant 2$ et $\lim_{s\to 0} Y(s) = 0$, le théorème de convergence dominée implique que $\lim_{s\to 0} \mathbb{E}[Y(s)] = 0$, ce qui établit la continuité uniforme.

Finalement, observons que

$$\sum_{j,k} \phi(t_j - t_k) z_j \bar{z}_k = \mathbb{E}\Big[\sum_{j,k} z_j e^{\mathrm{i}t_j X}\, \bar{z}_k e^{-\mathrm{i}t_k X}\Big] = \mathbb{E}\Big[\Big|\sum_j z_j e^{\mathrm{i}t_j X}\Big|^2\Big] \geqslant 0. \qquad \square$$

Remarque 5.7. *On peut également montrer la réciproque : toute fonction $\phi : \mathbb{R} \to \mathbb{C}$ possédant ces propriétés est la fonction caractéristique d'une variable aléatoire réelle. C'est ce que l'on appelle le Théorème de Bochner.*

Le résultat ci-dessous montre que la fonction caractéristique permet de calculer les moments de la variable aléatoire associée, mais il trouvera des usages bien plus importants au Chapitre 6.

Lemme 5.8. *Soit $n \in \mathbb{N}$. Si $\mathbb{E}[|X|^n] < \infty$, alors $\phi_X \in \mathcal{C}^n$ et*

$$\forall 0 \leqslant k \leqslant n, \qquad \phi_X^{(k)}(t) = \mathrm{i}^k \mathbb{E}[X^k e^{\mathrm{i}tX}].$$

Démonstration. On procède par récurrence. Le cas $k = 0$ suit immédiatement de la définition de ϕ_X. Supposons donc le résultat vrai pour $k < n$ et vérifions qu'il reste vrai pour $k + 1$. On a, par l'hypothèse de récurrence,

$$\frac{\phi_X^{(k)}(t+h) - \phi_X^{(k)}(t)}{h} = \mathrm{i}^k \mathbb{E}\Big[X^k \frac{e^{\mathrm{i}(t+h)X} - e^{\mathrm{i}tX}}{h}\Big] = \mathrm{i}^k \mathbb{E}\Big[X^{k+1} e^{\mathrm{i}tX} \frac{e^{\mathrm{i}hX} - 1}{hX}\Big].$$

D'une part, $\lim_{h\to 0} \frac{e^{\mathrm{i}hX}-1}{hX} = \mathrm{i}$. D'autre part, comme [1] $|e^{\mathrm{i}x} - 1| \leqslant |x|$ pour tout $x \in \mathbb{R}$,

$$|X^{k+1} e^{\mathrm{i}tX} \frac{e^{\mathrm{i}hX} - 1}{hX}| \leqslant |X^{k+1}|.$$

Par hypothèse, $\mathbb{E}[|X^{k+1}|] < \infty$. Il suit donc du théorème de convergence dominée que

$$\lim_{h\to 0} \mathrm{i}^k \mathbb{E}[X^{k+1} e^{\mathrm{i}tX} \frac{e^{\mathrm{i}hX} - 1}{hX}] = \mathrm{i}^{k+1} \mathbb{E}[X^{k+1} e^{\mathrm{i}tX}],$$

ce qui montre que ϕ_X est $k+1$ fois dérivable et que $\phi_X^{(k+1)}(t) = \mathrm{i}^{k+1}\mathbb{E}[X^{k+1}e^{\mathrm{i}tX}]$. La continuité de $\phi_X^{(k+1)}$ se démontre exactement comme la continuité uniforme dans le Théorème 5.6. On en conclut que $\phi_X \in \mathcal{C}^{k+1}$. □

Il suit immédiatement du Lemme 5.8 que si X possède un moment d'ordre k, alors $\mathbb{E}[X^k] = (-\mathrm{i})^k \phi_X^{(k)}(0)$.

Remarque 5.9. *Attention : l'existence de $\phi'_X(0)$ n'implique pas que $\mathbb{E}[X] = \phi'_X(0)$. On peut en effet construire des variables aléatoires sans espérance, mais telles que $\phi'_X(0)$ existe (cela sera fait en exercice).*

Le résultat élémentaire suivant est également souvent utile.

Lemme 5.10. *Si $a, b \in \mathbb{R}$ et $Y = aX + b$, alors*

$$\phi_Y(t) = e^{\mathrm{i}tb}\phi_X(at).$$

Démonstration.

$$\phi_Y(t) = \mathbb{E}[e^{\mathrm{i}t(aX+b)}] = e^{\mathrm{i}tb}\mathbb{E}[e^{\mathrm{i}(at)X}] = e^{\mathrm{i}tb}\phi_X(at).$$ □

Exemple 5.11. On vérifie facilement (exercice) que si $Z \sim \mathcal{N}(\mu, \sigma^2)$, alors $X := (Z - \mu)/\sigma \sim \mathcal{N}(0, 1)$. Il suit donc du Lemme 5.10 et de l'Exemple 5.4 que sa fonction caractéristique est donnée par

$$\phi_Z(t) = \phi_{\sigma X+\mu}(t) = e^{-\frac{1}{2}\sigma^2 t^2 + \mathrm{i}\mu t}.$$ ⌟

Un des nombreux intérêts des fonctions caractéristiques est qu'elles fournissent un outil très efficace pour étudier les sommes de variables aléatoires indépendantes.

Lemme 5.12. *Soit $X_1, \ldots, X_n$ des variables aléatoires réelles indépendantes. Alors*

$$\phi_{X_1+\cdots+X_n}(t) = \phi_{X_1}(t) \cdots \phi_{X_n}(t).$$

Démonstration.

$$\phi_{X_1+\cdots+X_n}(t) = \mathbb{E}[e^{\mathrm{i}tX_1} \cdots e^{\mathrm{i}tX_n}] = \mathbb{E}[e^{\mathrm{i}tX_1}] \cdots \mathbb{E}[e^{\mathrm{i}tX_n}] = \phi_{X_1}(t) \cdots \phi_{X_n}(t).$$

(La seconde identité suit de l'indépendance, après avoir décomposé chacune des exponentielles en sinus et cosinus, effectué la multiplication et regroupé les termes.) □

1. Par exemple : $|e^{\mathrm{i}x} - 1|^2 = 2(1 - \cos x) \leqslant x^2$.

5.2 Théorèmes d'inversion et de continuité

Nous allons à présent montrer qu'une variable aléatoire réelle est complètement caractérisée par sa fonction caractéristique : deux variables aléatoires réelles possédant la même fonction caractéristique ont la même loi. La preuve repose sur le théorème suivant.

Théorème 5.13 (Théorème d'inversion). *Soit X une variable aléatoire réelle de fonction de répartition F_X et de fonction caractéristique ϕ_X. Alors,*

$$F_X(b) - F_X(a) = \lim_{N\to\infty} \int_{-N}^{N} \frac{e^{-\mathrm{i}at} - e^{-\mathrm{i}bt}}{2\mathrm{i}\pi t} \phi_X(t)\,\mathrm{d}t.$$

pour chaque paire de points a, b auxquels F_X est continue.

Dans le cas d'une variable aléatoire à densité dont la densité est continue et bornée, cette dernière peut être récupérée à partir de la fonction caractéristique via le théorème d'inversion de Fourier. Le résultat ci-dessus fournit une généralisation à la transformée de Fourier d'une mesure de probabilité quelconque.

Démonstration.

$$\begin{aligned}
\lim_{N\to\infty} \frac{1}{2\pi} \int_{-N}^{N} \mathrm{d}t\, \frac{e^{-\mathrm{i}at} - e^{-\mathrm{i}bt}}{\mathrm{i}t} \int e^{\mathrm{i}tx}\,\mathrm{d}\mathbb{P}_X &= \lim_{N\to\infty} \frac{1}{2\pi} \int \mathrm{d}\mathbb{P}_X \int_{-N}^{N} \frac{e^{\mathrm{i}t(x-a)} - e^{\mathrm{i}t(x-b)}}{\mathrm{i}t}\,\mathrm{d}t \\
&= \lim_{N\to\infty} \frac{1}{2\pi} \int \mathrm{d}\mathbb{P}_X \int_{-N}^{N} \frac{\sin\left(t(x-a)\right) - \sin\left(t(x-b)\right)}{t}\,\mathrm{d}t \\
&= \tfrac{1}{2} \int \left\{\mathrm{signe}(x-a) - \mathrm{signe}(x-b)\right\} \mathrm{d}\mathbb{P}_X \\
&= F_X(b) - F_X(a),
\end{aligned}$$

pourvu que a et b soient des points de continuité de F. On a utilisé le théorème de Fubini et le théorème de convergence dominée pour prendre la limite $N \to \infty$. En effet, la fonction de Dirichlet

$$u(N, z) := \int_0^N \frac{\sin tz}{t}\,\mathrm{d}t$$

satisfait $\sup_{N,z} |u(N,z)| \leqslant C$ et[2]

$$\lim_{N\to\infty} u(N,z) = \begin{cases} \pi/2 & \text{si } z > 0, \\ -\pi/2 & \text{si } z < 0, \\ 0 & \text{si } z = 0. \end{cases}$$

□

Corollaire 5.14. *Deux variables aléatoires X et Y ont la même fonction caractéristique si et seulement si elles ont la même loi.*

Démonstration. Si $\phi_X = \phi_Y$, alors le Théorème d'inversion implique que

$$F_X(b) - F_X(a) = F_Y(b) - F_Y(a),$$

en toute paire de points de continuité a et b de F_X et F_Y. En laissant $a \to -\infty$ (se rappeler que l'ensemble des points de discontinuité d'une fonction croissante est au plus dénombrable), on obtient

$$F_X(b) = F_Y(b),$$

en tout point de continuité de F_X et F_Y. On en déduit que $F_X = F_Y$, par continuité à droite des fonctions de répartition. □

2. Poser, pour $n \geqslant 1$, $u_n = \int_0^{\pi/2} \sin\big((2n-1)x\big)/\sin(x)\,\mathrm{d}x$ et $v_n = \int_0^{\pi/2} \sin(2nx)/x\,\mathrm{d}x$. Montrer que : 1. $u_{n+1} = u_n, \forall n \geqslant 1$ (observez que $\sin\big((2n+1)x\big) - \sin\big((2n-1)x\big) = 2\cos(2nx)\sin(x)$); 2. $u_1 = \pi/2$; 3. $\lim_{n\to\infty}(u_n - v_n) = 0$ (intégration par parties en observant que $1/x - 1/\sin(x)$ est continûment différentiable sur $[0, \pi/2]$); 4. $\lim_{N\to\infty} u(N,1) = \lim_{n\to\infty} v_n = \pi/2$.

Exemple 5.15. Soit $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ deux variables aléatoires indépendantes. Alors, il suit du Lemme 5.12 et de l'Exemple 5.11 que

$$\phi_{X_1+X_2}(t) = \phi_{X_1}(t)\phi_{X_2}(t) = e^{-\frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2+\mathrm{i}(\mu_1+\mu_2)t}.$$

Par conséquent, il suit du Corollaire 5.14 que $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. ⌟

Exemple 5.16. Soit $X_1, \dots, X_n$ des variables aléatoires indépendantes suivant la loi de Cauchy. Alors, il suit du Lemme 5.12 et de l'Exemple 5.5 que

$$\phi_{(X_1+\cdots+X_n)/n}(t) = \phi_{X_1}(t/n)\cdots\phi_{X_n}(t/n) = e^{-|t|}.$$

Par conséquent, il suit du Corollaire 5.14 que $\frac{1}{n}(X_1 + \cdots + X_n) \sim \mathsf{cauchy}$. ⌟

Les fonctions caractéristiques sont aussi très utiles pour étudier la convergence de variables aléatoires (nous discuterons les différents modes de convergence au Chapitre 6).

Définition 5.17. *On dit qu'une suite de fonction de répartition F_n converge vers une fonction de répartition F si $F(x) = \lim_{n\to\infty} F_n(x)$, en chaque point x où F est continue. On notera cette convergence $F_n \to F$.*

Remarque 5.18. *La raison pour laquelle on exclut les points de discontinuité de F et que leur inclusion empêcherait la convergence même dans des cas très simples. Soit $(X_n)_{n\geqslant 1}$ des variables aléatoires telles que, pour tout $n \geqslant 1$, $\mathbb{P}(X_n = 1/n) = 1$. Évidemment, toute notion raisonnable de convergence des variables aléatoires $(X_n)_{n\geqslant 1}$ devrait garantir que la limite est la variable aléatoire X telle que $\mathbb{P}(X = 0) = 1$. Or, $F_{X_n}(x) = \mathbf{1}_{\{x\geqslant \frac{1}{n}\}}$ et $F_X(x) = \mathbf{1}_{\{x\geqslant 0\}}$. On voit donc $\lim_{n\to\infty} F_{X_n}(0) = 0 \neq 1 = F_X(0)$.*

Théorème 5.19 (Théorème de continuité de Lévy[3])**.** *Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires réelles. Si $\phi(t) = \lim_{n\to\infty} \phi_{X_n}(t)$ existe et est continue en $t = 0$, alors ϕ est la fonction caractéristique associée à une variable aléatoire X et $F_{X_n} \to F_X$.*

Démonstration. Nous ne la ferons pas ici, la preuve étant de nature purement analytique. □

5.3 Fonction caractéristique conjointe

On peut également définir une notion de fonction caractéristique conjointe pour une famille de variables aléatoires réelles.

Définition 5.20. *La **fonction caractéristique conjointe** associée au vecteur aléatoire $\mathbf{X} = (X_1, \dots, X_n)$ est définie par*

$$\phi_{\mathbf{X}}(\mathbf{t}) \coloneqq \mathbb{E}[e^{\mathbf{i}\mathbf{t}\cdot\mathbf{X}}],$$

pour tout $\mathbf{t} = (t_1, \dots, t_n) \in \mathbb{R}^n$.

Remarque 5.21. *Il est utile d'observer que $\phi_{(X_1,\dots,X_n)}(t_1, \dots, t_n) = \phi_{t_1X_1+\cdots+t_nX_n}(1)$.*

Des résultats analogues à ceux des Théorèmes 5.13 et 5.19 et du Corollaire 5.14 sont également vrais pour les fonctions caractéristiques conjointes. Nous ne les énoncerons pas explicitement.

La fonction caractéristique conjointe fournit une nouvelle caractérisation de l'indépendance.

Théorème 5.22. *Les variables aléatoires réelles $X_1, \dots, X_n$ sont indépendantes si et seulement si*

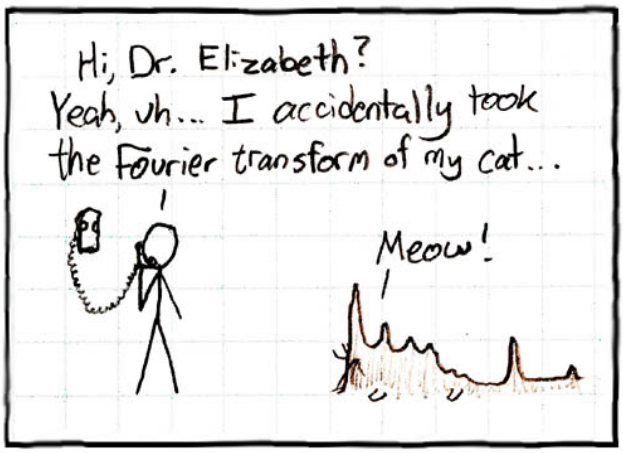
$$\phi_{(X_1,\dots,X_n)}(t_1, \dots, t_n) = \prod_{j=1}^{n} \phi_{X_j}(t_j).$$

3. Paul Pierre Lévy (1886, Paris – 1971, Paris), mathématicien français.

Démonstration. Si les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes, alors le résultat suit des Lemmes 5.12 et 5.10 :

$$\phi_{(X_1,\ldots,X_n)}(t_1,\ldots,t_n) = \phi_{t_1X_1+\cdots+t_nX_n}(1) = \prod_{j=1}^{n} \phi_{t_jX_j}(1) = \prod_{j=1}^{n} \phi_{X_j}(t_j).$$

La réciproque suit (de la version à n variables) du théorème d'inversion. □



https://xkcd.com/26

6 Théorèmes limites

I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." [...] The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.

Francis Galton

Le but de ce chapitre est d'étudier un certain nombre de résultats classiques de théorie des probabilités, dont les lois des grands nombres (faible et forte), le théorème central limite et la loi zéro–un de Kolmogorov.

Les théorèmes limites sont omniprésents en théorie des probabilités. Une raison de leur importance réside dans leur aptitude à transformer des événements de probabilité $p \in [0,1]$ en des événements de probabilité proche de 0 ou 1. En effet, la théorie des probabilités n'est pas uniquement une théorie mathématique, mais relève également de la science expérimentale : elle fait de nombreuses assertions que l'on peut tester empiriquement. Le problème, c'est qu'une assertion du type « ce dé équilibré va donner un 1 au prochain lancer avec probabilité $1/6$ » est infalsifiable (si le dé affiche un 5 après le lancer, qu'est-ce que cela dit sur l'assertion ? Rien du tout.). Les seules assertions probabilistes que l'on peut tester empiriquement sont celles qui concernent les événements de probabilité très proche de 1 ou 0. Pour revenir à l'exemple du dé, une assertion du type « si on lance le dé $6\,000$ fois, le nombre de fois où l'on a aura obtenu un 1 est compris entre 880 et $1\,120$ » est, elle, falsifiable : si vous lancez le dé $6\,000$ fois et n'obtenez pas un nombre de 1 compris dans l'intervalle annoncé, la conclusion rationnelle est que l'une des hypothèses est fausse (par exemple, le dé n'est peut-être pas équilibré), puisque la probabilité de commettre une erreur est approximativement $3 \cdot 10^{-5}$, soit bien plus petite que ce qui est toléré dans la grande majorité des sciences expérimentales.

6.1 Quelques outils

6.1.1 Les lemmes de Borel–Cantelli

Soit $(A_k)_{k\geqslant 1}$ une suite d'événements sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. L'événement « une infinité des A_k sont réalisés » peut s'écrire

$$\{\omega \in \Omega \,|\, \forall n \geqslant 1, \exists m \geqslant n, \omega \in A_m\} = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m =: \limsup_{n\to\infty} A_n.$$

Remarque 6.1. *La notation* $\limsup_{n\to\infty} A_n$ *est motivée par l'observation que*

$$\limsup_{n\to\infty} \mathbf{1}_{A_n} = \lim_{n\to\infty} \sup_{k\geqslant n} \mathbf{1}_{A_k} = \lim_{n\to\infty} \mathbf{1}_{\bigcup_{k\geqslant n} A_k} = \mathbf{1}_{\bigcap_{n\geqslant 1}\bigcup_{k\geqslant n} A_k} = \mathbf{1}_{\limsup_{n\to\infty} A_n}.$$

Il est souvent nécessaire de déterminer si un événement de ce type est réalisé.

Lemme 6.2 (Lemmes de Borel–Cantelli[1])**.** *Soit* $(A_k)_{k\geqslant 1}$ *une suite d'événements sur un espace de probabilité* $(\Omega, \mathcal{F}, \mathbb{P})$ *et* $A := \limsup_{n\to\infty} A_n$*. Alors,*

1. $\mathbb{P}(A) = 0$ *si* $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$*.*
2. $\mathbb{P}(A) = 1$ *si* $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$*, lorsque les événements* $(A_k)_{k\geqslant 1}$ *sont indépendants.*

Démonstration. *1.* Puisque la suite $(\bigcup_{m=n}^{\infty} A_m)_{n\geqslant 1}$ est décroissante, il suit du Lemme 1.4 que

$$\mathbb{P}(A) = \mathbb{P}\Big(\lim_{n\to\infty} \bigcup_{m=n}^{\infty} A_m\Big) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcup_{m=n}^{\infty} A_m\Big) \leqslant \lim_{n\to\infty} \sum_{m=n}^{\infty} \mathbb{P}(A_m) = 0.$$

2. Il est plus simple de considérer l'événement complémentaire

$$A^{\mathrm{c}} = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m^{\mathrm{c}}.$$

Observons à présent que la suite $(\bigcap_{m=n}^{N} A_m^{\mathrm{c}})_{N\geqslant n}$ est décroissante, ce qui implique que

$$\begin{aligned}
\mathbb{P}\Big(\bigcap_{m=n}^{\infty} A_m^{\mathrm{c}}\Big) &= \lim_{N\to\infty} \mathbb{P}\Big(\bigcap_{m=n}^{N} A_m^{\mathrm{c}}\Big) && \text{(Lemme 1.4)} \\
&= \lim_{N\to\infty} \prod_{m=n}^{N} (1 - \mathbb{P}(A_m)) && \text{(indépendance)} \\
&\leqslant \lim_{N\to\infty} \prod_{m=n}^{N} \exp\big(-\mathbb{P}(A_m)\big) && (1 - x \leqslant e^{-x}) \\
&= \lim_{N\to\infty} \exp\Big(-\sum_{m=n}^{N} \mathbb{P}(A_m)\Big) = 0,
\end{aligned}$$

dès que $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$. Comme la suite $(\bigcap_{m=n}^{\infty} A_m^{\mathrm{c}})_{n\geqslant 1}$ est croissante, il suit du Lemme 1.4 que

$$\mathbb{P}(A^{\mathrm{c}}) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcap_{m=n}^{\infty} A_m^{\mathrm{c}}\Big) = 0. \qquad \square$$

Remarque 6.3. *Sans l'hypothèse d'indépendance, la seconde partie peut être fausse : il suffit de considérer un événement* B *tel que* $\mathbb{P}(B) \in (0,1)$ *et la suite d'événements donnée par* $A_k = B$ *pour tout* $k \geqslant 1$*; on a alors* $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$*, mais* $\mathbb{P}(A) = \mathbb{P}(B) < 1$*. On peut toutefois remplacer cette condition par l'indépendance deux à deux (voire même des conditions encore plus faibles), mais la preuve devient alors moins simple.*

1. Francesco Paolo Cantelli (1875, Palerme – 1966, Rome), mathématicien italien.

6.1.2 Quelques inégalités

En dépit de sa simplicité, l'inégalité suivante (et ses conséquences) est un outil essentiel du probabiliste. On en a déjà vu une variante dans le Lemme (2.65).

Lemme 6.4 (Inégalité de Markov). *Soit X une variable aléatoire réelle à valeurs positives. Alors,*

$$\forall a > 0, \qquad \mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a}.$$

Démonstration.

$$\mathbb{E}[X] \geqslant \mathbb{E}[X\mathbf{1}_{\{X\geqslant a\}}] \geqslant a\mathbb{E}[\mathbf{1}_{\{X\geqslant a\}}] = a\mathbb{P}(X \geqslant a). \qquad \square$$

Examinons à présent quelques-unes des principales conséquences de l'inégalité de Markov.

Exemple 6.5. Soit $X \in \mathscr{L}^2$. Alors, pour tout $a > 0$,

$$\mathbb{P}(|X - \mathbb{E}[X]| \geqslant a) = \mathbb{P}((X - \mathbb{E}[X])^2 \geqslant a^2) \leqslant \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{a^2} = \frac{\mathrm{Var}[X]}{a^2},$$

où l'inégalité suit d'une application du Lemme 6.4. On retrouve l'inégalité de Bienaymé–Tchebychev, déjà rencontrée dans le Lemme 2.65. ⌟

Exemple 6.6. Soit X une variable aléatoire réelle. La fonction $H_X : \mathbb{R} \to [0, +\infty]$ définie par

$$H_X(t) := \begin{cases} \log \mathbb{E}[e^{tX}] & \text{si } \mathbb{E}[e^{tX}] < \infty, \\ \infty & \text{sinon.} \end{cases}$$

est appelée la **fonction génératrice des cumulants de X**. L'inégalité

$$\forall a \in \mathbb{R}, \qquad \mathbb{P}(X \geqslant a) \leqslant \exp\big(-\sup_{t>0}(ta - H_X(t))\big) \tag{6.1}$$

est appelée **inégalité de Chernoff**[2]. Pour la démontrer, observez qu'une application du Lemme 6.4 donne

$$\forall t > 0, \qquad \mathbb{P}(X \geqslant a) = \mathbb{P}(e^{tX} \geqslant e^{ta}) \leqslant \frac{\mathbb{E}[e^{tX}]}{e^{ta}} = e^{-(ta - H_X(t))}.$$

(6.1) suit en prenant l'infimum du membre de droite sur les $t > 0$. La fonction $a \mapsto \sup_{t>0}(ta - H_X(t))$ est la **transformée de Legendre**[3]**–Fenchel**[4] de la fonction H_X. ⌟

6.2 Modes de convergence

Le but de ce chapitre est d'étudier le comportement asymptotique de certaines variables aléatoires. Pour ce faire, nous allons avoir besoin de notions de convergence appropriées pour une suite de variables aléatoires.

2. Herman Chernoff (1923, New York –), mathématicien, physicien et statisticien américain.

3. Adrien-Marie Legendre (1752, Paris – 1833, Paris), mathématicien français. Stendhal écrivit de lui, dans son roman autobiographique *Vie de Henry Brulard* : « Le célèbre Legendre, géomètre de premier ordre, recevant la croix de la Légion d'honneur, l'attacha à son habit, se regarda au miroir et sauta de joie. L'appartement était bas, sa tête heurta le plafond, il tomba à moitié assommé. Digne mort c'eût été pour ce successeur d'Archimède ! ».

4. Moritz Werner Fenchel (1905, Berlin – 1988, Copenhague), mathématicien danois d'origine allemande.

Définition 6.7. *Soit $(X_k)_{k\geqslant 1}$ et X des variables aléatoires réelles sur $(\Omega, \mathcal{F}, \mathbb{P})$. On dit que*

1. *X_n converge vers X **presque sûrement**, noté $X_n \xrightarrow{\text{p.s.}} X$, si*

$$\mathbb{P}\big(\{\omega \in \Omega \,|\, \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\big) = 1.$$

2. *X_n converge vers X **en moyenne d'ordre r** ($r \geqslant 1$), noté $X_n \xrightarrow{r} X$, si $X_n \in \mathscr{L}^r$ pour tout n, et*

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

3. *X_n converge vers X **en probabilité**, noté $X_n \xrightarrow{\mathbb{P}} X$, si*

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

4. *X_n converge vers X **en loi**, noté $X_n \xrightarrow{\text{loi}} X$, si*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x),$$

en chaque point x auquel la fonction de répartition F_X est continue.

Remarque 6.8. *Lorsque $X_n \xrightarrow{1} X$, on parle de **convergence en moyenne**. Lorsque $X_n \xrightarrow{2} X$, on parle de **convergence en moyenne quadratique**.*

La plupart de ces modes de convergence devraient vous être familiers du cours d'analyse fonctionnelle, quoique sous d'autres noms : convergence presque partout au lieu de presque sûre, convergence dans L^p au lieu de convergence en moyenne d'ordre p, convergence en mesure au lieu de convergence en probabilité.

Observez que la convergence en loi se distingue nettement des autres modes de convergence en ce qu'elle ne dépend que de la loi de X (pas de l'espace $(\Omega, \mathcal{F}, \mathbb{P})$ sous-jacent). En particulier, on n'a évidemment pas unicité de la limite dans ce cas : si $X \overset{\text{loi}}{=} Y$ sont deux variable aléatoire réelles, potentiellement définies sur des espaces de probabilité différents, et si $X_n \xrightarrow{\text{loi}} X$, alors $X_n \xrightarrow{\text{loi}} Y$, puisque $F_X = F_Y$.

Remarque 6.9. *Soit $(X_n)_{n\geqslant 1}$ telles que $X_n \xrightarrow{\text{loi}} X$. Soit $I \subset \mathbb{R}$ un intervalle dont les extrémités sont des points de continuité de F_X. On vérifie facilement (exercice) que*

$$\lim_{n\to\infty} \mathbb{P}(X_n \in I) = \mathbb{P}(X \in I).$$

Plus généralement, on peut montrer que, pour tout $B \in \mathcal{B}(\mathbb{R})$ tel que $\mathbb{P}_X(\partial B) = 0$,

$$\lim_{n\to\infty} \mathbb{P}(X_n \in B) = \mathbb{P}(X \in B).$$

*On peut également montrer que $X_n \xrightarrow{\text{loi}} X$ si et seulement si $\mathbb{P}_{X_n}$ **converge faiblement** vers $\mathbb{P}_X$:*

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

pour toute fonction $\varphi : \mathbb{R} \to \mathbb{R}$ continue et bornée.

Les relations entre ces divers modes de convergence sont résumées dans le théorème suivant.

Théorème 6.10. *Les implications suivantes sont vérifiées :*

$$\begin{array}{ccc} (X_n \xrightarrow{\text{p.s.}} X) & & \\ \Downarrow & & \\ (X_n \xrightarrow{\mathbb{P}} X) & \Rightarrow & (X_n \xrightarrow{\text{loi}} X) \\ \Uparrow & & \\ (X_n \xrightarrow{s} X) & & \\ \Uparrow & & \\ (X_n \xrightarrow{r} X) & & \end{array}$$

pour tout $r > s \geqslant 1$. Aucune autre implication n'est vraie en général.

Démonstration. Sera faite en exercices. □

Certaines implications dans l'autre sens deviennent possibles si l'on ajoute des conditions supplémentaires. Le lemme suivant contient quelques résultats de ce type qui se révèlent particulièrement utiles.

Lemme 6.11. *1. Si $X_n \xrightarrow{\text{loi}} c$, avec c une constante, alors $X_n \xrightarrow{\mathbb{P}} c$.*
2. Si $\sum_n \mathbb{P}(|X_n - X| > \epsilon) < \infty$, pour tout $\epsilon > 0$, alors $X_n \xrightarrow{\text{p.s.}} X$.

Démonstration. *1.* Puisque $F_c(x) = \mathbf{1}_{\{x \geqslant c\}}$, on a

$$\mathbb{P}(|X_n - c| > \epsilon) = \mathbb{P}(X_n < c - \epsilon) + \mathbb{P}(X_n > c + \epsilon) \leqslant F_{X_n}(c - \epsilon) + (1 - F_{X_n}(c + \epsilon))$$

et le membre de droite tend vers 0 lorsque $X_n \xrightarrow{\text{loi}} c$.
2. Soit $A_M := \{\omega \in \Omega \,|\, \exists N \geqslant 1, \forall n \geqslant N, |X_n(\omega) - X(\omega)| \leqslant 1/M\} = \bigcup_{N \geqslant 1} \bigcap_{n \geqslant N} \{|X_n - X| \leqslant 1/M\}$. Par le Lemme 1.4,

$$\begin{aligned} \mathbb{P}\big(\{\omega \in \Omega \,|\, X_n(\omega) \to X(\omega)\}\big) &= \mathbb{P}\Big(\big\{\omega \in \Omega \,\big|\, \forall M \geqslant 1, \exists N \geqslant 1, \forall n \geqslant N, |X_n(\omega) - X(\omega)| \leqslant 1/M\big\}\Big) \\ &= \mathbb{P}\Big(\bigcap_{M \geqslant 1} \bigcup_{N \geqslant 1} \bigcap_{n \geqslant N} \big\{|X_n - X| \leqslant \frac{1}{M}\big\}\Big) = \lim_{M \to \infty} \mathbb{P}(A_M). \end{aligned}$$

La conclusion suit donc de l'hypothèse et du premier lemme de Borel–Cantelli : pour tout $M \geqslant 1$,

$$\mathbb{P}(A_M^{\text{c}}) = \mathbb{P}\Big(\bigcap_{N \geqslant 1} \bigcup_{n \geqslant N} \big\{|X_n - X| > \tfrac{1}{M}\big\}\Big) = \mathbb{P}(\limsup_{n \to \infty} \{|X_n - X| > \tfrac{1}{M}\}) = 0.$$ □

6.3 La loi des grands nombres

6.3.1 Loi faible des grands nombres

Nous avons vu une version de la loi faible des grands nombres, pour des variables aléatoires dans $\mathscr{L}^2$ non corrélées, dans le Théorème 2.83. Nous allons à présent en montrer une autre version, valables pour des variables aléatoires indépendantes, mais n'exigeant que l'existence du premier moment.

Théorème 6.12 (Loi faible des grands nombres)**.** *Soit $(X_k)_{k \geqslant 1}$ des variables aléatoires réelles i.i.d. d'espérance μ. Alors, $\overline{X}_n := (X_1 + \cdots + X_n)/n \xrightarrow{\mathbb{P}} \mu$.*

Démonstration. Le point 1 du Lemme 6.11 implique qu'il suffit de démontrer que $\overline{X}_n \xrightarrow{\text{loi}} \mu$. Pour ce faire, on observe tout d'abord que $\phi_X \in \mathcal{C}^1$ par le Lemme 5.8. Le théorème de Taylor–Young donne donc

$$\phi_X(t) = 1 + \mathrm{i}t\mu + \mathsf{o}(t).$$

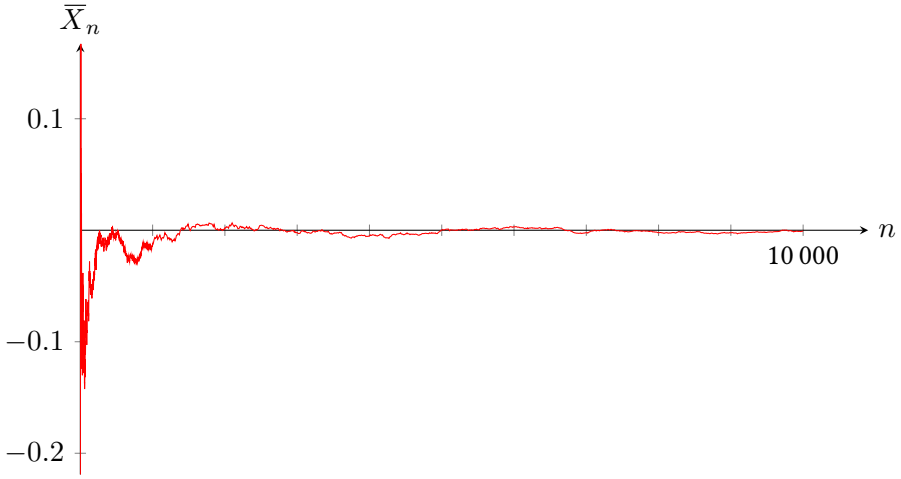


FIGURE 6.1: La moyenne empirique d'une famille de variables aléatoires de loi $\mathsf{U}(-1,1)$ (n allant de 1 à 10 000).

Les Lemmes 5.12 et 5.10 impliquent alors que la fonction caractéristique de la variable aléatoire $\overline{X}_n$ satisfait

$$\phi_{\overline{X}_n}(t) = \big(\phi_X(t/n)\big)^n = \big(1 + \frac{\mathrm{i}t\mu}{n} + \mathsf{o}(\frac{t}{n})\big)^n \xrightarrow{n\to\infty} e^{\mathrm{i}t\mu}.$$

$e^{\mathrm{i}t\mu}$ étant la fonction caractéristique de la variable aléatoire constante μ, la convergence en loi suit du théorème de continuité (Théorème 5.19). □

Remarque 6.13. *L'existence du premier moment est nécessaire : une suite de variables aléatoires indépendantes dont l'espérance n'existe pas ne satisfait pas la loi des grands nombres. Un exemple simple est donné par une suite de variables aléatoires i.i.d. suivant la loi de Cauchy. En effet, dans ce cas, il suit de l'Exemple 5.16 que $\phi_{\overline{X}_n}(t) = e^{-|t|}$, ce qui montre que $\overline{X}_n \sim$ cauchy et ne peut donc pas converger en loi vers une constante ! La Figure 6.2 montre le comportement d'une réalisation de $\overline{X}_n$ pour n allant de 1 à 10 000 ; comparez ce comportement à celui de la Figure 6.1 qui montre le comportement correspondant pour des variables aléatoires uniformes sur $[-1,1]$!*

Cet exemple montre également que la condition $\mathbb{E}[|X|] < \infty$ pour qu'une variable aléatoire possède une espérance n'est pas purement technique : on pourrait être tenté de poser $\mathbb{E}[X] = 0$ lorsque $X \sim$ cauchy (après tout, sa fonction de densité est paire), mais l'espérance ne serait alors plus la limite des moyennes empiriques de variables aléatoires i.i.d. suivant cette loi !

Pour être utile en pratique (en particulier, pour déterminer quelle doit être la taille minimale d'un échantillon si l'on désire obtenir un degré de certitude donné pour une précision donnée), il est important d'obtenir des estimations précises de la vitesse de convergence.

Exemple 6.14. Pour illustrer ce point, reprenons la discussion entamée dans l'Exemple 2.84, portant sur 10 000 jets d'une pièce équilibrée. Afin de travailler avec des variables centrées, on encode le résultat du k-ième jet par une variable X_k telle que $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = \frac{1}{2}$ (au lieu de 0 et 1).

On applique l'inégalité de Chernoff. Il suffit de déterminer la fonction génératrice des cumulants H associée à $\overline{X}_n$:

$$e^{H(t)} = \mathbb{E}[e^{t\overline{X}_n}] = \mathbb{E}[\prod_{k=1}^{n} e^{tX_k/n}] = \prod_{k=1}^{n} \mathbb{E}[e^{tX_k/n}] = \cosh(t/n)^n.$$

On a donc

$$\mathbb{P}(\overline{X}_n \geqslant x) \leqslant \inf_{t>0} e^{n\log\cosh(t/n) - tx}.$$

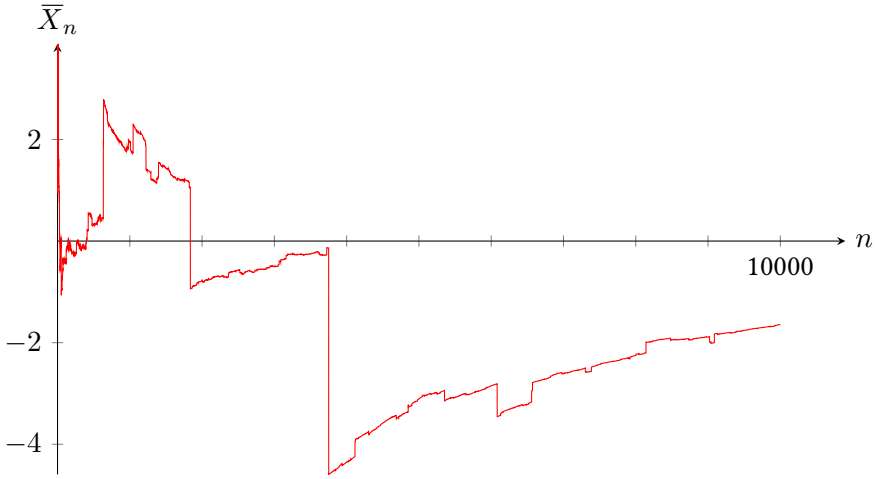


FIGURE 6.2: La moyenne empirique d'une famille de variables aléatoires suivant une loi de Cauchy (n allant de 1 à 10 000).

Un calcul[5] montre que la fonction $f(t) := \log\cosh(t/n) - x\,t/n$ atteint son unique minimum en $t^* := \frac{n}{2}\log[(1+x)/(1-x)]$. En introduisant

$$I(x) := -f(t^*) = \frac{1}{2}\big((1+x)\log(1+x) + (1-x)\log(1-x)\big),$$

et en utilisant la symétrie pour estimer $\mathbb{P}(\overline{X}_n \leqslant -x)$, on a finalement

$$\mathbb{P}\big(|\overline{X}_n| \geqslant x\big) \leqslant 2\,e^{-nI(x)}. \tag{6.2}$$

En posant $n = 10\,000$ et $x = 0{,}1$, on trouve $I(0{,}1) \simeq 0{,}005$, et par conséquent

$$\mathbb{P}(\overline{X}_{10\,000} \notin (-0{,}1, 0{,}1)) \leqslant 3{,}5 \cdot 10^{-22}.$$

Comparez ce résultat avec l'estimée de l'Exemple 2.84 !

Un résultat du type (6.2) est ce qu'on appelle une estimée de **grande déviation**. La théorie des grandes déviations est un domaine important de la théorie des probabilités, pour le développement duquel l'un de ses principaux artisans, S.R.S. Varadhan[6], a été récompensé par le prix Abel en 2007. ⌟

6.3.2 La loi forte des grands nombres

La loi faible des grands nombres nous fournit des informations sur le comportement de $\overline{X}_n$ (pour n grand) lorsqu'on considère de nombreuses répétitions de l'expérience : pour tout n *fixé* suffisamment grand, $\overline{X}_n$ est proche de μ pour la plupart des réalisations. Elle n'affirme cependant pas que, pour une réalisation ω donnée, la fonction $n \mapsto \overline{X}_n(\omega)$ reste forcément proche de μ lorsque n augmente : elle laisse ouverte la possibilité qu'il existe $\epsilon > 0$ et une sous-suite $(n_k)_{k\geqslant 1}$, $n_k \to \infty$, telle que $|\overline{X}_{n_k}(\omega) - \mu| > \epsilon$, pour tout $k \geqslant 1$. La **loi forte des grands nombres** montre que l'ensemble des réalisations ω pour lesquelles ceci se produit a probabilité nulle : pour tout $\epsilon > 0$, avec probabilité 1, seul un nombre fini des événements $\big\{|\overline{X}_n - \mu| > \epsilon\big\}$ sont réalisés.

Théorème 6.15 (Loi forte des grands nombres). *Soit $(X_k)_{k\geqslant 1}$ une suite de variables aléatoires réelles i.i.d.. Alors, lorsque $n \to \infty$,*

$$\overline{X}_n \xrightarrow{\text{p.s.}} \mu$$

pour une certaine constante μ si et seulement si $X_1 \in \mathscr{L}^1$. Dans ce cas, $\mu = \mathbb{E}[X_1]$.

5. Se rappeler que $\cosh(u) = 1/\sqrt{1-\tanh^2(u)}$ et que $\operatorname{artanh}(u) = \frac{1}{2}\log\big((1+x)/(1-x)\big)$.

6. Sathamangalam Ranga Iyengar Srinivasa Varadhan (1940, Chennai –), probabiliste américain d'origine indienne. Lauréat du prix Abel en 2007.

Démonstration. Nous nous contenterons de démontrer la convergence et ne le ferons que sous l'hypothèse plus forte que $X_1 \in \mathscr{L}^4$. On peut supposer sans perte de généralité que $\mathbb{E}[X_1] = 0$ (sinon on considère les variables aléatoires $Y_k \coloneqq X_k - \mathbb{E}[X_k]$). Dans ce cas, le Lemme 6.4 implique que $\overline{X}_n$ satisfait

$$\mathbb{P}(|\overline{X}_n| \geqslant \epsilon) = \mathbb{P}(\overline{X}_n^4 \geqslant \epsilon^4) \leqslant \frac{\mathbb{E}[\overline{X}_n^4]}{\epsilon^4}.$$

Les variables aléatoires $(X_k)_{k\geqslant 1}$ étant i.i.d. avec $\mathbb{E}[X_1] = 0$, on a

$$\mathbb{E}[\overline{X}_n^4] = n^{-3}\mathbb{E}[X_1^4] + 3n^{-3}(n-1)\mathbb{E}[X_1^2]\mathbb{E}[X_2^2].$$

Il existe donc une constante C telle que

$$\forall \epsilon > 0, \qquad \sum_{n\geqslant 1} \mathbb{P}(|\overline{X}_n| \geqslant \epsilon) \leqslant \frac{C}{\epsilon^4} \sum_{n\geqslant 1} \frac{1}{n^2} < \infty.$$

La convergence presque sûre de $\overline{X}_n$ vers 0 suit donc du point 2 du Lemme 6.11. □

À présent que l'on sait que la moyenne empirique d'une suite $(X_k)_{k\geqslant 1}$ de variables aléatoires i.i.d. appartenant à $\mathscr{L}^1$ se concentre autour de leur espérance $\mu = \mathbb{E}[X_1]$, la question suivante est naturelle : que peut-on dire des fluctuations de la moyenne empirique autour de l'espérance, c'est-à-dire de la distribution de $\overline{X}_n - \mu$? Observons que si l'on a également $\mathrm{Var}[X_1] = \sigma^2 < \infty$, alors

$$\mathrm{Var}[\overline{X}_n - \mu] = \mathrm{Var}[\overline{X}_n] = \frac{1}{n^2}\mathrm{Var}\Big[\sum_{i=1}^n X_i\Big] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[X_i] = \frac{\sigma^2}{n}.$$

En particulier,

$$\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu) = \frac{1}{\sqrt{n\sigma^2}}\sum_{i=1}^n (X_i - \mu)$$

est d'ordre 1. Le **théorème central limite** fournit une version précise de cette affirmation et identifie même la loi asymptotique.

6.4 Le Théorème Central Limite

Théorème 6.16 (Théorème Central Limite, TCL). *Soit $(X_k)_{k\geqslant 1}$ une suite de variables aléatoires réelles i.i.d. telles que $\mathbb{E}[X_1] = \mu$ et $\mathrm{Var}[X_1] = \sigma^2 \in (0,\infty)$. Alors,*

$$\frac{1}{\sqrt{n\sigma^2}}\sum_{i=1}^n (X_i - \mu) \xrightarrow{\text{loi}} \mathcal{N}(0,1).$$

Démonstration. La preuve est presque identique à celle du Théorème 6.12. On peut supposer, sans perte de généralité, que $\mu = 0$ et $\sigma^2 = 1$ (sinon il suffit de considérer les variables aléatoires $Y_i \coloneqq (X_i - \mu)/\sigma$). Dans ce cas, il suit du Lemme 5.8 que $\phi_X \in \mathcal{C}^2$. Le théorème de Taylor–Young implique donc que

$$\phi_X(t) = 1 - \tfrac{1}{2}t^2 + \mathsf{o}(t^2).$$

D'autre part, les Lemmes 5.12 et 5.10 impliquent que la fonction caractéristique de la variable aléatoire $\widehat{X}_n \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$ satisfait

$$\phi_{\widehat{X}_n}(t) = \big(\phi_X(t/\sqrt{n})\big)^n = \Big(1 - \frac{t^2}{2n} + \mathsf{o}(\frac{t^2}{n})\Big)^n \xrightarrow{n\to\infty} e^{-t^2/2}.$$

On reconnaît, dans le membre de droite, la fonction caractéristique d'une variable aléatoire de loi $\mathcal{N}(0,1)$. Le résultat suit donc du Théorème de continuité (Théorème 5.19). □

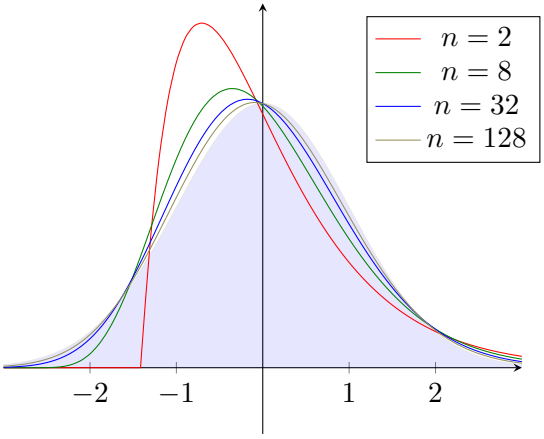


FIGURE 6.3: Convergence vers une loi normale pour une suite de variables aléatoires X_i de loi $\mathsf{exp}(1)$. Les courbes correspondent aux densités des variables $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - 1)$, pour $n = 2, 8, 32, 128$. La densité de la loi $\mathcal{N}(0,1)$ est indiquée par la région bleutée.

Insistons sur la caractère remarquable de ce résultat : quelle que soit la loi des variables aléatoires i.i.d. $(X_k)_{k\geqslant 1}$, pourvu que ces dernières aient une variance finie, les fluctuations de leur moyenne empirique suivent toujours une loi normale. C'est un exemple de phénomène d'**universalité** ; ces derniers intéressent beaucoup les physiciens et les probabilistes depuis quelques décennies.

Remarque 6.17. *Notons également que le théorème central limite montre que la variance est la mesure naturelle de la dispersion : c'est elle qui apparaît dans la normalisation de la somme dans l'énoncé du TCL, et non pas d'autres grandeurs qu'on aurait pu trouver plus naturelles, comme* $\mathbb{E}[|X - \mathbb{E}[X]|]$.

Exemple 6.18. Le Théorème Central Limite montre que, pour $n \gg 1$, on a

$$\mathbb{P}\Big(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \in [a,b]\Big) \simeq \Phi(b) - \Phi(a),$$

pour tout $a < b$ (ne dépendant pas de n), où l'on a utilisé la notation standard Φ pour la fonction de répartition associée à la loi normale standard. Ceci peut se réécrire,

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i \in [\widehat{a},\widehat{b}]\Big) \simeq \Phi\Big(\frac{\widehat{b} - n\mu}{\sqrt{n\sigma^2}}\Big) - \Phi\Big(\frac{\widehat{a} - n\mu}{\sqrt{n\sigma^2}}\Big).$$

Considérons une application élémentaire. Une chaîne de montage produit des pièces défectueuses avec un taux de 10%. Quelle est la probabilité d'obtenir au moins 50 pièces défectueuses parmi 400 ?

Modélisons cette situation par une suite d'épreuves de Bernoulli de paramètre $p = 0{,}1$. Avec $n = 400$, $n\mu = np = 40$ et $n\sigma^2 = np(1-p) = 36$, et en notant N le nombre de pièces défectueuses, on obtient

$$\mathbb{P}(N \geqslant 50) = \mathbb{P}(N \in [50, 400]) \simeq \Phi(\infty) - \Phi\Big(\frac{50 - 40}{\sqrt{36}}\Big) \simeq 0{,}05.$$

Il y a environ 5% de chances d'obtenir au moins 50 pièces défectueuses.

À titre de comparaison, N suivant une loi $\mathsf{binom}(400, 0{,}1)$, un calcul exact donne

$$\mathbb{P}(N \geqslant 50) = \sum_{k=50}^{400} \binom{400}{k} (0{,}1)^k\,(0{,}9)^{400-k} \simeq 0{,}06,$$

ce qui est assez proche de l'approximation précédente. Malheureusement, le Théorème 6.16 ne fournit pas de contrôle de l'erreur commise. ⌟

Pour les applications pratiques du théorème central limite (en particulier en statistique), comme dans l'exemple précédent, il est utile d'avoir des estimées quantitatives sur l'erreur commise lorsque l'on considère un n fini.

Théorème 6.19 (inégalité de Berry[7]–Esseen[8]). *Soit $(X_k)_{k\geqslant 1}$ une suite de variables aléatoires réelles i.i.d. telles que $\mathbb{E}[X_1] = \mu$, $0 < \mathrm{Var}[X_1] = \sigma^2 < \infty$ et $\mathbb{E}[|X_1|^3] < \infty$. Alors,*

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}\Big(\frac{1}{\sqrt{n\sigma^2}}\sum_{k=1}^{n}(X_k-\mu)\leqslant x\Big) - \Phi(x)\right| \leqslant C\frac{\mathbb{E}(|X_1-\mathbb{E}(X_1)|^3)}{\sigma^3\sqrt{n}},$$

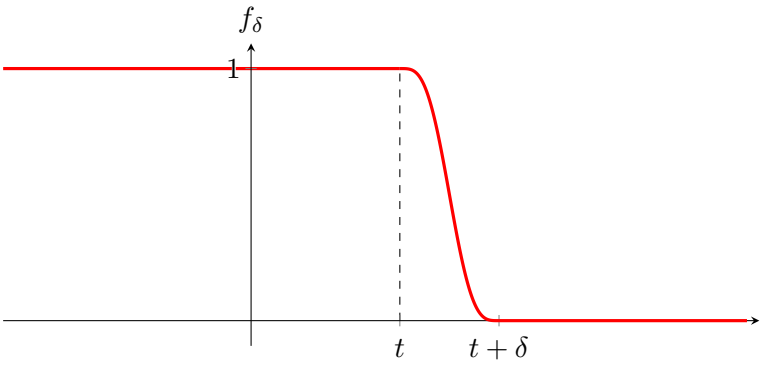
pour une certaine constante universelle $C < 0{,}4748$.

Démonstration. Pour simplifier la preuve, on ne démontre qu'une borne plus grossière sur l'erreur (d'ordre $n^{-1/8}$ plutôt que $n^{-1/2}$). On peut supposer, sans perte de généralité, que $\mu = 0$ et $\sigma^2 = 1$. Soit $(Z_k)_{k\geqslant 1}$ une suite de variables aléatoires i.i.d. de loi $\mathcal{N}(0,1)$, indépendantes des variables aléatoires X_k. On pose

$$\widehat{X}_n \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i, \qquad T_n \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i.$$

Observez que $T_n \sim \mathcal{N}(0,1)$, par l'Exemple 5.15. Soit $h : \mathbb{R} \to [0,1]$ une fonction de classe $\mathcal{C}^3$, telle que $h(s) = 1$ si $s \leqslant 0$, et $h(s) = 0$ si $s \geqslant 1$. Étant donné $t \in \mathbb{R}$ et $0 < \delta \leqslant 1$, on définit une nouvelle fonction $f_\delta : \mathbb{R} \to [0,1]$ par

$$f_\delta(x) \coloneqq h(\delta^{-1}(x-t)).$$



Par construction, $\mathbf{1}_{(-\infty,t]}(x) \leqslant f_\delta(x)$, pour tout $x \in \mathbb{R}$, et donc

$$\mathbb{P}(\widehat{X}_n \leqslant t) = \mathbb{E}[\mathbf{1}_{(-\infty,t]}(\widehat{X}_n)] \leqslant \mathbb{E}[f_\delta(\widehat{X}_n)].$$

Puisque $\Phi(t) = \mathbb{E}[\mathbf{1}_{(-\infty,t]}(T_n))$, on obtient

$$\mathbb{P}(\widehat{X}_n \leqslant t) - \Phi(t) \leqslant \mathbb{E}[f_\delta(\widehat{X}_n)] - \mathbb{E}[f_\delta(T_n)] + \mathbb{E}[f_\delta(T_n)] - \mathbb{E}[\mathbf{1}_{(-\infty,t]}(T_n)].$$

Comme $T_n \sim \mathcal{N}(0,1)$,

$$\mathbb{E}[f_\delta(T_n)] - \mathbb{E}[\mathbf{1}_{(-\infty,t]}(T_n)] = \frac{1}{\sqrt{2\pi}}\int_t^{t+\delta} h(\delta^{-1}(x-t))\, e^{-x^2/2}\mathrm{d}x \leqslant \frac{\delta}{\sqrt{2\pi}}.$$

7. Andrew C. Berry (1906, Somerville – 1998, Appleton), mathématicien américain.
8. Carl-Gustav Esseen (1918, Linköping – 2001, Uppsala), mathématicien suédois.

Il reste donc à estimer $\mathbb{E}[f_\delta(\widehat{X}_n)] - \mathbb{E}[f_\delta(T_n)]$. On le fait en réécrivant cette quantité sous la forme d'une somme télescopique, dans laquelle on remplace successivement une variable aléatoire X_i par une variable aléatoire Z_i :

$$\mathbb{E}[f_\delta(\widehat{X}_n)) - \mathbb{E}[f_\delta(T_n)] = \sum_{k=1}^{n} \big(\mathbb{E}[f_\delta(U_k + \frac{X_k}{\sqrt{n}})] - \mathbb{E}[f_\delta(U_k + \frac{Z_k}{\sqrt{n}})]\big),$$

où $U_k \coloneqq (Z_1 + Z_2 + \cdots + Z_{k-1} + X_{k+1} + X_{k+2} + \cdots + X_n)/\sqrt{n}$. Les variables aléatoires U_k, X_k et Z_k sont indépendantes. Par un développement de Taylor de f_δ autour de U_k, on peut écrire

$$f_\delta(U_k + (X_k/\sqrt{n})) = f_\delta(U_k) + \frac{X_k}{\sqrt{n}} f'_\delta(U_k) + \frac{X_k^2}{2n} f''_\delta(U_k) + \frac{X_k^3}{6n^{3/2}} f'''_\delta(Y),$$

où Y est entre U_k et $U_k + (X_k/\sqrt{n})$. On traite de la même façon le terme $f_\delta(U_k + (Z_k/\sqrt{n}))$. On obtient ainsi

$$\mathbb{E}[f_\delta(U_k + \frac{X_k}{\sqrt{n}})] - \mathbb{E}[f_\delta(U_k + \frac{Z_k}{\sqrt{n}})] \leqslant \frac{A}{6\delta^3 n^{3/2}} \big(\mathbb{E}[|X_k|^3] + \mathbb{E}[|Z_k|^3]\big),$$

où $A \coloneqq \sup_{y\in\mathbb{R}} |h'''(y)| = \delta^3 \sup_{y\in\mathbb{R}} |f'''_\delta(y)|$. En choisissant $\delta \coloneqq n^{-1/8}$, on obtient donc

$$\sup_{t\in\mathbb{R}} \big(\mathbb{P}(\widehat{X}_n \leqslant t) - \Phi(t)\big) \leqslant C n^{-1/8}.$$

La borne inférieure est prouvée de façon similaire, en remplaçant la fonction f_δ par la fonction $g_\delta(x) \coloneqq h(\delta^{-1}(x - t + \delta))$; observez que $g_\delta(x) \leqslant \mathbf{1}_{(-\infty,t]}(x)$ pour tout x. □

6.5 La loi zéro–un de Kolmogorov

Définition 6.20. *Soit $(X_k)_{k\geqslant 1}$ une suite de variables aléatoires sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. Soit $\mathcal{T}_n \coloneqq \sigma((X_k)_{k>n})$. La tribu $\mathcal{T}_\infty \coloneqq \bigcap_{n\geqslant 1} \mathcal{T}_n$ est appelée la **tribu asymptotique**. Les éléments de cette tribu sont les **événements asymptotiques**.*

La tribu asymptotique contient des événements comme

$$\Big\{(\sum_{i=1}^{n} X_i)_n \text{ converge}\Big\}, \quad \Big\{\lim_n X_n \text{ existe}\Big\} \quad \text{ou} \quad \Big\{\lim_n \frac{1}{n}(X_1 + \cdots + X_n) = 0\Big\}.$$

Ceux-ci sont indépendants des valeurs prises par les X_i, $i \in I$, pour tout ensemble fini I.

Théorème 6.21 (loi zéro–un de Kolmogorov)**.** *Soit $(X_k)_{k\geqslant 1}$ une suite variables aléatoires indépendantes. Alors tout événement $A \in \mathcal{T}_\infty$ satisfait $\mathbb{P}(A) \in \{0, 1\}$.*

Définition 6.22. *Une tribu dont tous les éléments sont de $\mathbb{P}$-probabilité 0 ou 1 est dite $\mathbb{P}$-**triviale**.*

Démonstration. Soit $A \in \mathcal{T}_\infty$. Étant donné que $A \in \mathcal{T}_n$, pour tout n, et que les tribus $\mathcal{T}_n$ et $\sigma(X_1, X_2, \ldots, X_n)$ sont indépendantes, il suit que A est indépendant de tout événement dans $\bigcup_n \sigma(X_1, X_2, \ldots, X_n)$.

On vérifie aisément que la classe des événements indépendants de A forme une classe monotone. Puisque cette classe contient le π-système $\bigcup_n \sigma(X_1, X_2, \ldots, X_n)$, il suit du théorème des classes monotones qu'elle contient également la tribu engendrée $\sigma((X_k)_{k\geqslant 1})$. Par conséquent, A est indépendant de $\sigma((X_k)_{k\geqslant 1})$.

Or, $A \in \mathcal{T}_\infty \subset \sigma((X_k)_{k\geqslant 1})$. On en déduit donc que A est indépendant de lui-même. Ceci implique que

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2,$$

et donc $\mathbb{P}(A) \in \{0, 1\}$. □

Définition 6.23. *Une variable aléatoire mesurable par rapport à la tribu asymptotique $\mathcal{T}_\infty$ est dite* ***asymptotique****.*

Corollaire 6.24. *Soit Y une variable aléatoire asymptotique réelle. Alors, il existe $y \in \mathbb{R}$ tel que*

$$\mathbb{P}(Y = y) = 1.$$

Démonstration. Y étant asymptotique,

$$\{\omega \in \Omega \,|\, Y(\omega) \leqslant x\} \in \mathcal{T}_\infty,$$

pour tout $x \in \mathbb{R}$. La loi zéro–un de Kolmogorov implique la $\mathbb{P}$-trivialité de $\mathcal{T}_\infty$. Par conséquent, la fonction de répartition de Y satisfait

$$F_Y(x) = \mathbb{P}(Y \leqslant x) \in \{0, 1\}.$$

Soit $y \coloneqq \inf\{x \,|\, \mathbb{P}(Y \leqslant x) = 1\}$. Comme $\lim_{x\to\infty} F_Y(x) = 1$ et $\lim_{x\to-\infty} F_Y(x) = 0$, il suit que $|y| < \infty$. On a donc $F_Y(x) = \mathbf{1}_{[y,\infty)}(x)$, ce qui implique que $Y = y$ presque sûrement. □

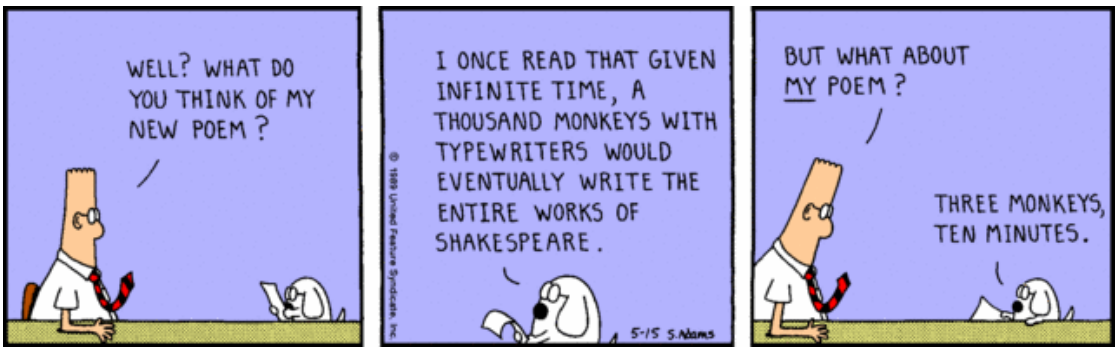
Exemple 6.25. Soit $(X_k)_{k\geqslant 1}$ une suite de variables aléatoires réelles indépendantes. Considérons la série aléatoire

$$P(z, \omega) \coloneqq \sum_{k=1}^{\infty} X_k(\omega) z^k.$$

Alors, le rayon de convergence $r(\omega)$ de cette série est presque sûrement constant, puisque

$$r(\omega) = (\limsup_{n\to\infty} |X_n(\omega)|^{1/n})^{-1},$$

est une variable aléatoire asymptotique. ⌟



©Scott Adams

7 Chaînes de Markov

Prediction is very difficult, especially about the future.

Niels Bohr

Dans ce chapitre, nous allons introduire une classe très importante de processus stochastiques : les chaînes de Markov. De manière informelle, une chaîne de Markov décrit un système dont l'évolution aléatoire est telle que la loi du système dans le futur ne dépend que de son état présent et pas de son histoire.

Les chaînes de Markov ont été introduites par Andreï Markov en 1902 (travail publié en 1906), afin de formaliser des problèmes d'épistémologie et de cryptage. En 1913, il les applique en linguistique (en particulier en critique textuelle) ; il a examiné, entre autres, la dépendance dans l'alternance des voyelles et des consonnes dans les premiers chapitres de « Eugène Onéguine » d'Alexandre Pouchkine et « Les Années d'enfance de Bagrov petit-fils » de Sergueï Aksakov.

Aujourd'hui, les chaînes de Markov ont de très nombreuses applications. Elles fournissent la base de la *méthode de Monte Carlo par chaînes de Markov* (MCMC), utilisée pour échantillonner selon des lois de probabilité complexes ou en très haute dimension ; elles sont employées en statistiques bayésiennes ; elles servent d'outil de modélisation en sciences naturelles et en sciences humaines ; elles jouent même un rôle central dans l'algorithme employé par google pour son moteur de recherche (PageRank).

7.1 Définition et exemples

Soit $(X_k)_{k\in\mathbb{N}}$ une suite de variables aléatoires à valeurs dans un ensemble S fini ou dénombrable. Nous noterons X le processus stochastique correspondant et $\mathbb{P}$ sa loi.

Définition 7.1. *Le processus X est une **chaîne de Markov** s'il possède la **propriété de Markov**,*

$$\mathbb{P}(X_n = s_n \,|\, X_0 = s_0, X_1 = s_1, \dots, X_{n-1} = s_{n-1}) = \mathbb{P}(X_n = s_n \,|\, X_{n-1} = s_{n-1}),$$

*pour tout $n \geqslant 1$ et tout $s_0, s_1, \dots, s_n \in S$. S est appelé l'**espace des états** de la chaîne.*

Les marches aléatoires du chapitre 3 fournissent un exemple de chaîne de Markov, avec $S = \mathbb{Z}$. La taille de la population dans le processus de branchement étudié dans la Section 4.2 est un autre exemple de processus de Markov, cette fois avec $S = \mathbb{N}$.

Définition 7.2. *Une chaîne de Markov X est **homogène** si*

$$\forall n \geqslant 1, \forall i, j \in S, \qquad \mathbb{P}(X_n = j \,|\, X_{n-1} = i) = \mathbb{P}(X_1 = j \,|\, X_0 = i).$$

Dorénavant, par souci de simplicité, nous allons supposer que S est un ensemble fini et que la chaîne de Markov est homogène. Dans ce cas, on voit que, si on connaît la loi de la chaîne en un temps donné, son évolution ultérieure est entièrement caractérisée par la matrice $\mathrm{P} = (p(i,j))_{i,j\in S}$ définie par

$$p(i,j) \coloneqq \mathbb{P}(X_1 = j \,|\, X_0 = i).$$

Définition 7.3. *La matrice P est appelée **matrice de transition** de la chaîne, et les probabilités $p(i,j)$ sont appelées **probabilités de transition** (de i à j).*

Lemme 7.4. *Une matrice de transition est caractérisée par les deux propriétés suivantes :*

1. *$p(i,j) \geqslant 0$ pour tout $i,j \in S$;*
2. *$\sum_{j\in S} p(i,j) = 1$ pour tout $i \in S$.*

*Une matrice possédant ces deux propriétés est appelée une **matrice stochastique**.*

Démonstration. Exercice élémentaire. □

Afin de spécifier complètement une chaîne de Markov, il reste à décrire son état initial.

Définition 7.5. *Soit μ une mesure de probabilité sur S et P une matrice stochastique. La chaîne de Markov (P,μ) est la chaîne de Markov (homogène dans le temps) de matrice de transition P et de loi initiale μ, c'est-à-dire telle que $\mathbb{P}(X_0 = i) = \mu(\{i\})$, pour tout $i \in S$. On écrira simplement $X \sim (\mathrm{P},\mu)$.*

Dans la suite, nous utiliserons les notations suivantes : la loi de la chaîne de Markov (P,μ) sera notée $\mathbb{P}_\mu$ et l'espérance correspondante $\mathbb{E}_\mu$. Dans le cas particulier où la loi initiale est concentrée sur un état $i \in S$, c'est-à-dire lorsque $\mu = \delta_i$, nous écrirons simplement $\mathbb{P}_i$ et $\mathbb{E}_i$.

Remarque 7.6. *L'existence d'un tel processus suit aisément du théorème d'extension de Kolmogorov (Théorème 2.86). En effet, les mesures μ_n décrivant les n premiers pas d'une chaîne de Markov $X \sim (\mathrm{P},\mu)$ sont données par*

$$\begin{aligned}\mu_n(\{(s_0,s_1,\dots,s_n)\}) &\coloneqq \mathbb{P}(X_0 = s_0,\dots,X_n = s_n)\\ &= \mathbb{P}(X_0 = s_0)\prod_{i=1}^{n}\mathbb{P}(X_i = s_i \,|\, X_0 = s_0,\dots,X_{i-1} = s_{i-1})\\ &= \mu(\{s_0\})\prod_{i=1}^{n} p(s_{i-1},s_i),\end{aligned}$$

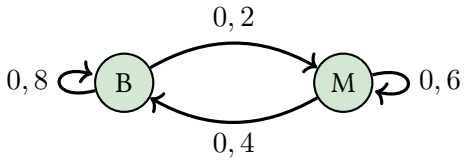
pour toute suite $(s_0,\dots,s_n)$. La consistance est immédiate, puisque

$$\begin{aligned}\sum_{s_{n+1}\in S}\mu_{n+1}(\{(s_0,s_1,\dots,s_n,s_{n+1})\}) &= \mu(\{s_0\})p(s_0,s_1)\dots p(s_{n-1},s_n)\sum_{s_{n+1}\in S}p(s_n,s_{n+1})\\ &= \mu(\{s_0\})p(s_0,s_1)\dots p(s_{n-1},s_n) = \mu_n(\{(s_0,s_1,\dots,s_n)\}).\end{aligned}$$

Exemple 7.7. Après une longue collecte de données, Robinson a conçu le modèle suivant pour prédire le temps qu'il fera sur son île :

$$S \coloneqq \{\mathrm{B},\mathrm{M}\}, \qquad \text{et} \qquad \mathrm{P} = \begin{array}{c} \\ \mathrm{B} \\ \mathrm{M}\end{array}\!\!\begin{array}{c}\begin{array}{cc}\mathrm{B} & \mathrm{M}\end{array}\\ \begin{pmatrix}0{,}8 & 0{,}2\\ 0{,}4 & 0{,}6\end{pmatrix}\end{array}.$$

où B représente l'état « beau temps » et M « mauvais temps ». La matrice P est stochastique et encode donc bien les probabilités de transition d'une chaîne de Markov sur S. Il est usuel de représenter de telles chaînes par un graphe comme sur la figure suivante :



Vendredi, quant à lui, a élaboré un modèle plus complexe, prédisant le temps du lendemain à partir du temps du jour et de celui de la veille. Le processus X qu'il obtient n'est plus une chaîne de Markov sur S, puisque la propriété de Markov n'est plus vérifiée. Il est cependant possible d'en déduire une chaîne de Markov sur un espace d'états étendu, en l'occurrence $S \times S$, en considérant les variables aléatoires $Y_n \coloneqq (X_{n-1}, X_n)$. En effet, la connaissance du couple $Y_n = (X_{n-1}, X_n)$ détermine X_n, et donc il ne reste plus qu'à prédire X_{n+1}, dont la probabilité est fonction uniquement de X_{n-1} et X_n. Les détails seront faits en exercice. ⌟

La matrice P contient toute l'information sur les probabilités de transition d'un état s au temps n vers un état s' au temps $n+1$. On peut facilement l'utiliser pour déterminer également les probabilités de transition d'un état s au temps m vers un état s' en un temps ultérieur $m+n$ quelconque. Notons

$$p_n(i,j) \coloneqq \mathbb{P}_i(X_n = j).$$

Alors, pour tout $n \geqslant 1$,

$$\begin{aligned} p_n(i,j) = \mathbb{P}_i(X_n = j) &= \sum_{k\in S} \mathbb{P}_i(X_n = j, X_{n-1} = k) \\ &= \sum_{k\in S} \mathbb{P}_i(X_n = j \mid X_{n-1} = k)\, \mathbb{P}_i(X_{n-1} = k) \\ &= \sum_{k\in S} \mathbb{P}_k(X_1 = j)\, \mathbb{P}_i(X_{n-1} = k) = \sum_{k\in S} p(k,j) p_{n-1}(i,k). \end{aligned}$$

Cette relation est connue sous le nom d'**équation de Chapman**[1]**–Kolmogorov**. On en déduit facilement le résultat fondamental suivant.

Théorème 7.8. *La **matrice de transition en n pas**, $\mathrm{P}_n = (p_n(i,j))_{i,j\in S}$, est donnée par la n-ième puissance de la matrice de transition P,*

$$\mathrm{P}_n = \mathrm{P}^n.$$

Démonstration. On peut réécrire l'équation de Chapman–Kolmogorov sous la forme

$$(\mathrm{P}_n)_{ij} = \sum_{k\in S} (\mathrm{P}_{n-1})_{ik} (\mathrm{P})_{kj} = (\mathrm{P}_{n-1}\mathrm{P})_{ij}.$$

En particulier, $\mathrm{P}_n = \mathrm{P}_{n-1}\mathrm{P} = \mathrm{P}_{n-2}\mathrm{P}^2 = \cdots = \mathrm{P}^n$. □

Ceci permet d'exprimer aisément la loi de la chaîne au temps n à partir de la loi de la chaîne au temps 0.

Corollaire 7.9. *Soit $X \sim (\mathrm{P}, \mu_0)$.*

$$\mu_n(\{i\}) \coloneqq \mathbb{P}_{\mu_0}(X_n = i) = \sum_{j\in S} \mu_0(\{j\}) p_n(j,i).$$

Démonstration.

$$\mu_n(\{i\}) = \mathbb{P}_{\mu_0}(X_n = i) = \sum_{j\in S} \mathbb{P}_{\mu_0}(X_n = i \mid X_0 = j) \mathbb{P}_{\mu_0}(X_0 = j) = \sum_{j\in S} p_n(j,i) \mu_0(\{j\}).$$ □

1. Sydney Chapman (1888, Eccles – 1970, Boulder), astronome et géophysicien britannique.

Toute mesure de probabilité μ sur S est entièrement déterminée par les valeurs prises sur les singletons, $\mu(i) := \mu(\{i\})$. Il sera souvent utile, dans ce chapitre, d'identifier une mesure de probabilité μ sur S avec le vecteur-ligne $(\mu(i))_{i\in S}$. Avec une telle convention, le résultat du corollaire précédent prend une forme plus simple :

$$\mu_n = \mu_0 \mathrm{P}^n,$$

où le membre de droite dénote le produit du vecteur-ligne μ_0 et de la matrice P^n.

Plus généralement, si l'on considère une fonction $f : S \to \mathbb{R}$, alors

$$\mathbb{E}_{\mu_0}[f(X_n)] = \sum_{i\in S} \mathbb{P}_{\mu_0}(X_n = i) f(i) = \sum_{i\in S} \mu_n(i) f(i) = \sum_{i\in S} (\mu_0 \mathrm{P}^n)(i) f(i) = \mu_0 \mathrm{P}^n f, \tag{7.1}$$

où l'on a interprété f comme un vecteur-colonne $(f(i))_{i\in S}$. Nous ferons de nombreuses identifications de ce type tout au long de ce chapitre.

Nous nous intéresserons principalement à deux classes particulières, mais très importantes, de chaînes de Markov : les chaînes irréductibles et les chaînes absorbantes.

Définition 7.10. *Soit* P *une matrice stochastique sur un ensemble* S.

- ▷ *Un état* $j \in S$ *est* ***atteignable*** *depuis l'état* $i \in S$, *noté* $i \to j$, *s'il existe* $n \geqslant 0$ *tel que* $p_n(i,j) > 0$.
- ▷ *Un état* $i \in S$ *est* ***absorbant*** *si* $p(i,i) = 1$.
- ▷ P *est* ***irréductible*** *si, pour tout* $i, j \in S$, *on a* $i \to j$.
- ▷ P *est* ***absorbante*** *si, pour tout* $i \in S$, *il existe* $j \in S$ *absorbant avec* $i \to j$.

Si X *est une chaîne de Markov de matrice de transition* P, *on dira que* X *est irréductible, resp. absorbante, lorsque* P *est irréductible, resp. absorbante.*

Par la suite, on notera $n(i,j) := \inf\{n \geqslant 1 \,|\, p_n(i,j) > 0\}$ le nombre minimal de pas permettant de passer de i à j avec probabilité positive ; en particulier, si $j \neq i$, alors $n(i,j) < \infty$ si et seulement si $i \to j$.

Exemple 7.11. On positionne un cavalier sur une des cases d'un échiquier (Fig. 7.1). À chaque pas, on déplace le cavalier aléatoirement sur une des cases accessibles depuis sa position actuelle (en respectant les règles de déplacement de cette pièce), de façon équiprobable. Combien de pas en moyenne faudra-t-il pour que le cavalier retourne à son point de départ ? On a ici un exemple de chaîne de Markov irréductible (vérifiez-le !). Nous verrons des méthodes permettant de répondre facilement à ce type de questions (*cf.* Exemple 7.39). ⌟

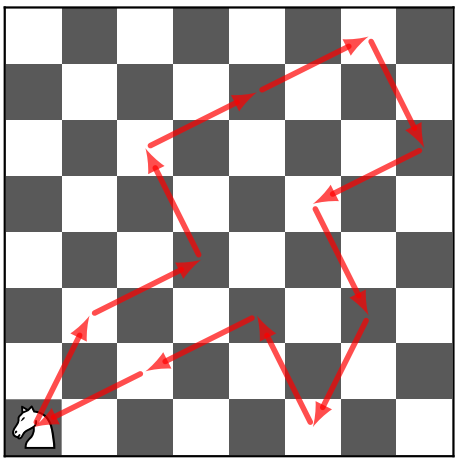


FIGURE 7.1: Quel est le nombre moyen de pas nécessaires pour que le cavalier se déplaçant au hasard sur l'échiquier se retrouve à son point de départ ?

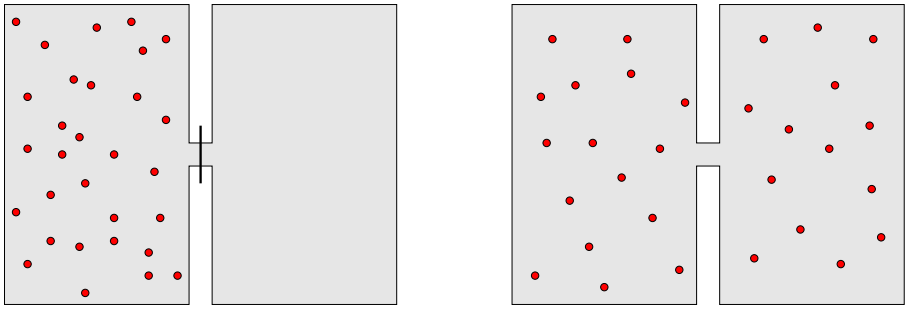
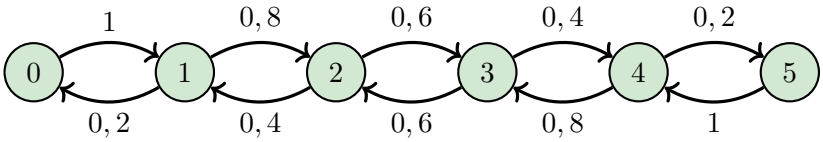


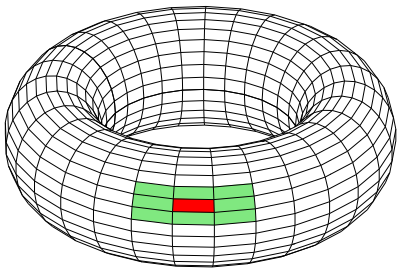
FIGURE 7.2: Au début de l'expérience, toutes les molécules du gaz sont confinées dans le récipient de gauche. Lorsque l'on retire la paroi séparant les deux récipients, les molécules se répartissent uniformément dans tout le volume disponible. Comment un tel comportement irréversible peut-il être compatible avec la réversibilité des équations d'évolution microscopiques ?

Exemple 7.12. Le **modèle des urnes d'Ehrenfest** a été introduit par Paul[2] et Tatiana[3] Ehrenfest en 1907 afin d'illustrer certains « paradoxes » liés à l'irréversibilité dans les fondements de la mécanique statistique, encore toute jeune. Le but est de modéliser l'évolution des molécules d'un gaz à l'intérieur d'un récipient. Plus particulièrement, on est intéressé au nombre de molécules se trouvant dans la moitié gauche et dans la moitié droite du récipient (voir Fig. 7.2). Leur modèle, très simplifié, de cette situation peut être formulé comme suit. On considère deux urnes A et B, et N boules numérotées de 1 à N. Initialement, toutes les boules se trouvent dans l'urne A. Ensuite, aux temps $1, 2, 3, \ldots$, un numéro entre 1 et N est tiré au hasard (uniformément) et la boule correspondante est déplacée de l'urne qu'elle occupe en ce moment vers l'autre. On note X_n le nombre de boules présentes dans l'urne A au temps n. La suite $X_0, X_1, \ldots$ est une chaîne de Markov sur $S = \{0, \ldots, N\}$. Le graphe correspondant, pour $N = 5$ est représenté dans la figure suivante :



La chaîne de Markov X est clairement irréductible. Nous retournerons à l'analyse de cette chaîne de Markov dans l'Exemple 7.37. ⌟

Exemple 7.13 (Modèle du votant)**.** Des modèles du type de celui que nous allons considérer à présent ont été utilisés entre autres en génétique. Ce modèle possède plusieurs noms, dont celui de modèle du votant. On considère une grille $n \times n$, dont chaque case est initialement peinte avec une couleur choisie parmi k. On suppose que cette grille est enroulée sur elle même de façon à former un tore. De cette manière, chaque case possède précisément 8 cases voisines :



La dynamique est la suivante : à chaque pas,

2. Paul Ehrenfest (1880, Vienne – 1933, Amsterdam), physicien théoricien autrichien.
3. Tatiana Alexeyevna Afanaseva (1876, Kiev – 1964, Leiden), mathématicienne russe et danoise.

1. on tire une case x au hasard (uniformément) ;
2. on choisit une de ses 8 voisines, y, au hasard (uniformément) ;
3. on repeint x de la couleur de y.

On vérifie aisément que la chaîne de Markov ainsi définie est absorbante, avec k états absorbants (les k configurations où toutes les cases sont de la même couleur).

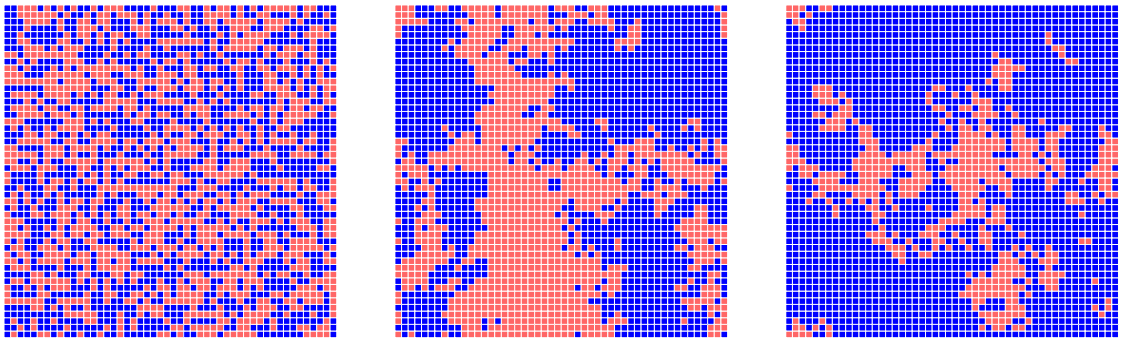


FIGURE 7.3 : Le modèle du votant (Exemple 7.13), pour $k = 2$, sur une grille 50×50 (représentée « à plat »). Gauche : état initial ; milieu : après $1\,000\,000$ de pas ; droite : après $10\,000\,000$ de pas.

La terminologie « modèle du votant » provient de l'interprétation suivante : chaque case représente un individu, et chaque couleur une opinion possible sur un certain sujet. À chaque itération du processus, un des individus discute avec l'un de ses voisins, se laisse convaincre par ce dernier et prend la même opinion. Les états absorbants correspondent alors aux k consensus possibles.

La figure 7.3 montre l'état initial de la chaîne, et deux états ultérieurs. Nous démontrerons plus tard (Exemple 7.21) que la probabilité que la chaîne finisse absorbée dans un état d'une certaine couleur est donnée par la fraction de cases de cette couleur, indépendamment de leur répartition géométrique ! ⌟

7.2 Chaînes de Markov absorbantes

Soit P une matrice stochastique absorbante et notons $\mathcal{A} \coloneqq \{i \in S \mid p(i,i) = 1\}$ l'ensemble des états absorbants. L'analyse des chaînes de Markov absorbantes est simplifiée si l'on écrit la matrice de transition sous sa forme **canonique**, c'est-à-dire en plaçant les états absorbants en dernier,

$$\mathrm{P} = \begin{pmatrix} \mathrm{Q} & \mathrm{R} \\ 0 & \mathrm{I} \end{pmatrix}.$$

Si $|S| = m$ et $|\mathcal{A}| = r$, Q est donc une matrice $(m-r) \times (m-r)$, R une matrice $(m-r) \times r$, 0 est la matrice nulle $r \times (m-r)$ et I la matrice identité $r \times r$.

Lemme 7.14. *Soit P une matrice de transition sous sa forme canonique. Alors, pour tout $n \geqslant 1$,*

$$\mathrm{P}^n = \begin{pmatrix} \mathrm{Q}^n & (\mathrm{I} + \mathrm{Q} + \cdots + \mathrm{Q}^{n-1})\mathrm{R} \\ 0 & \mathrm{I} \end{pmatrix}.$$

Démonstration. L'affirmation est triviale lorsque $n = 1$. Le cas $n \geqslant 2$ suit par récurrence puisque

$$\mathrm{P}^n = \mathrm{P}\mathrm{P}^{n-1} = \begin{pmatrix} \mathrm{Q} & \mathrm{R} \\ 0 & \mathrm{I} \end{pmatrix} \begin{pmatrix} \mathrm{Q}^{n-1} & (\mathrm{I} + \mathrm{Q} + \cdots + \mathrm{Q}^{n-2})\mathrm{R} \\ 0 & \mathrm{I} \end{pmatrix} = \begin{pmatrix} \mathrm{Q}^n & (\mathrm{I} + \mathrm{Q} + \cdots + \mathrm{Q}^{n-1})\mathrm{R} \\ 0 & \mathrm{I} \end{pmatrix}. \qquad \square$$

Théorème 7.15. *Soit* P *une matrice de transition mise sous forme canonique. Alors,*

$$\lim_{n\to\infty} \mathrm{Q}^n = 0.$$

En particulier, une chaîne de Markov absorbante finit toujours par se retrouver dans un état absorbant :

$$\forall i \notin \mathcal{A}, \qquad \lim_{n\to\infty} \mathbb{P}_i(X_n \notin \mathcal{A}) = 0.$$

Démonstration. Observons qu'il suit du Lemme 7.14 que

$$\forall i, j \notin \mathcal{A}, \forall n \geqslant 1, \qquad (\mathrm{Q}^n)_{ij} = (\mathrm{P}^n)_{ij} = \mathbb{P}_i(X_n = j) \leqslant \max_{i\in S} \mathbb{P}_i(X_n \notin \mathcal{A}).$$

Soit $M \coloneqq \max_{j\in S} \min\{m \,|\, \mathbb{P}_j(X_m \in \mathcal{A}) > 0\}$ et $p = \min_{i\in S} \mathbb{P}_i(X_M \in \mathcal{A}) > 0$. On a alors,

$$\max_{i\in S} \mathbb{P}_i(X_M \notin \mathcal{A}) = 1 - p.$$

Par conséquent, on déduit de la propriété de Markov que, pour tout $i \in S$ et tout $n \geqslant M$,

$$\begin{aligned}
\max_{i\in S} \mathbb{P}_i(X_n \notin \mathcal{A}) &= \max_{i\in S} \sum_{j\notin\mathcal{A}} \mathbb{P}_i(X_n \notin \mathcal{A} \,|\, X_M = j)\, \mathbb{P}_i(X_M = j) \\
&= \max_{i\in S} \sum_{j\notin\mathcal{A}} \mathbb{P}_j(X_{n-M} \notin \mathcal{A})\, \mathbb{P}_i(X_M = j) \\
&\leqslant \max_{j\in S} \mathbb{P}_j(X_{n-M} \notin \mathcal{A}) \max_{i\in S} \mathbb{P}_i(X_M \notin \mathcal{A}) \\
&= (1-p) \max_{j\notin\mathcal{A}} \mathbb{P}_j(X_{n-M} \notin \mathcal{A}) \\
&\leqslant \cdots \\
&\leqslant (1-p)^{\lfloor n/M \rfloor}, \qquad (7.2)
\end{aligned}$$

et la première affirmation suit en prenant la limite $n \to \infty$.

En particulier, pour tout $i \notin \mathcal{A}$,

$$\mathbb{P}_i(X_n \notin \mathcal{A}) = \sum_{j\notin\mathcal{A}} \mathbb{P}_i(X_n = j) = \sum_{j\notin\mathcal{A}} (\mathrm{Q}^n)_{ij} \xrightarrow{n\to\infty} 0. \qquad \square$$

Corollaire 7.16. *Soit* P *la matrice de transition d'une chaîne de Markov absorbante, sous forme canonique. Alors la matrice* $\mathrm{I} - \mathrm{Q}$ *est inversible et son inverse est donné par*

$$N \coloneqq (\mathrm{I} - \mathrm{Q})^{-1} = \mathrm{I} + \mathrm{Q} + \mathrm{Q}^2 + \cdots .$$

Démonstration. Soit v un vecteur tel que $(\mathrm{I} - \mathrm{Q})v = 0$, c'est-à-dire tel que $Qv = v$. Alors,

$$\forall n \geqslant 1, \qquad \mathrm{Q}^n v = \mathrm{Q}^{n-1}\mathrm{Q}v = \mathrm{Q}^{n-1}v = \cdots = Qv = v.$$

Il suit donc du Théorème 7.15 que

$$v = \lim_{n\to\infty} \mathrm{Q}^n v = 0,$$

ce qui montre que la matrice $\mathrm{I} - \mathrm{Q}$ n'admet pas 0 comme valeur propre et est donc inversible. À présent, il suffit d'observer que

$$(\mathrm{I} - \mathrm{Q})(\mathrm{I} + \mathrm{Q} + \mathrm{Q}^2 + \cdots + \mathrm{Q}^n) = \mathrm{I} - \mathrm{Q}^{n+1}$$

et, par conséquent,

$$\mathrm{I} + \mathrm{Q} + \mathrm{Q}^2 + \cdots + \mathrm{Q}^n = \mathrm{N}(\mathrm{I} - \mathrm{Q}^{n+1}).$$

En prenant la limite $n \to \infty$, on obtient donc

$$\mathrm{N} = \lim_{n\to\infty} \sum_{k=0}^{n} \mathrm{Q}^k. \qquad \square$$

Définition 7.17. *La matrice* N *est appelée* ***matrice fondamentale*** *de la chaîne.*

La matrice fondamentale d'une chaîne de Markov absorbante permet d'extraire de nombreuses propriétés de celle-ci. En particulier, elle permet de déterminer simplement le nombre moyen de visites en un état donné avant absorption, l'espérance du temps jusqu'à absorption partant d'un état donné, ainsi que les probabilités d'être absorbé dans un état donné k, étant parti d'un état i.

Théorème 7.18. *Soit* N *la matrice fondamentale de la chaîne et* $\tau_{\mathcal{A}} \coloneqq \min\{n \geqslant 0 \,|\, X_n \in \mathcal{A}\}$. *Alors,*

1. $\mathbb{E}_i[\sum_{k\geqslant 0} \mathbf{1}_{\{X_k=j\}}] = \mathrm{N}_{ij}$, *pour tout* $i, j \notin \mathcal{A}$;
2. $\mathbb{E}_i[\tau_{\mathcal{A}}] = \sum_{j\notin\mathcal{A}} \mathrm{N}_{ij}$, *pour tout* $i \notin \mathcal{A}$;
3. $\mathbb{P}_i(X_{\tau_{\mathcal{A}}} = j) = (\mathrm{NR})_{ij}$, *pour tout* $i \notin \mathcal{A}, j \in \mathcal{A}$.

Démonstration. *1.* Soit $i, j \notin \mathcal{A}$. Alors,

$$\mathbb{E}_i\Big[\sum_{n\geqslant 0} \mathbf{1}_{\{X_n=j\}}\Big] = \sum_{n\geqslant 0} \mathbb{P}_i(X_n = j) = \sum_{n\geqslant 0} (\mathrm{P}^n)_{ij} = \sum_{n\geqslant 0} (\mathrm{Q}^n)_{ij} = \mathrm{N}_{ij}.$$

2. Il suffit d'observer que, par le point précédent, pour tout $i \notin \mathcal{A}$,

$$\mathbb{E}_i[\tau_{\mathcal{A}}] = \mathbb{E}_i\Big[\sum_{n\geqslant 0} \mathbf{1}_{\{X_n\notin\mathcal{A}\}}\Big] = \sum_{j\notin\mathcal{A}} \mathbb{E}_i\Big[\sum_{n\geqslant 0} \mathbf{1}_{\{X_n=j\}}\Big] = \sum_{j\notin\mathcal{A}} \mathrm{N}_{ij}.$$

3. On a, pour tout $i \notin \mathcal{A}$ et $j \in \mathcal{A}$,

$$\begin{aligned}
\mathbb{P}_i(X_{\tau_{\mathcal{A}}} = j) &= \sum_{n\geqslant 1} \mathbb{P}_i(X_n = j, X_{n-1} \notin \mathcal{A}) \\
&= \sum_{n\geqslant 1} \sum_{k\notin\mathcal{A}} \mathbb{P}_i(X_n = j, X_{n-1} = k) \\
&= \sum_{n\geqslant 1} \sum_{k\notin\mathcal{A}} \mathbb{P}(X_n = j \,|\, X_{n-1} = k)\mathbb{P}_i(X_{n-1} = k) \\
&= \sum_{n\geqslant 1} \sum_{k\notin\mathcal{A}} \mathrm{R}_{kj}(\mathrm{Q}^{n-1})_{ik} \\
&= \sum_{n\geqslant 1} (\mathrm{Q}^{n-1}\mathrm{R})_{ij} \\
&= (\mathrm{NR})_{ij}. \qquad \square
\end{aligned}$$

Le théorème précédent permet en principe de calculer plusieurs quantités importantes. Dans la pratique cependant, le calcul peut se révéler laborieux, voire infaisable, en particulier lorsque la matrice de transition devient très grande. On doit alors recourir à d'autres outils...

Définition 7.19. *Soit* $f : S \to \mathbb{R}$ *et* $\mathrm{P} = (p(i,j))_{i,j\in S}$ *une matrice stochastique. On dit que* f *est une* ***fonction P-harmonique*** *si*

$$\forall i \in S, \qquad f(i) = \sum_{j\in S} p(i,j) f(j),$$

c'est-à-dire, sous forme vectorielle, $f = \mathrm{P}f$, *où* $f = (f(i))_{i\in S}$ *est vu comme un vecteur-colonne.*

Soit $f : S \to \mathbb{R}$ une fonction P-harmonique. Alors, il suit de (7.1) que, pour tout $n \in \mathbb{N}^*$ et tout $i \in S$,

$$\mathbb{E}_i[f(X_n)] = \delta_i \mathrm{P}^n f = \delta_i \mathrm{P}^{n-1} \mathrm{P} f = \delta_i \mathrm{P}^{n-1} f = \cdots = \delta_i f = f(i). \tag{7.3}$$

Le résultat suivant montre que cette dernière identité reste vraie lorsque le temps n est remplacé par le temps aléatoire $\tau_{\mathcal{A}}$ auquel la chaîne est absorbée.

Théorème 7.20. *Soient $(X_n)_{n\geqslant 0}$ une chaîne de Markov absorbante de matrice de transition P, $\tau_{\mathcal{A}}$ le temps d'absorption, et $\mathcal{A}$ l'ensemble des états absorbants. Alors, pour toute fonction f P-harmonique,*

$$\forall i \in S, \qquad \mathbb{E}_i[f(X_{\tau_{\mathcal{A}}})] = f(i).$$

Démonstration. L'affirmation est triviale lorsque $i \in \mathcal{A}$. Supposons donc que $i \notin \mathcal{A}$. Dans ce cas, il suit de (7.3) que

$$f(i) = \lim_{n\to\infty} \delta_i \mathrm{P}^n f = \delta_i \begin{pmatrix} 0 & \mathrm{NR} \\ 0 & \mathrm{I} \end{pmatrix} f = \sum_{j\in\mathcal{A}} (\mathrm{NR})_{ij} f(j) = \sum_{j\in\mathcal{A}} \mathbb{P}_i(X_{\tau_{\mathcal{A}}} = j) f(j) = \mathbb{E}_i[f(X_{\tau_{\mathcal{A}}})],$$

où la quatrième égalité suit du Théorème 7.18. □

Ce théorème (et ses extensions que l'on étudiera dans le Chapitre 10) se révèle souvent utile.

Exemple 7.21. Retournons au modèle introduit dans l'Exemple 7.13. On considère une grille $n \times n$ et k couleurs, notées $\{1, \dots, k\}$. On note P la matrice de transition associée. La fonction f donnant la fraction de cases de couleur 1 dans la configuration est P-harmonique. En effet, la dynamique revient à tirer au hasard (uniformément, donc avec une probabilité $1/(8n^2)$) une paire ordonnée (x, y) de cases voisines et à recolorier la case x avec la couleur de la case y. Le nombre de cases de couleur 1 va donc

- ▷ augmenter de 1 si la paire de sommets est telle que y soit de couleur 1, mais pas x ;
- ▷ diminuer de 1 si la paire de sommets est telle que x soit de couleur 1, mais pas y ;
- ▷ demeurer inchangé dans les autres cas.

On a donc, en notant $N_1(i)$ le nombre de 1 dans la configuration i,

$$\sum_{j\in S} p(i,j)\big(N_1(j) - N_1(i)\big) = \sum_{\substack{(x,y)\\ \text{voisins}}} \frac{1}{8n^2}\big(\mathbf{1}_{\{i(x)\neq 1, i(y)=1\}} - \mathbf{1}_{\{i(x)=1, i(y)\neq 1\}}\big),$$

où $i(x)$ est la couleur de la case x dans la configuration i. La dernière somme est nulle, puisque chaque contribution positive due à une paire (x, y) est compensée par la contribution négative de la paire (y, x). La fonction $f := N_1/n^2$ est donc bien P-harmonique.

Soit $\tau_{\mathcal{A}}$ le temps d'absorption de la chaîne, et notons $a_1, \dots, a_k$ les k états absorbants, a_ℓ représentant l'état où toutes les cases sont de couleur ℓ. Supposons à présent que la fraction de cases de couleur 1 dans l'état initial i_0 soit égale à ρ. Le Théorème 7.20 implique alors que

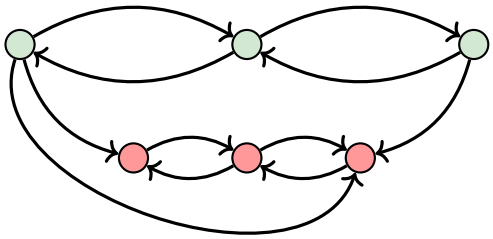
$$\rho = f(i_0) = \sum_{\ell=1}^{k} \mathbb{P}_{i_0}(X_{\tau_{\mathcal{A}}} = a_\ell) f(a_\ell).$$

Or, $f(a_1) = 1$ (puisque toutes les cases de a_1 sont de couleur 1) et $f(a_\ell) = 0$ pour $\ell \in \{2, \dots, k\}$. On a donc

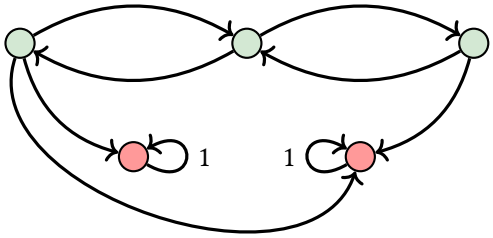
$$\rho = \mathbb{P}_{i_0}(X_{\tau_{\mathcal{A}}} = a_1).$$

En d'autres termes, à chaque instant, la probabilité que la chaîne finisse absorbée dans l'état absorbant de couleur ℓ est précisément donnée par la fraction de cases de couleur ℓ, un résultat qui serait probablement plus difficile à obtenir directement à partir du point 3 du Théorème 7.18. ⌟

Une remarque s'impose avant de conclure cette section. Considérons la chaîne de Markov correspondant au graphe suivant (la présence d'une flèche sans mention de probabilité de transition signifie que cette transition a probabilité strictement positive) :



Cette chaîne n'est pas absorbante, puisqu'aucun état n'est absorbant. Cependant, elle contient une sous-chaîne de laquelle il est impossible de s'échapper (les états représentés en rouge). L'analyse effectuée dans cette section permet d'obtenir très simplement des informations sur cette chaîne (par exemple, sur le temps moyen, ou le nombre de visites en un état donné, avant d'entrer dans cette sous-chaîne, ainsi que le point d'entrée) : il suffit de rendre absorbant chacun des états par lesquels on peut entrer dans la sous-chaîne ; on obtient ainsi la chaîne de Markov absorbante correspondant au graphe suivant :



Manifestement, si elles sont démarrées d'un même sommet vert, les trajectoires de ces deux chaînes ont la même loi, jusqu'au moment où elles atteignent pour la première fois un sommet rouge.

7.3 Chaînes de Markov irréductibles

Dans cette section, nous allons nous intéresser au cas des chaînes de Markov irréductibles. La terminologie suivante, déjà familière dans le contexte des marches aléatoires simples sur $\mathbb{Z}$, va se révéler utile.

Définition 7.22. *Soit $X = (X_n)_{n\geqslant 0}$ une chaîne de Markov sur S.*

- ▷ *Un état $i \in S$ est **récurrent** si $\mathbb{P}_i(\exists n \geqslant 1, X_n = i) = 1$. Sinon i est **transient**.*
- ▷ *X est **récurrente** si tous les états sont récurrents.*

Le résultat suivant donne une condition nécessaire et suffisante pour la récurrence d'un état (il ne suppose pas l'irréductibilité).

Lemme 7.23. *Un état j est récurrent si et seulement si $\sum_n p_n(j,j) = \infty$. Dans ce cas, $\sum_n p_n(i,j) = \infty$ pour tous les états i tels que j est accessible depuis i. Si j est transient, alors $\sum_n p_n(i,j) < \infty, \forall i \in S$.*

Observez qu'il suit que, si $j \in S$ est transient, alors

$$\forall i \in S, \qquad \lim_{n\to\infty} \mathbb{P}_i(X_n = j) = 0.$$

Démonstration. De façon similaire à ce que l'on a fait dans le cas des marches aléatoires, on introduit les fonctions génératrices

$$\mathbb{G}_{ij}(s) \coloneqq \sum_{n\geqslant 0} p_n(i,j)s^n, \qquad \mathbb{H}_{ij}(s) \coloneqq \sum_{n\geqslant 1} h_n(i,j)s^n,$$

où $h_n(i,j) \coloneqq \mathbb{P}_i(X_1 \neq j, X_2 \neq j, \dots, X_{n-1} \neq j, X_n = j)$. Notons que $\mathbb{H}_{ij}(1) = \mathbb{P}_i(\exists n \geqslant 1, X_n = j)$. En procédant exactement comme dans le Lemme 4.17, on obtient que, pour $i \neq j \in S$,

$$\mathbb{G}_{ii}(s) = 1 + \mathbb{H}_{ii}(s)\mathbb{G}_{ii}(s), \qquad \mathbb{G}_{ij}(s) = \mathbb{H}_{ij}(s)\mathbb{G}_{jj}(s).$$

Le lemme suit alors aisément. En effet,

$$\sum_n p_n(j,j) = \lim_{s\uparrow 1} \mathbb{G}_{jj}(s) = \lim_{s\uparrow 1}(1 - \mathbb{H}_{jj}(s))^{-1}$$

et cette dernière quantité est infinie si et seulement si $\mathbb{H}_{jj}(1) = 1$, ce qui est équivalent à dire que j est récurrent.

Pour les deux autres affirmations, on utilise $\sum_n p_n(i,j) = \mathbb{G}_{ij}(1) = \mathbb{H}_{ij}(1)\mathbb{G}_{jj}(1)$. Lorsque j est récurrent et accessible depuis i, $\mathbb{G}_{jj}(1) = \infty$ et $\mathbb{H}_{ij}(1) > 0$. Lorsque j est transient, $\mathbb{G}_{jj}(1) < \infty$ et $\mathbb{H}_{ij}(1) \leqslant 1$. □

Dans le cas d'une chaîne irréductible, on s'attend intuitivement à ce que tous les états soient visités infiniment souvent et donc que la chaîne soit récurrente.

Lemme 7.24. *Une chaîne X irréductible sur un espace d'états S fini est toujours récurrente. De plus, le **temps moyen de récurrence** dans l'état i est fini pour tout $i \in S$:*

$$\forall i \in S, \qquad \rho_i \coloneqq \mathbb{E}_i[T_i] < \infty,$$

*où $T_i \coloneqq \min\{n \geqslant 1 \,|\, X_n = i\}$. On dit que la chaîne est **récurrente-positive**.*

Démonstration. Observons tout d'abord qu'une telle chaîne possède toujours au moins un état récurrent. Si ce n'était pas le cas, on aurait, par le Lemme 7.23,

$$1 = \lim_{n\to\infty} \sum_{j\in S} p_n(i,j) = \sum_{j\in S} \lim_{n\to\infty} p_n(i,j) = 0,$$

puisque $\lim_{n\to\infty} p_n(i,j) = 0$ dès que j est transient.

Rappelons que $n(i,j) = \min\{n \geqslant 1 \,|\, p_n(i,j) > 0\}$. Montrons à présent que si $i \to j$, $j \to i$ et i est récurrent, alors j est également récurrent. Puisque $n(i,j) < \infty$ et $n(j,i) < \infty$, on a

$$\begin{aligned}\sum_{n\geqslant 1} p_n(j,j) &\geqslant \sum_{n\geqslant n(j,i)+n(i,j)} p_{n(j,i)}(j,i)p_{n-n(j,i)-n(i,j)}(i,i)p_{n(i,j)}(i,j) \\ &= p_{n(j,i)}(j,i)p_{n(i,j)}(i,j)\sum_{n\geqslant 0} p_n(i,i) = \infty\end{aligned}$$

et la première affirmation est démontrée.

Passons à la seconde. Soit $i \in S$ et $n \geqslant 1$. On a

$$\mathbb{P}_i(T_i > n) = \sum_{j\neq i} \mathbb{P}(T_i > n, X_1 = j) = \sum_{j\neq i} p(i,j)\mathbb{P}_j(T_i \geqslant n).$$

Afin de borner supérieurement le membre de droite, on considère la chaîne $(Y_n)_{n\geqslant 0}$ de matrice de transition $Q = (q(i,j))_{i,j\in S}$, où $q(k,j) \coloneqq p(k,j)$ pour tout $k \neq i$ et $q(i,j) \coloneqq \delta_{ij}$ (c'est-à-dire, on rend i absorbant) ;

notons $\mathbb{Q}$ la loi de cette chaîne. La chaîne $(X_n)_{n\geqslant 0}$ étant irréductible, il suit que $(Y_n)_{n\geqslant 0}$ est absorbante, avec $\mathcal{A} = \{i\}$. Par (7.2), il existe $p > 0$ et $M \in \mathbb{N}^*$ tels que

$$\forall n \in \mathbb{N}^*, \qquad \mathbb{P}_j(T_i \geqslant n) = \mathbb{Q}_j(\tau_{\mathcal{A}} \geqslant n) \leqslant (1-p)^{\lfloor (n-1)/M \rfloor}.$$

Ainsi,

$$\forall n \in \mathbb{N}^*, \qquad \mathbb{P}_i(T_i > n) \leqslant \sum_{j\neq i} p(i,j)(1-p)^{\lfloor (n-1)/M \rfloor} \leqslant (1-p)^{\lfloor (n-1)/M \rfloor}.$$

Par conséquent, on a bien $\rho_i = \mathbb{E}_i(T_i) = \sum_{n\geqslant 0} \mathbb{P}_i(T_i > n) < \infty$. □

Lemme 7.25. *Soit X une chaîne de Markov irréductible sur un espace d'états S fini. Alors,*

$$\forall i, j \in S, \qquad \mathbb{P}_j(\exists n \geqslant 1, X_n = i) = 1.$$

Démonstration. Soit $i, j \in S$, $i \neq j$. Manifestement, si $X_0 = i$ et $X_{n(i,j)} = j$, alors, $X_k \neq i$, pour tout $1 \leqslant k \leqslant n(i,j)$ (sinon $n(i,j)$ ne serait pas minimal). On a donc

$$\begin{aligned}\mathbb{P}_i(\forall n \geqslant 1, X_n \neq i) &\geqslant \mathbb{P}_i(\forall n \geqslant 1, X_n \neq i, X_{n(i,j)} = j) \\ &= \mathbb{P}_i(\forall n > n(i,j), X_n \neq i, X_{n(i,j)} = j) \\ &= p_{n(i,j)}(i,j)\, \mathbb{P}_j(\forall n \geqslant 1, X_n \neq i).\end{aligned}$$

On sait du Lemme 7.24 que X est récurrente, et par conséquent, le membre de gauche est nul. Puisque $p_{n(i,j)}(i,j) > 0$ par définition, on conclut que $\mathbb{P}_j(\forall n \geqslant 1, X_n \neq i) = 0$. □

7.3.1 Distribution stationnaire

Pour une chaîne de Markov irréductible X, le processus ne va pas s'arrêter dans un certain état, mais va continuer à évoluer éternellement. Une question fondamentale est alors de déterminer son comportement asymptotique : si l'on observe une telle chaîne après un temps très long, quelle est la probabilité qu'elle se trouve dans un état donné ? Avec quelle fréquence visite-t-elle chaque état ? La réponse à ces questions est étroitement liée à la notion de *distribution stationnaire*.

Supposons pour un instant que la loi de X_n se stabilise lorsque n devient très grand, c'est-à-dire que, pour un certain $i \in S$, il existe une mesure π sur S tel que $\lim_{n\to\infty} p_n(i,j) = \pi(j)$ pour tout $j \in S$. Alors, on devrait nécessairement avoir, d'une part, $\sum_{j\in S} \pi(j) = \lim_{n\to\infty} \sum_{j\in S} p_n(i,j) = 1$ et, d'autre part, pour tout $k \in S$,

$$\sum_{j\in S} \pi(j)p(j,k) = \lim_{n\to\infty} \sum_{j\in S} p_n(i,j)p(j,k) = \lim_{n\to\infty} p_{n+1}(i,k) = \pi(k).$$

Ceci motive la définition suivante.

Définition 7.26. *Une mesure de probabilité π sur S est une* ***distribution stationnaire*** *associée à la matrice de transition P si*

$$\pi = \pi \mathrm{P},$$

où l'on a identifié $(\pi(i))_{i\in S}$ à un vecteur-ligne.

La raison derrière cette terminologie est la suivante : si $X \sim (\mathrm{P}, \pi)$, alors il suit du Corollaire 7.9 que les probabilités d'occupation au temps n sont données par

$$\pi \mathrm{P}^n = (\pi \mathrm{P})\mathrm{P}^{n-1} = \pi \mathrm{P}^{n-1} = \cdots = \pi.$$

Ainsi, la distribution π est stationnaire : elle ne change pas lorsque le temps passe.

Nous allons à présent montrer que toute chaîne de Markov irréductible sur un espace des états fini possède une unique distribution stationnaire. Pour ce faire, introduisons, pour chaque $k \in S$, le vecteur-ligne $\gamma^k := (\gamma^k(i))_{i\in S}$ défini par

$$\gamma^k(i) := \mathbb{E}_k\Big[\sum_{n=0}^{T_k-1} \mathbf{1}_{\{X_n=i\}}\Big].$$

En d'autre termes, $\gamma^k(i)$ est le nombre moyen de visites en i, partant de k, avant le premier retour en k. Le théorème suivant montre qu'une distribution stationnaire existe toujours, lorsque la chaîne est irréductible et l'espace des états fini.

Théorème 7.27. *Soit* P *irréductible sur* S *fini. Alors, pour tout* $k \in S$,

1. $\gamma^k(k) = 1$;
2. $\gamma^k \mathrm{P} = \gamma^k$;
3. $\sum_{i\in S} \gamma^k(i) = \rho_k$;
4. $0 < \gamma^k(i) < \infty$, *pour tout* $i \in S$.

En particulier, pour tout $k \in S$, *la mesure de probabilité* π *sur* S *définie par*

$$\forall i \in S, \qquad \pi(i) := \gamma^k(i)/\rho_k$$

est une distribution stationnaire.

Démonstration. *1.* suit immédiatement de la définition.
2. D'une part, il suit du Lemme 7.24 qu'avec probabilité 1, $T_k < \infty$ et donc $X_0 = X_{T_k} = k$. D'autre part, l'événement $\{T_k \geqslant n\} = \{X_1 \neq k, X_2 \neq k, \dots, X_{n-1} \neq k\}$ ne dépendant que de $X_1, \dots, X_{n-1}$, la propriété de Markov au temps $n-1$ (et le fait que les deux membres sont nuls lorsque $i = k$) donne

$$\mathbb{P}_k(X_{n-1} = i, X_n = j, T_k \geqslant n) = \mathbb{P}_k(X_{n-1} = i, T_k \geqslant n)\, p(i,j), \qquad \forall i, j \in S.$$

On peut donc écrire

$$\begin{aligned}
\gamma^k(j) &= \mathbb{E}_k\big[\sum_{n=0}^{T_k-1} \mathbf{1}_{\{X_n=j\}}\big] = \mathbb{E}_k\big[\sum_{n=1}^{T_k} \mathbf{1}_{\{X_n=j\}}\big] = \mathbb{E}_k\big[\sum_{n=1}^{\infty} \mathbf{1}_{\{X_n=j, T_k\geqslant n\}}\big] \\
&= \sum_{n=1}^{\infty} \mathbb{P}_k(X_n = j, T_k \geqslant n) = \sum_{i\in S}\sum_{n=1}^{\infty} \mathbb{P}_k(X_{n-1} = i, X_n = j, T_k \geqslant n) \\
&= \sum_{i\in S} p(i,j) \sum_{n=1}^{\infty} \mathbb{P}_k(X_{n-1} = i, T_k \geqslant n) = \sum_{i\in S} p(i,j)\, \mathbb{E}_k\big(\sum_{n=1}^{\infty} \mathbf{1}_{\{X_{n-1}=i, T_k\geqslant n\}}\big) \\
&= \sum_{i\in S} p(i,j)\, \mathbb{E}_k\big(\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n=i, T_k-1\geqslant n\}}\big) = \sum_{i\in S} p(i,j)\, \mathbb{E}_k\big(\sum_{n=0}^{T_k-1} \mathbf{1}_{\{X_n=i\}}\big) \\
&= \sum_{i\in S} p(i,j)\, \gamma^k(i)\,.
\end{aligned}$$

3. suit directement de la définition :

$$\sum_{i\in S} \gamma^k(i) = \sum_{i\in S} \mathbb{E}_k\Big[\sum_{n=0}^{T_k-1} \mathbf{1}_{\{X_n=i\}}\Big] = \mathbb{E}_k\Big[\sum_{n=0}^{T_k-1}\sum_{i\in S} \mathbf{1}_{\{X_n=i\}}\Big] = \mathbb{E}_k[T_k] = \rho_k.$$

4. D'une part, il suit du point précédent que $\gamma^k(i) \leqslant \sum_{j\in S} \gamma^k(j) = \rho_k < \infty$. D'autre part, il suit de l'irréductibilité de la chaîne que, pour tout $i \in S$, $n(k,i) < \infty$. Par conséquent, on a

$$\gamma^k(i) = \sum_{j\in S} \gamma^k(j)\, p_{n(k,i)}(j,i) \geqslant \gamma^k(k)\, p_{n(k,i)}(k,i) = p_{n(k,i)}(k,i) > 0. \qquad \square \quad (7.4)$$

Le résultat suivant montre l'unicité de la distribution stationnaire d'une chaîne irréductible sur un espace des états finis et fournit une formule alternative utile, pour cette distribution.

Théorème 7.28. *Soit* P *une matrice stochastique irréductible sur un espace des états* S *fini. Alors,* P *possède une unique distribution stationnaire* π*. De plus,*

$$\forall i \in S, \qquad \pi(i) = \frac{1}{\rho_i},$$

où $\rho_i = \mathbb{E}_i[T_i] < \infty$ *est le temps moyen de récurrence dans l'état* i*.*

Démonstration. On commence par démontrer l'unicité. Fixons $k \in S$. Soit $\lambda = (\lambda(i))_{i\in S}$ un vecteur-ligne non-nul satisfaisant $\lambda(i) \geqslant 0$ pour tout $i \in S$ et $\lambda = \lambda\mathrm{P}$. L'argument utilisé en (7.4) implique que $\lambda(i) > 0$, pour tout $i \in S$; on peut donc supposer sans perte de généralité que $\lambda(k) = 1$. On a donc, pour tout $j \in S$,

$$\begin{aligned}
\lambda(j) &= \sum_{i_1\in S} \lambda(i_1)p(i_1,j) = \sum_{i_1\in S\setminus\{k\}} \lambda(i_1)p(i_1,j) + \lambda(k)p(k,j) \\
&= \sum_{i_1,i_2\in S\setminus\{k\}} \lambda(i_2)p(i_2,i_1)p(i_1,j) + \Big(p(k,j) + \sum_{i_1\in S\setminus\{k\}} p(k,i_1)p(i_1,j)\Big) \\
&= \dots \\
&= \sum_{i_1,\dots,i_n\in S\setminus\{k\}} \lambda(i_n)p(i_n,i_{n-1})\cdots p(i_1,j) \\
&\qquad + \Big(p(k,j) + \sum_{i_1\in S\setminus\{k\}} p(k,i_1)p(i_1,j) + \cdots + \sum_{i_1,\dots,i_{n-1}\in S\setminus\{k\}} p(k,i_{n-1})\cdots p(i_2,i_1)p(i_1,j)\Big) \\
&\geqslant \mathbb{P}_k(X_1 = j, T_k \geqslant 1) + \mathbb{P}_k(X_2 = j, T_k \geqslant 2) + \cdots + \mathbb{P}_k(X_n = j, T_k \geqslant n).
\end{aligned}$$

Cette dernière expression convergeant vers $\gamma^k(j)$ lorsque $n \to \infty$, on en déduit que $\lambda(j) \geqslant \gamma^k(j)$, pour tout $j \in S$. Par conséquent, le vecteur $\mu = \lambda - \gamma^k$ satisfait $\mu(i) \geqslant 0$ pour tout $i \in S$.

Par irréductibilité, pour chaque $i \in S$, $n(i,k) < \infty$. Comme $\mu\mathrm{P} = \lambda\mathrm{P} - \gamma^k\mathrm{P} = \lambda - \gamma^k = \mu$, on en conclut que

$$0 = \mu(k) = \sum_{j\in S} \mu(j)p_{n(i,k)}(j,k) \geqslant \mu(i)p_{n(i,k)}(i,k),$$

ce qui implique $\mu(i) = 0$. L'unicité suit.

La première partie et le théorème 7.27 impliquent que $\pi(k) = \gamma^k(k)/\sum_{i\in S}\gamma^k(i) = 1/\rho_k$. Le choix de l'état k étant arbitraire, la seconde affirmation est démontrée. □

7.3.2 Convergence

On a vu que si la loi de X_n converge, ce ne peut être que vers son unique distribution stationnaire. Il n'est cependant pas garanti que la loi de X_n converge, comme on peut le voir simplement en considérant la matrice de transition $\mathrm{P} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, qui donne lieu à une chaîne de Markov irréductible de distribution stationnaire $(\frac{1}{2}, \frac{1}{2})$, et pour laquelle la loi de X_n ne converge manifestement pas. Le problème ici est que la chaîne de Markov X a un comportement périodique.

Définition 7.29. *Soit* X *une chaîne de Markov sur* S *de matrice de transition* P*.*

- ▷ *Le nombre* $d(i) := \mathrm{pgcd}\{n \mid p_n(i,i) > 0\}$ *est la* ***période*** *de l'état* $i \in S$*.*
- ▷ *Un état* i *est* ***apériodique*** *si* $d(i) = 1$ *et* ***périodique*** *sinon.*
- ▷ X *est* ***apériodique*** *si tous ses états sont apériodiques.*
- ▷ X *est dite* ***ergodique*** *si elle est récurrente-positive, irréductible et apériodique.*

Lorsque S est fini, comme on le suppose dans ce chapitre, le Lemme 7.24 montre qu'une chaîne de Markov X sur S est ergodique si et seulement si elle est irréductible et apériodique.

Lemme 7.30. *Soit X une chaîne de Markov irréductible et apériodique. Alors, il existe $N < \infty$ tel que, pour tout $i, j \in S$,*

$$\forall n \geqslant N, \qquad p_n(i,j) > 0.$$

Démonstration. Soit $j \in S$. Par apériodicité, il existe $t_1, t_2, \dots, t_\ell \in \mathbb{N}^*$ tels que $\operatorname{pgcd}\{t_1, t_2, \dots, t_\ell\} = 1$ et $p_{t_k}(j,j) > 0$, pour tout $1 \leqslant k \leqslant \ell$. Par le Théorème de Bézout[4], il existe un entier $M = M(j)$ tel que tout nombre entier $m \geqslant M(j)$ peut se décomposer comme $m = \sum_{k=1}^{\ell} a_k t_k$, pour une suite $a_1, \dots, a_\ell$ d'entiers positifs. Par conséquent, on a

$$\forall m \geqslant M(j), \qquad p_m(j,j) \geqslant \prod_{k=1}^{\ell} (p_{t_k}(j,j))^{a_k} > 0.$$

Soit $i \in S$, $i \neq j$. Par irréductibilité, $n(i,j) < \infty$, et donc

$$\forall m \geqslant M(j) + n(i,j) \equiv M'(i,j), \qquad p_m(i,j) \geqslant p_{n(i,j)}(i,j) p_{m-n(i,j)}(j,j) > 0.$$

Comme il y a un nombre fini de paires $(i,j) \in S \times S$, on peut prendre $N := \max_{i,j \in S} M'(i,j)$. □

Le théorème suivant montre que la périodicité est la seule entrave possible à la convergence.

Théorème 7.31. *Soit P irréductible et apériodique sur S fini et μ une mesure de probabilité sur S. Alors,*

$$\lim_{n \to \infty} \mu \mathrm{P}^n = \pi,$$

où π est l'unique distribution stationnaire associée à P.

En particulier, si $X \sim (\mathrm{P}, \mu)$ avec P comme ci-dessus, les Théorèmes 7.28 et 7.31 impliquent que

$$\forall i \in S, \qquad \lim_{n \to \infty} \mathbb{P}_\mu(X_n = i) = 1/\rho_i.$$

Remarque 7.32. *On peut vérifier (cela sera fait en exercice) que pour une chaîne irréductible, tous les états ont même période d. Il suit alors que, si X est une chaîne irréductible de période d, alors les chaînes $Y^{(r)}$, $0 \leqslant r < d$, définies par $Y_n^{(r)} := X_{nd+r}$, sont apériodiques. On peut donc leur appliquer le théorème.*

Démonstration. Soient X et Y deux copies indépendantes de la chaîne de Markov et posons $Z_n := (X_n, Y_n)$. La chaîne de Markov Z sur $S \times S$ est irréductible. En effet, par le Lemme 7.30 et l'irréductibilité et l'apériodicité des chaînes indépendantes X et Y, il suit qu'il existe $N > 0$ tel que, pour tout $i, j, k, l \in S$ et tout $n \geqslant N$,

$$\mathbb{P}(Z_n = (k,l) \,|\, Z_0 = (i,j)) = \mathbb{P}(X_n = k \,|\, X_0 = i)\, \mathbb{P}(Y_n = l \,|\, Y_0 = j) = p_n(i,k) p_n(j,l) > 0.$$

Notons $\mathbb{P}_{(i,j)}$ la loi de la chaîne Z partant de $Z_0 = (i,j)$. Fixons $s \in S$ et considérons le temps aléatoire $T := \min\{n \geqslant 1 \,|\, Z_n = (s,s)\}$. Z étant irréductible, $\mathbb{P}_{(i,j)}(T < \infty) = 1$, pour tout $i, j \in S$. L'observation cruciale est la suivante :

$$\begin{aligned}
\mathbb{P}_{(i,j)}(X_n = k, T \leqslant n) &= \sum_{m \leqslant n} \mathbb{P}_{(i,j)}(X_n = k \,|\, T = m) \mathbb{P}_{(i,j)}(T = m) = \sum_{m \leqslant n} \mathbb{P}_{(s,s)}(X_{n-m} = k) \mathbb{P}_{(i,j)}(T = m) \\
&= \sum_{m \leqslant n} \mathbb{P}_{(s,s)}(Y_{n-m} = k) \mathbb{P}_{(i,j)}(T = m) = \sum_{m \leqslant n} \mathbb{P}_{(i,j)}(X_n = k \,|\, T = m) \mathbb{P}_{(i,j)}(T = m) \\
&= \mathbb{P}_{(i,j)}(Y_n = k, T \leqslant n),
\end{aligned}$$

4. Théorème de Bézout : si $x_1, \dots, x_m \in \mathbb{N}^*$ sont tels que $\operatorname{pgcd}(x_1, \dots, x_m) = d$, alors, pour tout $n \geqslant 0$, $\exists a_1, \dots, a_m \in \mathbb{Z}$ tels que $a_1 x_1 + \cdots + a_m x_m = nd$. De plus, si $n \geqslant x_1 \cdots x_m$, $a_1, \dots, a_m$ peuvent être choisis tous positifs.

où la troisième égalité suit de $\mathbb{P}_{(s,s)}(X_{n-m} = k) = p_{n-m}(s,k) = \mathbb{P}_{(s,s)}(Y_{n-m} = k)$, et la deuxième et la quatrième de la propriété de Markov. On peut donc écrire

$$\begin{aligned} p_n(i,k) = \mathbb{P}_{(i,j)}(X_n = k) &= \mathbb{P}_{(i,j)}(X_n = k, T \leqslant n) + \mathbb{P}_{(i,j)}(X_n = k, T > n) \\ &= \mathbb{P}_{(i,j)}(Y_n = k, T \leqslant n) + \mathbb{P}_{(i,j)}(X_n = k, T > n) \\ &\leqslant \mathbb{P}_{(i,j)}(Y_n = k) + \mathbb{P}_{(i,j)}(T > n) = p_n(j,k) + \mathbb{P}_{(i,j)}(T > n). \end{aligned}$$

On obtient donc

$$\forall i,j,k \in S, \qquad |p_n(i,k) - p_n(j,k)| \leqslant \mathbb{P}_{(i,j)}(T > n) \xrightarrow{n\to\infty} 0.$$

On en déduit que

$$\forall j,k \in S, \qquad \pi(k) - p_n(j,k) = \sum_{i\in S} \pi(i)\big(p_n(i,k) - p_n(j,k)\big) \xrightarrow{n\to\infty} 0$$

et donc,

$$\forall k \in S, \qquad \pi(k) - \sum_{j\in S} \mu(j) p_n(j,k) = \sum_{j\in S} \mu(j)\big(\pi(k) - p_n(j,k)\big) \xrightarrow{n\to\infty} 0. \qquad \square$$

7.3.3 Réversibilité

Dans de nombreux cas, en particulier pour les chaînes de Markov provenant de la modélisation de phénomènes physiques, la chaîne possède la propriété remarquable d'être invariante sous le renversement du temps, dans le sens que si l'on filme son évolution et que l'on passe le film, il est impossible de déterminer si le film est passé à l'endroit ou à l'envers. Bien entendu, ceci n'est possible que si la chaîne se trouve dans le régime stationnaire (sinon la relaxation vers l'équilibre permet de déterminer le sens d'écoulement du temps).

Soit P une matrice stochastique irréductible et π son unique distribution stationnaire. Fixons $N \in \mathbb{N}^*$ et considérons la chaîne de Markov $(X_n)_{0\leqslant n\leqslant N} \sim (\pi, \mathrm{P})$. On définit la **chaîne renversée** $\overleftarrow{X}$ par

$$\forall n \in \{0,\dots,N\}, \qquad \overleftarrow{X}_n \coloneqq X_{N-n}.$$

Lemme 7.33. *La chaîne renversée est une chaîne de Markov :* $\overleftarrow{X} \sim (\pi, \overleftarrow{\mathrm{P}})$, *où* $\overleftarrow{\mathrm{P}} = ((\overleftarrow{p}(i,j))_{i,j\in S})$ *avec*

$$\forall i,j \in S, \qquad \overleftarrow{p}(i,j) \coloneqq \frac{\pi(j)}{\pi(i)} p(j,i).$$

Démonstration. Vérifions tout d'abord que $\overleftarrow{X}$ est une chaîne de Markov :

$$\begin{aligned} \mathbb{P}_\pi(\overleftarrow{X}_{n+1} = i_{n+1} \,|\, \overleftarrow{X}_0 = i_0, \dots, \overleftarrow{X}_n = i_n) & \\ &= \mathbb{P}_\pi(X_{N-n-1} = i_{n+1} \,|\, X_N = i_0, \dots, X_{N-n} = i_n) \\ &= \frac{\mathbb{P}_\pi(X_N = i_0, \dots, X_{N-n} = i_n, X_{N-n-1} = i_{n+1})}{\mathbb{P}_\pi(X_N = i_0, \dots, X_{N-n} = i_n)} \\ &= \frac{\mathbb{P}_\pi(X_N = i_0, \dots, X_{N-n+1} = i_{n-1} \,|\, X_{N-n} = i_n, X_{N-n-1} = i_{n+1})}{\mathbb{P}_\pi(X_N = i_0, \dots, X_{N-n+1} = i_{n-1} \,|\, X_{N-n} = i_n)} \\ &\qquad \times \mathbb{P}_\pi(X_{N-n-1} = i_{n+1} \,|\, X_{N-n} = i_n) \\ &= \mathbb{P}_\pi(X_{N-n-1} = i_{n+1} \,|\, X_{N-n} = i_n) = \mathbb{P}_\pi(\overleftarrow{X}_{n+1} = i_{n+1} \,|\, \overleftarrow{X}_n = i_n), \end{aligned}$$

où l'on a utilisé la propriété de Markov pour $(X_n)_{0\leqslant n\leqslant N}$ pour la quatrième égalité.

Il reste à vérifier que $\overleftarrow{\mathrm{P}}$ est bien la matrice de transition de $\overleftarrow{X}$: par la formule de Bayes,

$$\begin{aligned} \mathbb{P}_\pi(\overleftarrow{X}_{n+1} = j \,|\, \overleftarrow{X}_n = i) &= \mathbb{P}_\pi(X_{-n-1} = j \,|\, X_{-n} = i) \\ &= \mathbb{P}_\pi(X_{-n} = i \,|\, X_{-n-1} = j) \frac{\mathbb{P}_\pi(X_{-n-1} = j)}{\mathbb{P}_\pi(X_{-n} = i)} = p(j,i)\frac{\pi(j)}{\pi(i)} = \overleftarrow{p}(i,j), \end{aligned}$$

où l'on a utilisé la stationnarité de π pour l'avant-dernière égalité. $\square$

Définition 7.34. *La chaîne X est **réversible** (à l'équilibre) si les matrices de transition de X et $\overleftarrow{X}$ sont identiques.*

La caractérisation suivante de la réversibilité est une conséquence immédiate du Lemme 7.33.

Théorème 7.35. *Soit $X \sim (\pi, \mathrm{P})$, où π est l'unique distribution stationnaire associée à P. Alors, la chaîne X est réversible si et seulement si la **condition d'équilibre local** est satisfaite :*

$$\forall i, j \in S, \qquad \pi(i)\, p(i,j) = \pi(j)\, p(j,i).$$

Une façon d'interpréter la condition d'équilibre local est comme suit. Imaginons que l'on répartisse un volume total d'eau égal à 1 entre les différents sommets du graphe associé à la chaîne de Markov. À chaque instant, une fraction $p(i,j)$ de l'eau se trouvant au sommet i est déplacée vers le sommet j (pour tous les sommets i, j simultanément). La distribution d'équilibre correspond à la répartition de l'eau sur les sommets telle que la quantité d'eau en chaque sommet est préservée : toute l'eau qui sort d'un sommet est compensée exactement par l'eau qui y entre ($\pi(i) = \sum_j \pi(j) p(j,i)$). La condition d'équilibre local est beaucoup plus forte : on demande à ce que, pour toute paire de sommets i, j, la quantité d'eau passant du sommet i au sommet j soit compensée exactement par la quantité d'eau passant du sommet j au sommet i ($\pi(i)\, p(i,j) = \pi(j)\, p(j,i)$).

Théorème 7.36. *Soit X une chaîne irréductible. S'il existe une mesure de probabilité π sur S telle que*

$$\forall i, j \in S, \qquad \pi(i)\, p(i,j) = \pi(j)\, p(j,i),$$

alors la chaîne est réversible (à l'équilibre) et de distribution stationnaire π.

Démonstration. Par hypothèse,

$$\sum_{j \in S} \pi(j)\, p(j,i) = \sum_{j \in S} \pi(i)\, p(i,j) = \pi(i),$$

ce qui montre que $\pi \mathrm{P} = \pi$. π est donc bien l'unique distribution stationnaire de la chaîne. □

Lorsqu'il s'applique, ce dernier théorème permet de déterminer beaucoup plus simplement la distribution stationnaire d'une chaîne de Markov irréductible : si l'on parvient à trouver une mesure de probabilité π sur S satisfaisant la condition d'équilibre local, alors π est nécessairement la mesure stationnaire de la chaîne.

Exemple 7.37. Retournons à la chaîne de Markov du modèle d'Ehrenfest (Exemple 7.12). Il est intuitivement clair que celle-ci devrait être réversible à l'équilibre. Il est donc naturel de chercher une mesure de probabilité π sur $\{0, \dots, N\}$ satisfaisant la condition d'équilibre local. Cette dernière peut s'écrire : pour tout $1 \leqslant k \leqslant N$,

$$\frac{\pi(k)}{\pi(k-1)} = \frac{p(k-1,k)}{p(k,k-1)} = \frac{(N-k+1)/N}{k/N} = \frac{N-k+1}{k}.$$

On en conclut que, pour tout $k \in \{1, \dots, N\}$, $\pi(k) = \binom{N}{k} m(0)$ et la condition de normalisation $1 = \sum_{k=0}^{N} \pi(k) = 2^N \pi(0)$ implique que la loi stationnaire est donnée par

$$\forall k \in \{0, \dots, N\}, \qquad \pi(k) = 2^{-N} \binom{N}{k}.$$

On peut à présent utiliser le Théorème 7.31 afin de déterminer les temps moyens de récurrence des divers états. Si, pour fixer les idées, on suppose qu'il y a une transition toutes les 10^{-10} secondes et $N = 10^{23}$ boules, on voit que le temps moyen nécessaire pour retourner dans l'état où toutes les boules sont dans l'urne A est

$$\mathbb{E}_N[T_N] = \frac{1}{\pi(N)} = \frac{2^N}{\binom{N}{N}} = 2^N \simeq 2^{10^{23}} \text{ secondes} \simeq 2^{10^{23}} \text{ âge de l'univers.}$$

À titre de comparaison, le temps moyen de récurrence de l'état dans lequel chacune des deux urnes contient la moitié des boules est de

$$\mathbb{E}_{N/2}[T_{N/2}] = \frac{1}{\pi(N/2)} = \frac{2^N}{\binom{N}{N/2}} \simeq \sqrt{\tfrac{1}{2}\pi N} \simeq 40 \text{ secondes}.$$

Ceci éclaire de manière particulièrement frappante le « paradoxe » entre la réversibilité microscopique et l'apparente irréversibilité macroscopique : si les molécules de gaz, toutes initialement contenues dans le récipient de gauche, se répartissent très rapidement de façon homogène entre les deux récipients, elles vont bien se retrouver dans l'état initial si l'on attend suffisamment longtemps, mais le temps nécessaire est tellement astronomique (immensément plus grand que l'âge de l'univers !), que cela n'aura jamais lieu en pratique. ⌟

Exemple 7.38. Soit $G = (S, A)$ un graphe fini simple[5] connexe. On considère la chaîne de Markov irréductible $(X_n)_{n\geqslant 0}$ sur S partant de $i_0 \in S$ et de matrice de transition $\mathrm{P} = (p(i,j))_{i,j\in S}$ donnée par

$$\forall i, j \in S, \qquad p(i,j) \coloneqq \frac{\mathbf{1}_{\{i\sim j\}}}{\deg(i)},$$

où $i \sim j$ signifie que i et j sont voisins, c'est-à-dire que $\{i,j\} \in A$, et $\deg(i) \coloneqq |\{j \in S \,|\, i \sim j\}|$ est le degré de i. $(X_n)_{n\geqslant 0}$ est appelée la **marche aléatoire simple sur G**.

Nous allons déterminer son unique distribution stationnaire à l'aide du Théorème 7.36. Cherchons donc une mesure de probabilité π sur S satisfaisant la condition d'équilibre local. Comme, pour tout $\{i,j\} \in A$,

$$\frac{p(i,j)}{p(j,i)} = \frac{\deg(j)}{\deg(i)},$$

on cherche une mesure de probabilité π de la forme $\pi(i) = C\deg(i)$, pour une constante $C > 0$. Cette dernière peut être facilement déterminée par la condition de normalisation :

$$1 = \sum_{i\in S} \pi(i) = C\sum_{i\in S} \deg(i) = C\sum_{i\in S}\sum_{j\in S} \mathbf{1}_{\{i\sim j\}} = C2|A|.$$

Par conséquent, la marche aléatoire simple sur G est réversible et son unique distribution stationnaire est donnée par

$$\forall i \in S, \qquad \pi(i) \coloneqq \frac{\deg(i)}{2|A|}. \tag{7.5}$$

⌟

Exemple 7.39. Revenons à l'Exemple 7.11. On considère le graphe fini simple connexe dont les sommets sont les 64 cases de l'échiquier et où deux sommets sont reliés par une arête si et seulement si le cavalier peut passer de l'un à l'autre en un déplacement. Quelques instants suffisent pour voir que les degrés des sommets du graphe sont donnés par

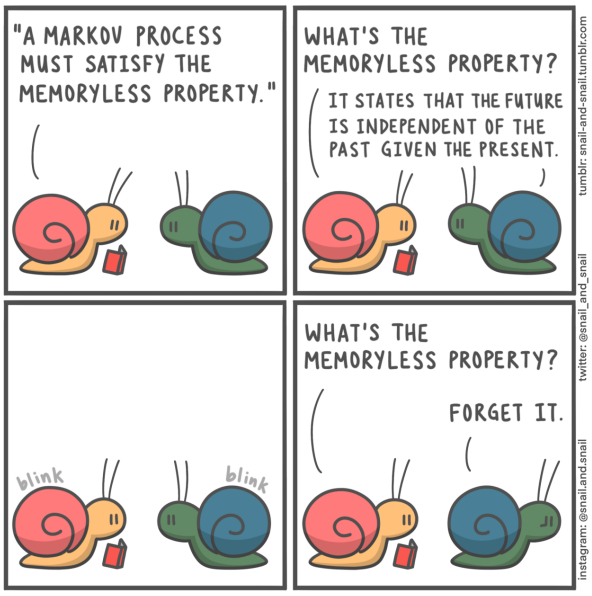
5. Rappelons que cela signifie que le graphe ne contient pas d'arête joignant un sommet à lui même et qu'une paire de sommets distincts est connectée par au plus une arête.

	a	b	c	d	e	f	g	h
8	2	3	4	4	4	4	3	2
7	3	4	6	6	6	6	4	3
6	4	6	8	8	8	8	6	4
5	4	6	8	8	8	8	6	4
4	4	6	8	8	8	8	6	4
3	4	6	8	8	8	8	6	4
2	3	4	6	6	6	6	4	3
1	2	3	4	4	4	4	3	2

En particulier, le nombre total d'arêtes du graphe est égal à $|A| = (4\cdot 2+8\cdot 3+20\cdot 4+16\cdot 6+16\cdot 8)/2 = 168$. La distribution stationnaire de cette chaîne de Markov est donnée par (7.5). Il est à présent très facile de déterminer le temps moyen de récurrence pour chaque case de départ du cavalier en utilisant le Théorème 7.28. On a, par exemple, lorsque le cavalier part de la case a1 :

$$\mathbb{E}_{\mathtt{a1}}[T_{\mathtt{a1}}] = \frac{1}{\pi(\mathtt{a1})} = \frac{2|A|}{\deg(\mathtt{a1})} = \frac{336}{2} = 168.$$

Il faut donc, en moyenne, 168 déplacements pour que le cavalier partant de la case a1 retourne à son point de départ. ⌟



https://snail-and-snail.tumblr.com

8 Marche aléatoire simple sur $\mathbb{Z}^d$

A drunk man will find his way home,
but a drunk bird may get lost forever.

Shizuo Kakutani

Dans le chapitre précédent, nous avons vu que toute chaîne de Markov irréductible sur un espace des états fini est nécessairement récurrente (positive). Dans ce chapitre, nous considérons à nouveau la marche aléatoire simple symétrique, mais sur le graphe $\mathbb{Z}^d$. Nous nous intéresserons à l'effet de la dimension d sur le comportement de la marche, en particulier en relation avec ses propriétés de récurrence ou transience et montrerons ainsi que, lorsque l'espace des états n'est plus fini, l'irréductibilité d'une chaîne de Markov ne suffit plus à garantir sa récurrence.

8.1 Définition du processus

Nous allons à présent brièvement décrire la marche aléatoire simple symétrique sur $\mathbb{Z}^d$. Ce processus (et ses généralisations) joue un rôle central en théorie des probabilités. Une interprétation naturelle est la description de la diffusion d'une particule (un tel modèle a par exemple été employé par Einstein[1] en 1905 afin d'expliquer le mouvement erratique des particules de pollen dans l'eau observé en 1827 par Brown[2] et de cette façon confirmer la théorie atomiste alors encore controversée en permettant à Perrin[3] de déterminer expérimentalement la constante d'Avogadro[4], un travail faisant partie de ceux pour lesquels Perrin recevra le prix Nobel en 1926).

Soit $(X_k)_{k\geqslant 1}$ une suite de variables aléatoires i.i.d. prenant valeurs dans l'ensemble $\{\pm\vec{e_i} \,|\, i = 1, \dots, d\}$ et de loi uniforme ; ici, $\vec{e_i} = (\delta_{ik})_{k=1,\dots,d}$ est le i-ième vecteur de la base canonique de $\mathbb{R}^d$. On appelle **marche aléatoire simple symétrique sur $\mathbb{Z}^d$ partant de $a \in \mathbb{Z}^d$** le processus

$$S_n \coloneqq a + \sum_{i=1}^{n} X_i.$$

Comme précédemment, on note $\mathbb{P}_a$ la loi de la marche partant de a.

1. Albert Einstein (1879, Ulm – 1955, Princeton), physicien allemand, puis apatride (1896), suisse (1899), et enfin suisse-américain (1940). Prix Nobel de physique en 1921.
2. Robert Brown (1773, Montrose – 1858, Londres), botaniste britannique.
3. Jean Baptiste Perrin (1870, Lille – 1942, New York), physicien français. Prix Nobel de Physique en 1926.
4. Lorenzo Romano Amedeo Carlo Avogadro, Comte de Quaregna et Cerreto (1776, Turin – 1856, Turin). Physicien et chimiste italien.

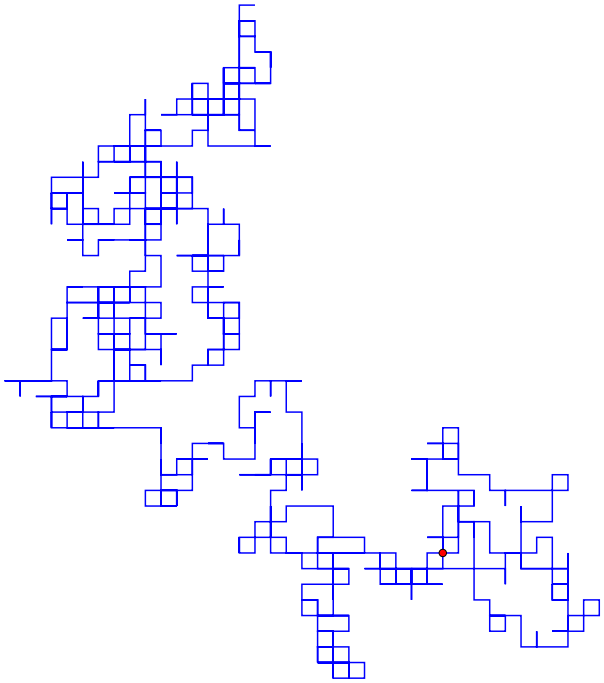


FIGURE 8.1: Ensemble des arêtes traversées lors des 1 000 premiers pas d'une marche aléatoire simple symétrique sur $\mathbb{Z}^2$ partant du point rouge.

Ce processus décrit donc une particule se déplaçant aléatoirement de proche en proche sur le réseau $\mathbb{Z}^d$. On vérifie aisément que les propriétés énoncées dans le Lemme 3.1 sont également vérifiées ici (la structure étant identique).

8.2 Récurrence et transience des marches aléatoires simples symétriques sur $\mathbb{Z}^d$

Nous allons nous intéresser à un problème classique : déterminer si la marche aléatoire simple symétrique sur $\mathbb{Z}^d$ est récurrente ou transiente. Notre étude est basée sur la caractérisation suivante de la récurrence.

Lemme 8.1. *Soit N le nombre de retours de la marche aléatoire simple symétrique à l'origine. Alors,*

$$S_n \text{ est récurrente} \iff \mathbb{E}_0[N] = \infty \iff \sum_{n\geqslant 1} \mathbb{P}_0(S_n = 0) = \infty.$$

Démonstration. Soit $r := \mathbb{P}_0(N \geqslant 1)$ la probabilité de retour à l'origine et soit $\tau_0^{(n)}$ le temps auquel se produit le n-ième retour en 0 (avec $\tau_0^{(n)} := \infty$ si $N < n$). Il suit de la propriété de Markov que, pour tout $n \geqslant 1$,

$$\begin{aligned}\mathbb{P}_0(N \geqslant n \,|\, N \geqslant n-1) &= \sum_{k\geqslant 2n-2} \mathbb{P}_0(N \geqslant n \,|\, \tau_0^{(n-1)} = k)\, \mathbb{P}_0(\tau_0^{(n-1)} = k \,|\, N \geqslant n-1) \\ &= r \sum_{k\geqslant 2n-2} \mathbb{P}_0(\tau_0^{(n-1)} = k \,|\, N \geqslant n-1) = r.\end{aligned}$$

Il suit donc que $\mathbb{P}_0(N \geqslant n) = r\,\mathbb{P}_0(N \geqslant n-1) = r^2\,\mathbb{P}_0(N \geqslant n-2) = \cdots = r^n$. Par conséquent,

$$\mathbb{E}_0[N] = \sum_{n\geqslant 1} \mathbb{P}_0(N \geqslant n) = \begin{cases} r/(1-r) & \text{si } r < 1 \\ \infty & \text{si } r = 1 \end{cases}$$

ce qui démontre la première équivalence. Comme

$$\mathbb{E}_0[N] = \mathbb{E}_0\big[\sum_{n\geqslant 1} \mathbf{1}_{\{S_n=0\}}\big] = \sum_{n\geqslant 1} \mathbb{P}_0(S_n = 0),$$

le lemme est démontré. □

Lorsque $d = 1$, nous avons établi que la marche symétrique est récurrente-nulle (Théorème 4.18). Le résultat suivant a été démontré par Pólya[5] en 1921 ; il montre que la dimension du réseau affecte crucialement le comportement de la marche aléatoire.

Théorème 8.2. *La marche aléatoire simple symétrique sur $\mathbb{Z}^d$ est récurrente si et seulement si $d \leqslant 2$.*

Démonstration. Il existe de nombreuses preuves de ce résultat. Une façon assez élémentaire de le démontrer est de déterminer exactement la probabilité de retour à l'origine et d'utiliser la formule de Stirling et des bornes appropriées.

Nous allons passer par les fonctions caractéristiques, car cet argument est beaucoup plus robuste (en particulier, il n'est pas limité aux marche aléatoires simples, comme on le verra en exercices).

En utilisant l'identité

$$\forall x \in \mathbb{Z}^d, \qquad \int_{[-\pi,\pi]^d} \frac{\mathrm{d}p}{(2\pi)^d} e^{\mathrm{i}p\cdot x} = \mathbf{1}_{\{x=0\}},$$

on obtient

$$\mathbb{P}_0(S_n = 0) = \int_{[-\pi,\pi]^d} \frac{\mathrm{d}p}{(2\pi)^d} \mathbb{E}_0[e^{\mathrm{i}p\cdot S_n}],$$

et $\mathbb{E}_0[e^{\mathrm{i}p\cdot S_n}] = (\mathbb{E}[e^{\mathrm{i}p\cdot X_1}])^n = (\phi_{X_1}(p))^n$. Un calcul élémentaire montre que la fonction caractéristique conjointe de X_1 satisfait

$$\forall p = (p_1, \dots, p_d), \qquad \phi_{X_1}(p) = \frac{1}{d}\sum_{i=1}^{d} \cos(p_i). \tag{8.1}$$

Par conséquent, pour tout $0 < \lambda < 1$,

$$\sum_{n\geqslant 0} \lambda^n \mathbb{P}_0(S_n = 0) = \int_{[-\pi,\pi]^d} \frac{\mathrm{d}p}{(2\pi)^d} \sum_{n\geqslant 0} (\lambda\phi_{X_1}(p))^n = \int_{[-\pi,\pi]^d} \frac{\mathrm{d}p}{(2\pi)^d} \frac{1}{1-\lambda\phi_{X_1}(p)}.$$

On souhaiterait prendre la limite $\lambda \uparrow 1$ à présent, afin de pouvoir utiliser le Lemme 8.1, mais cela nécessite quelques précautions. Pour le membre de gauche, cela suit immédiatement du théorème de convergence monotone. En ce qui concerne le membre de droite, on commence par observer que $\phi_{X_1}(p)$ est réelle et positive pour tout $p \in [-1,1]^d$. Par conséquent, il suit du Théorème de convergence monotone que

$$\lim_{\lambda\uparrow 1} \int_{[-1,1]^d} \frac{\mathrm{d}p}{(2\pi)^d} \frac{1}{1-\lambda\phi_{X_1}(p)} = \int_{[-1,1]^d} \frac{\mathrm{d}p}{(2\pi)^d} \frac{1}{1-\phi_{X_1}(p)}.$$

Pour traiter le reste du domaine d'intégration, on observe que la suite de fonctions $1/(1-\lambda\phi_{X_1}(p))$ converge ponctuellement et est uniformément bornée sur $[-\pi,\pi]^d \setminus [-1,1]^d$. Par conséquent, il suit du Théorème de convergence dominée que

$$\lim_{\lambda\uparrow 1} \int_{[-\pi,\pi]^d\setminus[-1,1]^d} \frac{\mathrm{d}p}{(2\pi)^d} \frac{1}{1-\lambda\phi_{X_1}(p)} = \int_{[-\pi,\pi]^d\setminus[-1,1]^d} \frac{\mathrm{d}p}{(2\pi)^d} \frac{1}{1-\phi_{X_1}(p)}.$$

On a donc bien

$$\sum_{n\geqslant 0} \mathbb{P}_0(S_n = 0) = \int_{[-\pi,\pi]^d} \frac{\mathrm{d}p}{(2\pi)^d} \frac{1}{1-\phi_{X_1}(p)}.$$

5. George Pólya (1887, Budapest – 1985, Palo Alto), mathématicien hongrois.

Le problème se réduit ainsi à l'analyse de la divergence de l'intégrande du membre de droite en $p = 0$. Par le théorème de Taylor–Lagrange,

$$\forall x \in \mathbb{R}_+, \exists x_0 \in (0, x), \qquad \cos(x) = 1 - \tfrac{1}{2}x^2 + \tfrac{1}{24}x_0^4.$$

Par conséquent, pour tout $x \in [-1, 1]$,

$$1 - \tfrac{1}{2}x^2 \leqslant \cos(x) \leqslant 1 - \tfrac{11}{24}x^2.$$

On déduit donc de (8.1) que $\frac{1}{2d}\|p\|_2^2 \geqslant 1-\phi_{X_1}(p) \geqslant \frac{11}{24d}\|p\|_2^2$ au voisinage de 0. Ainsi, l'intégrande se comporte comme $\|p\|_2^{-2}$ au voisinage de 0. Par conséquent, l'intégrale converge si et seulement si $d > 2$. □

Remarque 8.3. *Le résultat précédent montre que lorsque $d \geqslant 3$, la probabilité π_d de retour au point de départ est strictement inférieure à 1. Il est en fait facile de la déterminer. On montrera en exercice que $\pi_d = 1 - 1/u_d$, où*

$$u_d \coloneqq \int_{[-\pi,\pi]^d} \frac{\mathrm{d}p}{(2\pi)^d} \frac{1}{1 - \phi_{X_1}(p)}.$$

On obtient ainsi, par exemple : $\pi_3 \simeq 0{,}340$, $\pi_4 \simeq 0{,}193$, $\pi_5 \simeq 0{,}135$, etc.

8.3 Récurrence nulle de la marche aléatoire simple symétrique sur $\mathbb{Z}^2$

On a vu (dans la Remarque 3.5 et dans le Théorème 4.18) que la marche aléatoire simple symétrique sur $\mathbb{Z}$ est récurrente-nulle. Dans cette section, nous considérons le problème correspondant en dimension 2.

Théorème 8.4. *La marche aléatoire simple symétrique sur $\mathbb{Z}^2$ est récurrente-nulle.*

Démonstration. Notons $S_n = (S_n(1), S_n(2))$ la marche aléatoire simple symétrique sur $\mathbb{Z}^2$ et notons $X_k = (X_k(1), X_k(2))$, $k \geqslant 1$, les incréments correspondants. On a déjà vu que $(S_n)_{n\geqslant 0}$ est récurrente. Il suffit donc de montrer que $E_0[\tau_0] = \infty$.

On vérifie très facilement que le processus $\widetilde{S}_n \coloneqq S_n(1)+S_n(2)$ est une marche aléatoire simple symétrique sur $\mathbb{Z}$ (il suffit de voir que les variables aléatoires $(X_n(1)+X_n(2))_{n\geqslant 1}$ sont i.i.d. et de loi uniforme sur $\{-1, 1\}$). Par conséquent, si on note $\widetilde{\tau}_0$ le temps de premier retour de $\widetilde{S}_n$ en 0, on a

$$\mathbb{E}_0[\tau_0] = \mathbb{E}_0[\inf\{n \geqslant 1 \mid S_n(1) = S_n(2) = 0\}] \geqslant \mathbb{E}_0[\inf\{n \geqslant 1 \mid \widetilde{S}_n = 0\}]) = \mathbb{E}_0[\widetilde{\tau}_0] = \infty,$$

puisque la marche aléatoire simple symétrique sur $\mathbb{Z}$ est récurrente-nulle. □

8.4 Quelques commentaires sur la convergence vers le mouvement brownien

On considère la marche aléatoire simple symétrique $(S_n)_{n\geqslant 0}$ sur $\mathbb{Z}$. Le théorème central limite implique que,

$$\forall t \in \mathbb{R}_+, \qquad \frac{1}{\sqrt{N}} S_{[tN]} \xrightarrow{\text{loi}} \mathcal{N}(0, t) \qquad \text{lorsque } N \to \infty.$$

Il est en fait possible de démontrer (un résultat appelé **principe d'invariance**) qu'une convergence de ce type a lieu pour la loi des *trajectoires* du processus. On obtient ainsi, dans la limite, un processus $(B_t)_{t\in\mathbb{R}_+}$, dont chaque réalisation est presque sûrement une fonction continue, mais nulle part différentiable. Ce processus est appelé **mouvement brownien** ou **processus de Wiener**[6]. Une partie d'une trajectoire de ce processus est représentée sur la Figure 8.2.

De façon similaire, on peut montrer la convergence en loi de la marche aléatoire simple sur $\mathbb{Z}^d$ vers un processus limite $(B_t)_{t\in\mathbb{R}_+}$ à valeurs dans $\mathbb{R}^d$, dont les trajectoires sont, presque sûrement, continues mais nulle part différentiables. Sur la figure 8.3, on a tracé une portion de trajectoire dans le cas bidimensionnel.

6. Norbert Wiener (1894, Columbia – 1964, Stockholm), mathématicien américain.

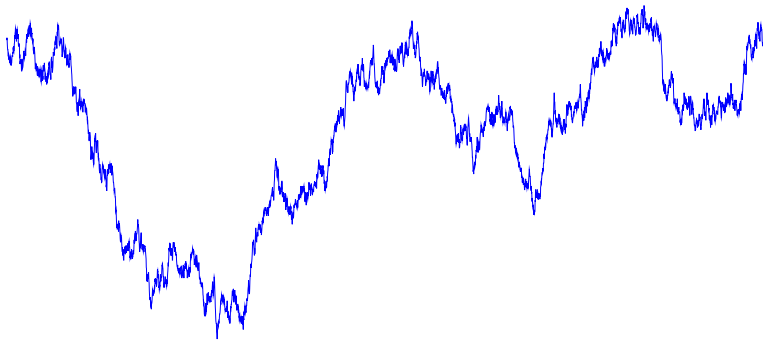


FIGURE 8.2: Partie d'une trajectoire du mouvement brownien en dimension 1.

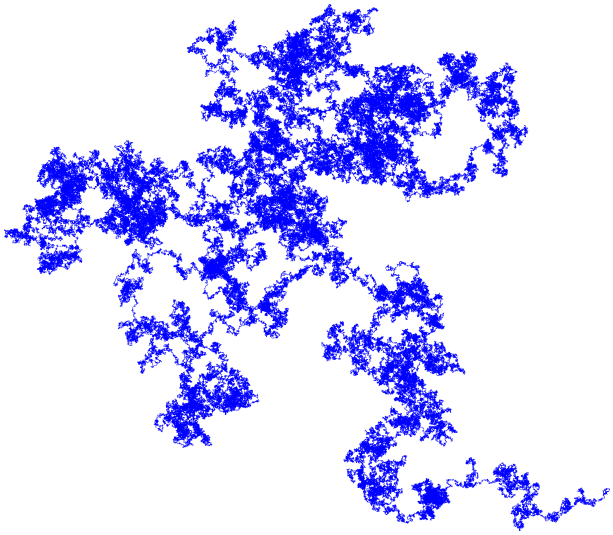
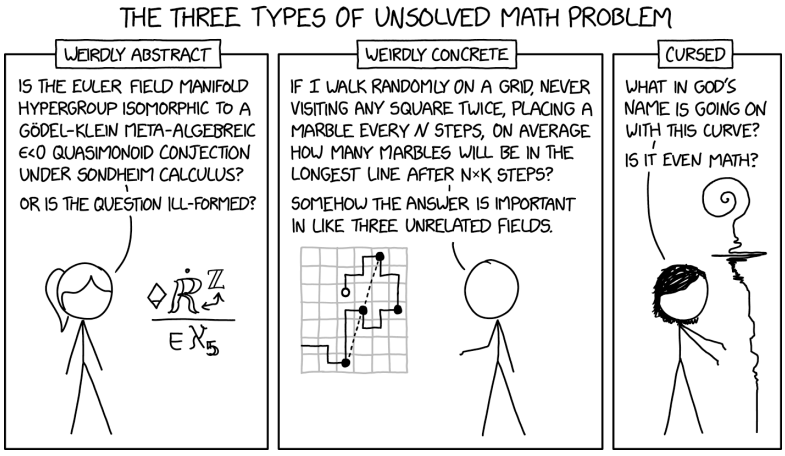


FIGURE 8.3: Ensemble des points visités par une partie de la trajectoire du mouvement brownien en dimension 2.



https://xkcd.com/2529

9 Espérance conditionnelle

La probabilité conditionnelle de gagner un milliard de dollars sachant que vous possédez un demi-milliard est la même que celle de gagner un million de dollars si vous en possédez déjà un demi-million. L'argent va à l'argent, la puissance à la puissance. Injuste, peut-être, mais vrai - à la fois socialement et mathématiquement.

Benoît Mandelbrot

9.1 Motivation, définition, existence et unicité

Considérons deux variables aléatoires réelles discrètes X et Y à valeurs dans deux ensembles dénombrables ou finis E et F. Notons $F' \coloneqq \{y \in F \mid \mathbb{P}(Y = y) > 0\}$. On suppose que $X, Y \in \mathscr{L}^1$. On peut alors écrire

$$\forall x \in E, \forall y \in F', \qquad \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}.$$

On peut alors considérer l'espérance de X par rapport à la mesure de probabilité $\mathbb{P}(\cdot \mid Y = y)$:

$$\forall y \in F', \qquad \mathbb{E}[X \mid Y = y] \coloneqq \sum_{x \in E} x \mathbb{P}(X = x \mid Y = y).$$

On définit une variable aléatoire $\mathbb{E}[X \mid Y]$ par

$$\mathbb{E}[X \mid Y](\omega) \coloneqq \begin{cases} \mathbb{E}[X \mid Y = Y(\omega)] & \text{si } Y(\omega) \in F' \\ 0 & \text{sinon.} \end{cases}$$

Cette variable aléatoire est une **version** de l'**espérance conditionnelle** de X sachant Y. Le mot version vient du fait qu'on aurait, par exemple, pu modifier la définition de cette variable lorsque $Y(\omega) \in F \setminus F'$, puisque cela n'affecte presque sûrement pas la valeur de $\mathbb{E}[X \mid Y]$.

On interprète cette variable aléatoire comme suit : on fait une expérience dont le résultat ω ne nous est pas connu, mais dont on connaît la valeur qu'il associe à la variable aléatoire Y ; ensuite, on calcule l'espérance de X conditionnellement au fait que Y prend la valeur $Y(\omega)$.

Observons que $\mathbb{E}[X \mid Y]$ ne dépend de la variable aléatoire Y qu'à travers la partition de Ω induite par Y :

$$\Omega = \bigcup_{y \in F} \{\omega \mid Y(\omega) = y\} =: \bigcup_{y \in F} G_y.$$

La tribu $\mathcal{G} := \sigma(Y)$ engendrée par la variable aléatoire Y est explicitement donnée par

$$\mathcal{G} = \{Y^{-1}(B) \mid B \subset F\} = \{\bigcup_{y \in B} G_y \mid B \subset F\}.$$

Par définition, $\mathbb{E}[X \mid Y]$ est constante sur chaque G_y, ce qui signifie que

$$\mathbb{E}[X \mid Y] \text{ est } \mathcal{G}\text{-mesurable.} \tag{9.1}$$

De plus, si $y \in F'$,

$$\begin{aligned}\int_{G_y} \mathbb{E}[X \mid Y](\omega)\, \mathrm{d}\mathbb{P}(\omega) = \mathbb{E}[X \mid Y = y]\mathbb{P}(Y = y) &= \sum_{x \in E} x\mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y) \\ &= \sum_{x \in E} x\mathbb{P}(X = x, Y = y) = \sum_{x \in E} x \int_{G_y} \mathbf{1}_{\{X(\omega) = x\}}\, \mathrm{d}\mathbb{P}(\omega) \\ &= \int_{G_y} \sum_{x \in E} x\mathbf{1}_{\{X(\omega) = x\}}\, \mathrm{d}\mathbb{P}(\omega) = \int_{G_y} X(\omega)\, \mathrm{d}\mathbb{P}(\omega).\end{aligned}$$

Cette identité reste vraie si $y \in F \setminus F'$, puisque les deux expressions sont nulles dans ce cas. Tout $G \in \mathcal{G}$ pouvant s'écrire comme union disjointe d'ensembles G_y, on obtient donc

$$\forall G \in \mathcal{G}, \qquad \mathbb{E}\big[\mathbb{E}[X \mid Y]\, \mathbf{1}_G\big] = \mathbb{E}\big[X\mathbf{1}_G\big]. \tag{9.2}$$

Les deux propriétés (9.1) et (9.2) peuvent être utilisées afin de donner un sens à la notion d'espérance conditionnelle pour des variables aléatoires générales. Notons que si sa définition peut sembler abstraite à première vue, cette quantité joue un rôle essentiel en théorie des probabilités.

Théorème 9.1. *Soit $X \in \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ et $\mathcal{G} \subset \mathcal{F}$ une sous-tribu. Il existe une variable aléatoire Z telle que*

1. *Z est $\mathcal{G}$-mesurable,*
2. *$\mathbb{E}[|Z|] < \infty$,*
3. *$\mathbb{E}[Z\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G]$ pour tout $G \in \mathcal{G}$.*

De plus, si $\tilde{Z}$ est une autre variable aléatoire avec ces propriétés, alors $\mathbb{P}(Z = \tilde{Z}) = 1$.

*Une variable aléatoire Z satisfaisant les trois propriétés ci-dessus est appelée une **version de l'espérance conditionnelle** $\mathbb{E}[X \mid \mathcal{G}]$ de X sachant $\mathcal{G}$ et on écrit $Z = \mathbb{E}[X \mid \mathcal{G}]$ presque sûrement.*

Notation : Si $\mathcal{G} = \sigma(Y)$, on écrit $\mathbb{E}[X \mid Y] \equiv \mathbb{E}[X \mid \mathcal{G}]$. Plus généralement, si $\mathcal{G} = \sigma(Y_i, i \in I)$, on notera $\mathbb{E}[X \mid Y_i, i \in I] \equiv \mathbb{E}[X \mid \mathcal{G}]$.

Exemple 9.2. Soit $X \in \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P})$.

- ▷ Soit $\mathcal{G} := \{\varnothing, \Omega\}$. Par la première propriété, $\mathbb{E}[X \mid \mathcal{G}]$ est presque sûrement constante. Par conséquent, $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}\big[\mathbb{E}[X \mid \mathcal{G}]\big] = \mathbb{E}[X]$ presque sûrement, la seconde égalité suivant de la 3e propriété avec $G = \Omega$.
- ▷ Soit $\mathcal{G} := \mathcal{F}$. Alors, $\mathbb{E}[X \mid \mathcal{G}] = X$, presque sûrement, puisque X satisfait aux 3 conditions.

▷ Soit X, Y deux variables aléatoires réelles possédant une densité de probabilité conjointe $f_{(X,Y)}$. Alors,

$$f_X(x) \coloneqq \int_{\mathbb{R}} f_{(X,Y)}(x,y)\,\mathrm{d}y \qquad \text{et} \qquad f_Y(y) \coloneqq \int_{\mathbb{R}} f_{(X,Y)}(x,y)\,\mathrm{d}x$$

sont des densités de probabilité marginales pour X et Y respectivement. On définit la **densité de probabilité conditionnelle** de X sachant Y par

$$f_{X\,|\,Y}(x\,|\,y) \coloneqq \begin{cases} f_{(X,Y)}(x,y)/f_Y(y) & \text{si } f_Y(y) > 0, \\ 0 & \text{sinon.} \end{cases}$$

Soit h telle que $h(X) \in \mathscr{L}^1$. Posons

$$g(y) \coloneqq \int_{\mathbb{R}} h(x) f_{X\,|\,Y}(x\,|\,y)\,\mathrm{d}x.$$

Alors, $g(Y)$ est une version de $\mathbb{E}[h(X)\,|\,Y]$. En effet, pour tout $G \coloneqq \{\omega\,|\,Y(\omega) \in B\}$, $B \in \mathcal{B}(\mathbb{R})$, on a

$$\begin{aligned} \mathbb{E}[g(Y)\mathbf{1}_G] = \mathbb{E}[g(Y)\mathbf{1}_B(Y)] &= \int_{\mathbb{R}} g(y)\mathbf{1}_B(y) f_Y(y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} h(x) f_{X\,|\,Y}(x\,|\,y)\,\mathrm{d}x\,\mathbf{1}_B(y) f_Y(y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} h(x) f_{(X,Y)}(x,y)\mathbf{1}_B(y)\,\mathrm{d}x\mathrm{d}y = \mathbb{E}[h(X)\mathbf{1}_B(Y)] = \mathbb{E}[h(X)\mathbf{1}_G], \end{aligned}$$

par le théorème de Fubini–Tonelli. ⌟

Preuve du Théorème 9.1. ▷ *Unicité presque sûre.* Soit Y_1 et Y_2 deux versions de $\mathbb{E}[X\,|\,\mathcal{G}]$. Il suit alors des propriétés de l'espérance conditionnelle que $Y_1, Y_2 \in \mathcal{L}^1(\Omega, \mathcal{G}, \mathbb{P})$ et

$$\forall G \in \mathcal{G}, \qquad \mathbb{E}[(Y_1 - Y_2)\mathbf{1}_G] = \mathbb{E}[Y_1\mathbf{1}_G] - \mathbb{E}[Y_2\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G] - \mathbb{E}[X\mathbf{1}_G] = 0. \tag{9.3}$$

Supposons que $\mathbb{P}(Y_1 \neq Y_2) \neq 0$. Sans perte de généralité, on considère le cas où $\mathbb{P}(Y_1 > Y_2) > 0$. Comme $\{Y_1 > Y_2 + \frac{1}{n}\} \uparrow \{Y_1 > Y_2\}$, il suit que $\mathbb{P}(Y_1 > Y_2 + \frac{1}{n}) > 0$ lorsque n est suffisamment grand. Évidemment, $\{Y_1 - Y_2 > \frac{1}{n}\} \in \mathcal{G}$, puisque Y_1, Y_2 sont $\mathcal{G}$-mesurables. Par conséquent, il suit de (9.3) que

$$0 = \mathbb{E}[(Y_1 - Y_2)\mathbf{1}_{\{Y_1 - Y_2 > \frac{1}{n}\}}] \geqslant \tfrac{1}{n}\mathbb{P}(Y_1 > Y_2 + \tfrac{1}{n}) > 0,$$

ce qui est absurde.

▷ *Existence lorsque* $X \in \mathscr{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. On suppose tout d'abord que $X \in \mathscr{L}^2(\Omega, \mathcal{F}, \mathbb{P})$.

Les espaces $L^2(\Omega, \mathcal{F}, \mathbb{P})$ et $L^2(\Omega, \mathcal{G}, \mathbb{P})$ sont des espaces de Hilbert pour la norme $\|\cdot\|_2$ induite par le produit scalaire

$$\langle Y, Z\rangle \coloneqq \int Y(\omega)Z(\omega)\,\mathrm{d}\mathbb{P}(\omega).$$

De plus, $L^2(\Omega, \mathcal{G}, \mathbb{P})$ est un sous-espace vectoriel fermé de $L^2(\Omega, \mathcal{F}, \mathbb{P})$. On peut donc considérer la projection orthogonale $\pi_{\mathcal{G}} : L^2(\Omega, \mathcal{F}, \mathbb{P}) \to L^2(\Omega, \mathcal{G}, \mathbb{P})$. On a alors, pour tout $G \in \mathcal{G}$,

$$\int_G \pi_{\mathcal{G}} X\,\mathrm{d}\mathbb{P} = \int \mathbf{1}_G\,\pi_{\mathcal{G}} X\,\mathrm{d}\mathbb{P} = \langle \mathbf{1}_G, \pi_{\mathcal{G}} X\rangle = \langle \pi_{\mathcal{G}}\mathbf{1}_G, X\rangle = \langle \mathbf{1}_G, X\rangle = \int_G X\,\mathrm{d}\mathbb{P},$$

où la 3e identité suit du fait que $\pi_{\mathcal{G}}$ est autoadjointe et la 4e du fait que $\mathbf{1}_G \in L^2(\Omega, \mathcal{G}, \mathbb{P})$. En outre, $\pi_{\mathcal{G}} X$ est $\mathcal{G}$-mesurable par définition et

$$\|\pi_{\mathcal{G}} X\|_1 \leqslant \|\pi_{\mathcal{G}} X\|_2 \leqslant \|X\|_2 < \infty.$$

On conclut que $\pi_{\mathcal{G}} X$ est bien une version de $\mathbb{E}[X\,|\,\mathcal{G}]$.

▷ *Existence lorsque* $X \in \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P})$.

Afin de passer du résultat pour $X \in \mathscr{L}^2$ à celui pour $X \in \mathscr{L}^1$, nous allons utiliser un argument classique de troncature. Supposons que $X \in \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ et $X \geqslant 0$. Pour tout $n \geqslant 1$, on définit $X_n \coloneqq X \wedge n$. Alors, $X_n \in \mathscr{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ et le point précédent implique qu'il existe une version Z_n de l'espérance conditionnelle $\mathbb{E}[X_n \,|\, \mathcal{G}]$.

Nous verrons plus bas que

$$\text{presque sûrement,} \quad Z_n \geqslant 0 \text{ et } Z_n \leqslant Z_{n+1} \text{ pour tout } n \geqslant 1. \tag{9.4}$$

On peut donc considérer la variable aléatoire $Z \coloneqq \lim_{n\to\infty} Z_n$. Alors, Z est $\mathcal{G}$-mesurable, $Z_n \uparrow Z$ et donc

$$\mathbb{E}[Z\mathbf{1}_G] = \int_G \lim_{n\to\infty} Z_n \,\mathrm{d}\mathbb{P} = \lim_{n\to\infty} \int_G Z_n \,\mathrm{d}\mathbb{P} = \lim_{n\to\infty} \int_G X_n \,\mathrm{d}\mathbb{P} = \int_G \lim_{n\to\infty} X_n \,\mathrm{d}\mathbb{P} = \mathbb{E}[X\mathbf{1}_G],$$

pour tout $G \in \mathcal{G}$, par la théorème de convergence monotone. En particulier, $Z \in \mathscr{L}^1(\Omega, \mathcal{G}, \mathbb{P})$ (prendre $G = \Omega$). On en conclut que Z est une version de $\mathbb{E}[X \,|\, \mathcal{G}]$.

Pour lever la restriction $X \geqslant 0$, on considère la décomposition habituelle $X = X^+ - X^-$. Alors, l'analyse ci-dessus garantit l'existence de $\mathbb{E}[X^+ \,|\, \mathcal{G}]$ et $\mathbb{E}[X^- \,|\, \mathcal{G}]$. En particulier,

$$\mathbb{E}\big[(\mathbb{E}[X^+ \,|\, \mathcal{G}] - \mathbb{E}[X^- \,|\, \mathcal{G}])\mathbf{1}_G\big] = \mathbb{E}\big[\mathbb{E}[X^+ \,|\, \mathcal{G}]\mathbf{1}_G\big] - \mathbb{E}\big[\mathbb{E}[X^- \,|\, \mathcal{G}]\mathbf{1}_G\big] = \mathbb{E}[X^+\mathbf{1}_G] - \mathbb{E}[X^-\mathbf{1}_\mathcal{G}] = \mathbb{E}[X\mathbf{1}_G]$$

et on en conclut que $\mathbb{E}[X^+ \,|\, \mathcal{G}] - \mathbb{E}[X^- \,|\, \mathcal{G}]$ fournit une version de $\mathbb{E}[X \,|\, \mathcal{G}]$.

▷ *Preuve de* (9.4). On observe pour commencer que $Z_{n+1} - Z_n = \mathbb{E}[X_{n+1} - X_n \,|\, \mathcal{G}]$ p.s., puisque

$$\mathbb{E}[(Z_{n+1} - Z_n)\mathbf{1}_G] = \mathbb{E}[Z_{n+1}\mathbf{1}_G] - \mathbb{E}[Z_n\mathbf{1}_G] = \mathbb{E}[X_{n+1}\mathbf{1}_G] - \mathbb{E}[X_n\mathbf{1}_G] = \mathbb{E}[(X_{n+1} - X_n)\mathbf{1}_G],$$

pour tout $G \in \mathcal{G}$. Comme $X_{n+1} - X_n \geqslant 0$ et $X_n \geqslant 0$, il suffit de montrer que, pour toute variable aléatoire bornée V telle que $V \geqslant 0$, on a $\mathbb{E}[V \,|\, \mathcal{G}] \geqslant 0$ p.s..

Soit W une version de $\mathbb{E}[V \,|\, \mathcal{G}]$ et supposons que $\mathbb{P}(W < 0) > 0$. Alors, il existe $k \geqslant 1$ tel que $G_k \coloneqq \{W < -\frac{1}{k}\} \in \mathcal{G}$ satisfasse $\mathbb{P}(G_k) > 0$. Mais ceci implique

$$0 \leqslant \mathbb{E}[V\mathbf{1}_{G_k}] = \mathbb{E}[W\mathbf{1}_{G_k}] \leqslant -\tfrac{1}{k}\mathbb{P}(G_k) < 0,$$

ce qui est absurde. □

Remarque 9.3. *Il suit de la preuve que, lorsque* $X \in \mathscr{L}^2(\Omega, \mathcal{F}, \mathbb{P})$, $\mathbb{E}[X \,|\, \mathcal{G}]$ *minimise* $\|X - Y\|_2$ *parmi toutes les fonctions* $Y \in \mathscr{L}^2(\Omega, \mathcal{G}, \mathbb{P})$. *En ce sens, l'espérance conditionnelle fournit la meilleure approximation de* X *dans* $\mathscr{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ *au sens des moindres carrés.*

9.2 Propriétés

9.2.1 Propriétés de base

Lemme 9.4. *Soit* $X \in \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ *et* $\mathcal{G} \subset \mathcal{F}$ *une sous-tribu. Alors :*

1. *Si* X *est* $\mathcal{G}$*-mesurable, alors* $\mathbb{E}[X \,|\, \mathcal{G}] = X$ *presque sûrement.*
2. $X \mapsto \mathbb{E}[X \,|\, \mathcal{G}]$ *est linéaire.*
3. *Si* $X \geqslant 0$ *presque sûrement, alors* $\mathbb{E}[X \,|\, \mathcal{G}] \geqslant 0$ *presque sûrement.*
4. $\mathbb{E}[\mathbb{E}[X \,|\, \mathcal{G}]] = \mathbb{E}[X]$.
5. *Si* Y *est* $\mathcal{G}$*-mesurable et bornée, alors* $\mathbb{E}[XY \,|\, \mathcal{G}] = Y\mathbb{E}[X \,|\, \mathcal{G}]$ *presque sûrement.*
   L'hypothèse que Y *est bornée peut être remplacée par* $X \in \mathscr{L}^p$ *et* $Y \in \mathscr{L}^q$ *avec* $p > 1$ *et* $\frac{1}{p} + \frac{1}{q} = 1$.

Démonstration. Laissée en exercice. □

Lemme 9.5. *Soit $\mathcal{G} \subset \mathcal{F}$ une sous-tribu et $A \in \mathcal{F}$. Supposons qu'il existe $\epsilon < 1$ tel que $\mathbb{E}[\mathbf{1}_A \,|\, \mathcal{G}] \leqslant \epsilon$. Alors, $\mathbb{P}(A \,|\, B) \leqslant \epsilon$, pour tout $B \in \mathcal{G}$ tel que $\mathbb{P}(B) > 0$.*

Démonstration. Soit $B \in \mathcal{G}$ tel que $\mathbb{P}(B) > 0$. Par définition de la probabilité conditionnelle,

$$\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{E}[\mathbf{1}_A \mathbf{1}_B]}{\mathbb{E}[\mathbf{1}_B]} = \frac{\mathbb{E}\big[\mathbb{E}[\mathbf{1}_A \,|\, \mathcal{G}]\mathbf{1}_B\big]}{\mathbb{E}[\mathbf{1}_B]} \leqslant \frac{\mathbb{E}[\epsilon \mathbf{1}_B]}{\mathbb{E}[\mathbf{1}_B]} = \epsilon. \qquad \square$$

Lemme 9.6. *Soit $X \in \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ et $\mathcal{G} \subset \mathcal{F}$ une sous-tribu. Alors :*

1. *(Convergence monotone conditionnelle) Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires positives dans $\mathscr{L}^1$ telles que $X_n \uparrow X$. Alors, $\lim_{n\to\infty} \mathbb{E}[X_n \,|\, \mathcal{G}] = \mathbb{E}[X \,|\, \mathcal{G}]$ presque sûrement.*
2. *(Fatou conditionnel) Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires positives dans $\mathscr{L}^1$. Alors, $\mathbb{E}[\liminf_{n\to\infty} X_n \,|\, \mathcal{G}] \leqslant \liminf_{n\to\infty} \mathbb{E}[X_n \,|\, \mathcal{G}]$ presque sûrement.*
3. *(Convergence dominée conditionnelle) Soit $(X_n)_{n\geqslant 1}$ et $Y \in \mathscr{L}^1$ tels que $|X_n| \leqslant Y$ pour tout $n \geqslant 1$. Supposons que $X_n \to X$ presque sûrement. Alors, $\lim_{n\to\infty} \mathbb{E}[X_n \,|\, \mathcal{G}] = \mathbb{E}[X \,|\, \mathcal{G}]$ p.s. et dans $\mathscr{L}^1$.*

Démonstration. 1. Soit Z_n une version de $\mathbb{E}[X_n \,|\, \mathcal{G}]$. Par le point 3 du Lemme 9.4 et l'hypothèse que $X_{n+1} - X_n \geqslant 0$, on déduit que $Z_{n+1} - Z_n \geqslant 0$ presque sûrement. Posons $Z := \lim_{n\to\infty} Z_n$. Alors, Z est $\mathcal{G}$-mesurable et $Z_n \uparrow Z$ presque sûrement. On a donc, pour tout $G \in \mathcal{G}$,

$$\mathbb{E}[Z\mathbf{1}_G] = \lim_{n\to\infty} \mathbb{E}[Z_n \mathbf{1}_G] = \lim_{n\to\infty} \mathbb{E}[X_n \mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G],$$

où l'on a utilisé le théorème de convergence monotone pour la première et la troisième identités, et la définition de Z_n pour la deuxième. En particulier, $Z \in \mathscr{L}^1(\Omega, \mathcal{G}, \mathbb{P})$. On conclut que Z est une version de $\mathbb{E}[X \,|\, \mathcal{G}]$.

2. et *3.* sont laissés en exercices (leur dérivation à partir du point précédent se fait de la même façon que pour les versions non conditionnelles). $\square$

L'inégalité de Jensen s'étend également à l'espérance conditionnelle.

Lemme 9.7. *Soit $X \in \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{G} \subset \mathcal{F}$ une sous-tribu et φ une fonction convexe telle que $\varphi(X) \in \mathscr{L}^1$. Alors, presque sûrement,*

$$\mathbb{E}[\varphi(X) \,|\, \mathcal{G}] \geqslant \varphi(\mathbb{E}[X \,|\, \mathcal{G}]).$$

Démonstration. Admis. $\square$

En particulier, il suit du résultat précédent que, pour tout $p \geqslant 1$,

$$\|\mathbb{E}[X \,|\, \mathcal{G}]\|_p = \mathbb{E}[|\mathbb{E}[X \,|\, \mathcal{G}]|^p]^{1/p} \leqslant \mathbb{E}[\mathbb{E}[|X|^p \,|\, \mathcal{G}]]^{1/p} = \mathbb{E}[|X|^p]^{1/p} = \|X\|_p,$$

puisque $x \mapsto |x|^p$ est convexe. L'avant-dernière identité suit du point 4 du Lemme 9.4.

Lemme 9.8. *Soit $X \in \mathscr{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ et $\mathcal{G} \subset \mathcal{F}$ une sous-tribu. Alors :*

1. *Si $\mathcal{H} \subset \mathcal{G}$ est une sous-tribu, alors $\mathbb{E}\big[\mathbb{E}[X \,|\, \mathcal{G}] \,|\, \mathcal{H}\big] = \mathbb{E}[X \,|\, \mathcal{H}]$ presque sûrement.*
2. *Supposons X indépendant de $\mathcal{G}$. Alors, $\mathbb{E}[X \,|\, \mathcal{G}] = \mathbb{E}[X]$ presque sûrement.*

Démonstration. 1. Il suffit d'observer que, pour tout $H \in \mathcal{H} \subset \mathcal{G}$,

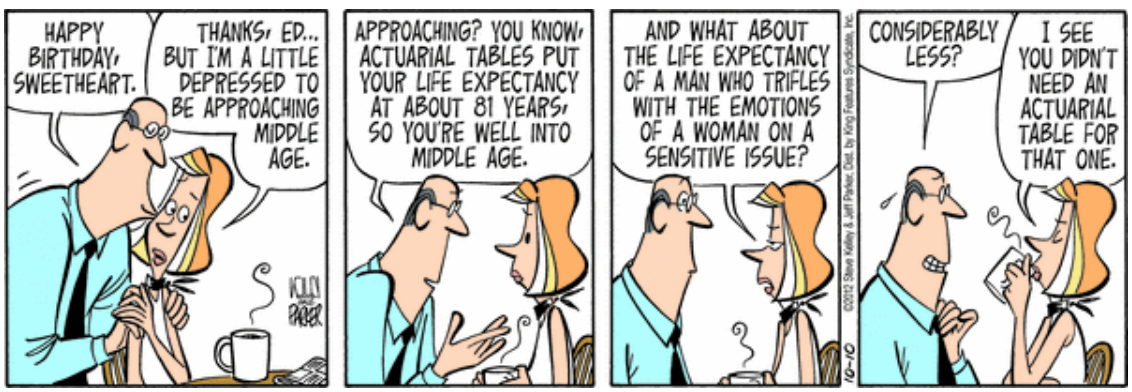
$$\mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[X \,|\, \mathcal{G}] \,|\, \mathcal{H}\big]\mathbf{1}_H\Big] = \mathbb{E}\big[\mathbf{1}_H \mathbb{E}[X \,|\, \mathcal{G}]\big] = \mathbb{E}\big[\mathbf{1}_H X\big],$$

où l'on a utilisé le fait que $H \in \mathcal{H}$ pour la première égalité, et le fait que $H \in \mathcal{G}$ pour la seconde. $\mathbb{E}\big[\mathbb{E}[X \,|\, \mathcal{G}] \,|\, \mathcal{H}\big]$ est donc bien une version de $\mathbb{E}[X \,|\, \mathcal{H}]$.

2. Évidemment, $\mathbb{E}[X]$ est $\mathcal{G}$-mesurable et appartient à $\mathscr{L}^1$. Il suffit donc d'observer que, pour tout $G \in \mathcal{G}$,

$$\mathbb{E}[\mathbb{E}[X]\mathbf{1}_G] = \mathbb{E}[X]\mathbb{E}[\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G],$$

où la dernière identité suit de l'indépendance de X et $\mathbf{1}_G$. On en conclut que $\mathbb{E}[X]$ est bien une version de $\mathbb{E}[X \mid \mathcal{G}]$. □



10 Martingales

There are two times in a man's life when he should not speculate; when he can't afford it, and when he can.

Mark Twain

Ce chapitre est consacré à une classe particulièrement importante de processus stochastiques : les martingales. Avec les processus markoviens (ceux pour auxquels la propriété de Markov s'applique, par exemple les chaînes de Markov), ils forment aujourd'hui les deux classes les plus importantes de processus en théorie des probabilités.

À l'origine, le terme *martingale* faisait référence à une classe de stratégies de paris qui était populaire dans la France du XVIII^{e} siècle. L'exemple le plus célèbre s'applique à un jeu dans lequel le joueur gagne sa mise si une pièce tombe sur « face » et la perd si la pièce tombe sur « pile » et consiste, pour le joueur, à doubler sa mise après chaque perte. Ainsi, lors de son premier gain, il récupère toutes les pertes précédentes et réalise un bénéfice égal à la mise initiale.

Bien que l'on puisse déjà reconnaître une martingale dans le traitement par Pascal du célèbre *problème des partis* en 1654, le concept abstrait de martingale a été introduit en théorie des probabilités dans les années 1930. Les principaux artisans du développement initial de la théorie des martingales sont Bernstein [1], Lévy, Ville [2], Borel et Doob [3].

En théorie des probabilités, les martingales jouent un rôle essentiel dans la théorie des processus et, en particulier, dans le développement du calcul stochastique. Elles sont également très utiles dans de nombreux domaines plus appliqués, par exemple en mathématiques financières, où les prix des actifs financiers sont souvent modélisés comme des martingales.

Afin de motiver la définition des martingales, retournons à un exemple de processus appartenant à cette classe que l'on a déjà rencontré : la marche aléatoire simple symétrique sur $\mathbb{Z}$.

Soit donc $(X_k)_{k\geqslant 1}$ i.i.d. telles que $\mathbb{P}(X_k = 1) = p = 1 - \mathbb{P}(X_k = -1)$ et soit $S_n \coloneqq X_1 + \cdots + X_n$ la marche aléatoire simple sur $\mathbb{Z}$ de paramètre p. Évidemment, $|S_n| \leqslant n$ et par conséquent

$$\forall n \geqslant 1, \qquad S_n \in \mathscr{L}^1. \tag{10.1}$$

De plus, si $\mathcal{F}_n \coloneqq \sigma(S_0, \ldots, S_n) = \sigma(X_1, \ldots, X_n)$, alors

$$\mathbb{E}[S_{n+1} \,|\, \mathcal{F}_n] = \mathbb{E}[S_n + X_{n+1} \,|\, \mathcal{F}_n] = \mathbb{E}[S_n \,|\, \mathcal{F}_n] + \mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] = S_n + \mathbb{E}[X_{n+1}] = S_n + (2p - 1),$$

1. Sergueï Natanovitch Bernstein (1880, Odessa – 1968, Moscou), mathématicien soviétique.
2. Jean Ville (1910, Marseille – 1989, Blois), mathématicien français.
3. Joseph Leo Doob (1910, Cincinnati – 2004, Clark-Lindsey Village), mathématicien américain.

presque sûrement. En particulier, dans le cas symétrique ($p = \frac{1}{2}$),

$$\mathbb{E}[S_{n+1} \,|\, \mathcal{F}_n] = S_n, \quad \text{presque sûrement.} \tag{10.2}$$

Les propriétés (10.1) et (10.2) caractérisent les martingales. Nous verrons dans ce chapitre que ces dernières possèdent un grand nombre de propriétés remarquables qui en font un outil crucial pour le probabiliste.

10.1 Définition et exemples

Définition 10.1. *Soit $(\Omega, \mathcal{F})$ un espace mesurable. Une **filtration** est une suite croissante $(\mathcal{F}_n)_{n\geqslant 0}$ de sous-tribus de $\mathcal{F}$: $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}$.*

*Une suite de variables aléatoires $(X_n)_{n\geqslant 0}$ est **adaptée** à la filtration $(\mathcal{F}_n)_{n\geqslant 0}$ si X_n est $\mathcal{F}_n$-mesurable pour tout $n \geqslant 0$.*

*La filtration $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$ est appelée la **filtration canonique** ou **naturelle** associée à $(X_n)_{n\geqslant 0}$.*

Intuitivement, $\mathcal{F}_n$ représente l'information disponible au temps n.

Définition 10.2. *Un processus stochastique $(X_n)_{n\geqslant 0}$ est une **martingale** relativement à la filtration $(\mathcal{F}_n)_{n\geqslant 0}$ si*

1. *$(X_n)_{n\geqslant 0}$ est adapté à $(\mathcal{F}_n)_{n\geqslant 0}$.*
2. *$X_n \in \mathscr{L}^1$ pour tout $n \geqslant 0$.*
3. *$\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] = X_n$ presque sûrement, pour tout $n \geqslant 0$.*

*Si la 3e condition est remplacée par $\forall n \geqslant 0$, $\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] \leqslant X_n$ p.s., $(X_n)_{n\geqslant 0}$ est une **sur-martingale**.*
*Si la 3e condition est remplacée par $\forall n \geqslant 0$, $\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] \geqslant X_n$ p.s., $(X_n)_{n\geqslant 0}$ est une **sous-martingale**.*

Exemple 10.3. Relativement à la filtration naturelle, la marche aléatoire du début du chapitre est une sous-martingale lorsque $p > \frac{1}{2}$, une martingale lorsque $p = \frac{1}{2}$ et une sur-martingale lorsque $p < \frac{1}{2}$. ⌟

Exemple 10.4. On modélise la fortune d'un joueur au cours d'un jeu de hasard à l'aide d'un processus stochastique $(X_n)_{n\geqslant 0}$. Le jeu est équitable si, même en utilisant toute l'information disponible après les manches précédentes, le gain net (le gain moins la mise) du joueur est nul en moyenne : presque sûrement, $\mathbb{E}[X_{n+1} - X_n \,|\, \mathcal{F}_n] = 0$ ou, ce qui est équivalent, $\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] = X_n$; en d'autres termes, si le processus est une martingale. De même, le jeu est défavorable au joueur si son gain net est strictement négatif en moyenne, ce qui donne lieu à une sur-martingale. ⌟

Remarque 10.5.
- ▷ *$(X_n)_{n\geqslant 0}$ est une sur-martingale si et seulement si $(-X_n)_{n\geqslant 0}$ est une sous-martingale.*
- ▷ *$(X_n)_{n\geqslant 0}$ est une martingale si et seulement si $(X_n)_{n\geqslant 0}$ est à la fois une sous-martingale et une sur-martingale.*
- ▷ *$(X_n)_{n\geqslant 0}$ est une martingale/sur-martingale/sous-martingale si et seulement si $X_0 \in \mathscr{L}^1$ et $(X_n - X_0)_{n\geqslant 0}$ est une martingale/sur-martingale/sous-martingale. En particulier, il suffira de considérer le cas $X_0 = 0$.*

L'observation suivante est importante. Soit $(X_n)_{n\geqslant 0}$ une martingale relativement à la filtration $(\mathcal{F}_n)_{n\geqslant 0}$ et $m > n$. Dans ce cas, $\mathcal{F}_n \subset \mathcal{F}_{m-1}$ et il suit donc du premier point du Lemme 9.8 que, presque sûrement,

$$\mathbb{E}[X_m \,|\, \mathcal{F}_n] = \mathbb{E}\big[\mathbb{E}[X_m \,|\, \mathcal{F}_{m-1}] \;\big|\; \mathcal{F}_n\big] = \mathbb{E}[X_{m-1} \,|\, \mathcal{F}_n] = \cdots = \mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] = X_n.$$

De plus, en prenant l'espérance dans les deux membres de cette identité lorsque $n = 0$, on obtient

$$\forall m \geqslant 0, \qquad \mathbb{E}[X_m] = \mathbb{E}[X_0].$$

En appliquant le même argument aux sur-/sous- martingales, on conclut que, presque sûrement,

$$\mathbb{E}[X_m \,|\, \mathcal{F}_n] \begin{cases} \leqslant X_n & \text{si } (X_n)_{n\geqslant 0} \text{ est une sur-martingale,} \\ \geqslant X_n & \text{si } (X_n)_{n\geqslant 0} \text{ est une sous-martingale.} \end{cases}$$

En particulier,

$$\forall m > n \geqslant 0, \qquad \mathbb{E}[X_m] \begin{cases} \leqslant \mathbb{E}[X_n] & \text{si } (X_k)_{k\geqslant 0} \text{ est une sur-martingale,} \\ \geqslant \mathbb{E}[X_n] & \text{si } (X_k)_{k\geqslant 0} \text{ est une sous-martingale.} \end{cases}$$

Exemple 10.6. Soit $(X_n)_{n\geqslant 1}$ des variables aléatoires i.i.d. telles que $X_1 \geqslant 0$ et $\mathbb{E}[X_1] = 1$. Alors, le processus stochastique $(M_n)_{n\geqslant 0}$ défini par $M_0 \coloneqq 1$ et

$$\forall n \geqslant 1, \qquad M_n \coloneqq X_1 X_2 \cdots X_n$$

est une martingale relativement à la filtration $(\mathcal{F}_n)_{n\geqslant 1}$, où $\mathcal{F}_n \coloneqq \sigma(X_1, \ldots, X_n)$. En effet,

- ▷ $\mathbb{E}[M_n] = \mathbb{E}[X_1]^n = 1$, ce qui implique que $M_n \in \mathscr{L}^1$ pour tout $n \geqslant 1$.
- ▷ M_n est $\mathcal{F}_n$-mesurable par le Lemme 2.31.
- ▷ $\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[M_n X_{n+1} \mid \mathcal{F}_n] = M_n \mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = M_n \mathbb{E}[X_{n+1}] = M_n$ presque sûrement. ⌟

Exemple 10.7. Soit $(\mathcal{F}_n)_{n\geqslant 0}$ une filtration et $X \in \mathscr{L}^1$. On considère le processus $X_n \coloneqq \mathbb{E}[X \mid \mathcal{F}_n]$. Alors $(X_n)_{n\geqslant 0}$ est une martingale relativement à $(\mathcal{F}_n)_{n\geqslant 0}$. En effet,

- ▷ X_n est $\mathcal{F}_n$-mesurable.
- ▷ $X_n \in \mathscr{L}^1$ pour tout $n \geqslant 0$.
- ▷ $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{F}_{n+1}] \mid \mathcal{F}_n] = \mathbb{E}[X \mid \mathcal{F}_n] = X_n$ p.s., par le point 1 du Lemme 9.8.

Une martingale construite de cette façon est appelée une **martingale fermée**. ⌟

Exemple 10.8. Soit $(X_k)_{k\geqslant 0}$ une chaîne de Markov à valeurs dans un espace des états S fini et de matrice de transition P. Soit $f : S \to \mathbb{R}$ une fonction P-harmonique (se rappeler la Définition 7.19). Alors $(f(X_n)_{n\geqslant 0})$ est une martingale relativement à la filtration $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$. En effet,

$$\mathbb{E}[f(X_{n+1}) \mid \mathcal{F}_n] = \mathbb{E}[f(X_{n+1}) \mid X_0, \ldots, X_n] = \mathbb{E}[f(X_{n+1}) \mid X_n] = \sum_{j\in S} f(j) p(X_n, j) = f(X_n),$$

où l'on a utilisé la propriété de Markov pour la deuxième égalité et la P-harmonicité de f pour la dernière. ⌟

Définition 10.9. *Soit $p \geqslant 1$ et $(X_k)_{k\geqslant 0}$ une martingale.*
▷ *$(X_k)_{k\geqslant 0}$ est une **$\mathscr{L}^p$-martingale** si $\|X_n\|_p < \infty$ pour tout $n \geqslant 0$.*
▷ *$(X_k)_{k\geqslant 0}$ est **bornée dans $\mathscr{L}^p$** si $\sup_{n\geqslant 0} \|X_n\|_p < \infty$.*

10.2 Transformées de martingales

Étant donné une martingale $(X_n)_{n\geqslant 0}$ et une fonction mesurable $\varphi : \mathbb{R} \to \mathbb{R}$, que peut-on dire de $(\varphi(X_n))_{n\geqslant 0}$?

Lemme 10.10. *Soit $(X_n)_{n\geqslant 0}$ une martingale (resp. une sous-martingale) et $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction convexe (resp. convexe et croissante) telle que $\varphi(X_n) \in \mathscr{L}^1$ pour tout $n \geqslant 0$. Alors, $(\varphi(X_n))_{n\geqslant 0}$ est une sous-martingale.*

Exemple 10.11.
- ▷ Soit $(X_n)_{n\geqslant 0}$ une martingale. Alors $(|X_n|)_{n\geqslant 0}$ est une sous-martingale.
- ▷ Soit $(X_n)_{n\geqslant 0}$ une $\mathscr{L}^2$-martingale. Alors, $(X_n^2)_{n\geqslant 0}$ est une sous-martingale.
- ▷ Soit $(X_n)_{n\geqslant 0}$ une sous-martingale. Alors $(X_n^+)_{n\geqslant 0}$ est une sous-martingale. ⌟

Démonstration. L'affirmation pour les martingales est une conséquence immédiate de l'inégalité de Jensen conditionnelle :

$$\varphi(X_n) = \varphi(\mathbb{E}[X_{n+1} \mid \mathcal{F}_n]) \leqslant \mathbb{E}[\varphi(X_{n+1}) \mid \mathcal{F}_n].$$

L'affirmation pour les sous-martingales se démontre de façon tout à fait similaire :

$$\varphi(X_n) \leqslant \varphi(\mathbb{E}[X_{n+1} \mid \mathcal{F}_n]) \leqslant \mathbb{E}[\varphi(X_{n+1}) \mid \mathcal{F}_n],$$

où la première inégalité suit de la croissance de φ. □

Définition 10.12. *Soit $(\mathcal{F}_n)_{n\geqslant 0}$ une filtration. Un processus $(C_n)_{n\geqslant 1}$ est **prévisible** si C_n est $\mathcal{F}_{n-1}$-mesurable pour tout $n \geqslant 1$.*

Intuitivement, un processus est prévisible si les informations disponibles à l'instant n permettent de déterminer son état au temps $n+1$.

Exemple 10.13. On considère un jeu dans lequel le croupier lance à chaque étape une pièce équilibrée, obtenant une suite i.i.d. $(\xi_k)_{k\geqslant 1}$, où $\mathbb{P}(\xi_k = 1) = \mathbb{P}(\xi_k = -1) = \frac{1}{2}$ (les deux valeurs possibles -1 et 1 de la variable aléatoire ξ_k correspondent, respectivement, au résultat « pile » ou « face » obtenu au k-ième lancer).

Le processus $(X_k)_{k\geqslant 0}$, où $X_k := \xi_1 + \cdots + \xi_k$, est une martingale relativement à sa filtration naturelle $(\mathcal{F}_n)_{n\geqslant 0}$. Avant le n-ième lancer, le parieur peut miser un montant C_n sur la sortie de « face » lors de ce lancer ; il double sa mise s'il a raison et la perd s'il a tort. Lors du choix de sa mise, le joueur ne peut se baser que sur les résultats des $n-1$ lancers précédents ; en particulier, sa stratégie de mise $(C_n)_{n\geqslant 1}$ est prévisible relativement à la filtration $(\mathcal{F}_n)_{n\geqslant 0}$.

Le gain (ou la perte selon le signe) du joueur lors du k-ième lancer est donc donné par $C_k\xi_k$ et son gain total au cours des n premiers lancers est

$$\sum_{k=1}^{n} C_k\xi_k = \sum_{k=1}^{n} C_k(X_k - X_{k-1}).$$ ⌟

Définition 10.14. *Soit $(X_k)_{k\geqslant 0}$ une sur-martingale et $(C_k)_{k\geqslant 1}$ un processus prévisible. La **transformée de la martingale** $(X_k)_{k\geqslant 0}$ par $(C_k)_{k\geqslant 1}$ est définie par $(C\bullet X)_0 := 0$ et*[4]

$$\forall n \geqslant 1, \qquad (C\bullet X)_n := \sum_{k=1}^{n} C_k(X_k - X_{k-1}).$$

Le résultat suivant montre qu'il n'existe pas de stratégie (nécessairement prévisible si le joueur ne triche pas) permettant de s'assurer de battre le casino, c'est-à-dire de transformer un jeu qui n'est pas favorable au joueur en un jeu favorable, s'il existe une limite sur le montant des mises autorisées...

Lemme 10.15. *Soit $(C_k)_{k\geqslant 1}$ un processus prévisible et localement borné, c'est-à-dire tel que, pour chaque $n \geqslant 1$, $\sup_\omega |C_n(\omega)| < \infty$. Alors :*

1. *Si $(C_k)_{k\geqslant 1}$ est positif et $(X_k)_{k\geqslant 0}$ est une sur-martingale, alors $C\bullet X$ est une sur-martingale.*
2. *Si $(X_k)_{k\geqslant 0}$ est une martingale, alors $C\bullet X$ est une martingale.*

On peut remplacer l'hypothèse que $(C_k)_{k\geqslant 1}$ est borné par $C_k \in \mathscr{L}^2$ pour tout $k \geqslant 1$, pourvu que $(X_k)_{k\geqslant 0}$ soit une $\mathscr{L}^2$-martingale.

4. On peut considérer cette quantité comme l'analogue discret de l'intégrale stochastique $\int C\,\mathrm{d}X$ que vous verrez peut-être dans certains cours de master.

Démonstration. Notons $Y_n \coloneqq (C \bullet X)_n$. C_n étant borné pour chaque $n \geqslant 1$, on a $Y_n \in \mathscr{L}^1$ pour tout $n \geqslant 0$. De plus,

$$\begin{aligned}\mathbb{E}[Y_{n+1} \,|\, \mathcal{F}_n] &= \mathbb{E}[Y_n \,|\, \mathcal{F}_n] + \mathbb{E}[Y_{n+1} - Y_n \,|\, \mathcal{F}_n] \\ &= Y_n + \mathbb{E}[C_{n+1}(X_{n+1} - X_n) \,|\, \mathcal{F}_n] = Y_n + C_{n+1}\mathbb{E}[X_{n+1} - X_n \,|\, \mathcal{F}_n].\end{aligned}$$

Le dernier terme du membre de droite est négatif ou nul si $(X_k)_{k\geqslant 0}$ est une sur-martingale ; il est nul si $(X_k)_{k\geqslant 0}$ est une martingale. Ceci démontre les deux points. Les hypothèses sur $(C_k)_{k\geqslant 1}$ sont utilisées pour justifier la troisième égalité (par le point 5 du Lemme 9.4). □

10.3 Temps d'arrêt

On a vu précédemment que si $(X_n)_{n\geqslant 0}$ est une martingale, alors $\mathbb{E}[X_n] = \mathbb{E}[X_0]$ pour tout $n \geqslant 0$. Il se révèle très utile de déterminer sous quelles conditions une égalité de ce type reste valide pour un temps *aléatoire*.

Définition 10.16. *Soit $(\mathcal{F}_n)_{n\geqslant 0}$ une filtration. Une variable aléatoire T à valeurs dans $\overline{\mathbb{N}}$ est un **temps d'arrêt** relativement à cette filtration si*

$$\forall n \in \overline{\mathbb{N}}, \qquad \{T = n\} \in \mathcal{F}_n,$$

où l'on a noté $\mathcal{F}_\infty \coloneqq \sigma(\bigcup_{n\in\mathbb{N}} \mathcal{F}_n) \subset \mathcal{F}$.

En restant sur l'interprétation en termes de jeu, un temps d'arrêt peut être vu comme un algorithme à l'aide duquel le joueur décide d'arrêter la partie. Il ne peut le faire que sur la base de l'information acquise au cours du jeu, sans « voir dans le futur ». Ainsi, à la fin du n-ième tour, il détermine si $T = n$ sur la base des n tours déjà effectués (possible puisque $\{T = n\} \in \mathcal{F}_n$) et arrête la partie si c'est le cas.

Lemme 10.17. *T est un temps d'arrêt si et seulement si $\forall n \in \overline{\mathbb{N}}$, $\{T \leqslant n\} \in \mathcal{F}_n$.*

Démonstration. Il suffit d'utiliser les identités

$$\{T = n\} = \{T \leqslant n\} \setminus \{T \leqslant n-1\} \qquad \text{et} \qquad \{T \leqslant n\} = \bigcup_{k=0}^{n} \{T = k\}.$$ □

Exemple 10.18. Soit $(X_n)_{n\geqslant 0}$ un processus à valeurs dans $(E, \mathcal{E})$, adapté à la filtration $(\mathcal{F}_n)_{n\geqslant 0}$ (par exemple, la marche aléatoire simple et sa filtration naturelle).

Soit $A \in \mathcal{E}$. La variable aléatoire $T \coloneqq \inf\{n \geqslant 0 \,|\, X_n \in A\}$, correspondant au temps de la première visite en A, est un temps d'arrêt. En effet,

$$\{T \leqslant n\} = \bigcup_{k\leqslant n} \underbrace{\{X_k \in A\}}_{\in \mathcal{F}_k \subset \mathcal{F}_n} \in \mathcal{F}_n.$$

Par contre, la variable aléatoire $L \coloneqq \sup\{n \geqslant 0 \,|\, X_n \in A\}$, correspondant au temps de la dernière visite en A, n'est généralement pas un temps d'arrêt (exercice). ⌟

Lemme 10.19. *Si T_1 et T_2 sont des temps d'arrêt, alors $T_1 \vee T_2$, $T_1 \wedge T_2$, $T_1 + T_2$ sont des temps d'arrêt.*

Démonstration. Laissée en exercice. □

10.4 Théorèmes d'arrêt et applications

Soit $(X_n)_{n\geqslant 0}$ une (sur-)martingale et T un temps d'arrêt. Supposons qu'un joueur mise toujours une unité et cesse de jouer après le temps T. Le processus décrivant ses mises est donc $(C_n^T)_{n\geqslant 1}$, où $C_n^T := \mathbf{1}_{\{n\leqslant T\}}$. Par conséquent, le processus décrivant ses gains est

$$Y_n := (C^T \bullet X)_n = \sum_{k=1}^{n} C_k^T (X_k - X_{k-1}) = \sum_{k=1}^{n\wedge T} (X_k - X_{k-1}) = X_{n\wedge T} - X_0.$$

Définition 10.20. *Soit $(X_n)_{n\geqslant 0}$ un processus adapté à la filtration $(\mathcal{F}_n)_{n\geqslant 0}$ et T un temps d'arrêt. Le processus $(X_{n\wedge T})_{n\geqslant 0}$ est appelé **processus arrêté au temps T**.*

Le processus $(C_n^T)_{n\geqslant 1}$ est positif, borné et prévisible. En effet, pour tout $n \geqslant 1$,

$$\{C_n^T = 0\} = \{T \leqslant n-1\} \in \mathcal{F}_{n-1} \qquad \text{et} \qquad \{C_n^T = 1\} = \{C_n^T = 0\}^{\mathrm{c}} \in \mathcal{F}_{n-1}.$$

Il suit donc du Lemme 10.15 que $(Y_n)_{n\geqslant 0}$ est une (sur-)martingale. Par conséquent, on a démontré le résultat suivant.

Lemme 10.21. *Si $(X_n)_{n\geqslant 0}$ est une sur-martingale et T un temps d'arrêt, alors $(X_{n\wedge T})_{n\geqslant 0}$ est une sur-martingale ; en particulier, $\mathbb{E}[X_{n\wedge T}] \leqslant \mathbb{E}[X_0]$ pour tout $n \geqslant 0$.*

Si $(X_n)_{n\geqslant 0}$ est une martingale et T un temps d'arrêt, alors $(X_{n\wedge T})_{n\geqslant 0}$ est une martingale ; en particulier, $\mathbb{E}[X_{n\wedge T}] = \mathbb{E}[X_0]$ pour tout $n \geqslant 0$.

Remarque 10.22. *Il est important d'observer que si $(X_n)_{n\geqslant 0}$ est une martingale et T un temps d'arrêt, le lemme précédent n'implique pas que $\mathbb{E}[X_T] = \mathbb{E}[X_0]$, même lorsque $\mathbb{P}(T < \infty) = 1$!*

Considérons, par exemple, la marche aléatoire simple symétrique $(S_n)_{n\geqslant 0}$ sur $\mathbb{Z}$ partant de 0 et le temps d'arrêt $T := \inf\{n \geqslant 0 \,|\, S_n = 1\}$ (presque sûrement fini par le Théorème 3.7). Alors, $1 = \mathbb{E}[S_T] \neq \mathbb{E}[S_0] = 0$, malgré le fait que $\mathbb{E}[S_{T\wedge n}] = \mathbb{E}[S_0]$ pour tout $n \geqslant 0$ par le lemme précédent.

Théorème 10.23 (Théorème d'arrêt de Doob)**.** *Soit $(X_n)_{n\geqslant 0}$ un processus adapté à la filtration $(\mathcal{F}_n)_{n\geqslant 0}$ et T un temps d'arrêt. On considère les conditions suivantes :*

(i) T est borné : $\exists K, \forall \omega,\ T(\omega) \leqslant K$.

(ii) $T < \infty$ presque sûrement et $(X_n)_{n\geqslant 0}$ est borné : $\exists K, \forall n \geqslant 0, \forall \omega,\ |X_n(\omega)| \leqslant K$.

(iii) $T \in \mathscr{L}^1$ et $(X_n)_{n\geqslant 0}$ est à accroissements bornés : $\exists K, \forall n \geqslant 0, \forall \omega,\ |X_{n+1}(\omega) - X_n(\omega)| \leqslant K$.

(iv) $T < \infty$ presque sûrement et $\forall n \geqslant 0,\ X_n \geqslant 0$.

Alors,

1. *Si $(X_n)_{n\geqslant 0}$ est une sur-martingale et une des conditions (i) à (iv) est satisfaite, alors $X_T \in \mathscr{L}^1$ et*

$$\mathbb{E}[X_T] \leqslant \mathbb{E}[X_0].$$

2. *Si $(X_n)_{n\geqslant 0}$ est une martingale et une des conditions (i) à (iii) est satisfaite, alors $X_T \in \mathscr{L}^1$ et*

$$\mathbb{E}[X_T] = \mathbb{E}[X_0].$$

Démonstration. *1.* Par le Lemme 10.21, $X_{T\wedge n} \in \mathscr{L}^1$ et $\mathbb{E}[X_{T\wedge n}] \leqslant \mathbb{E}[X_0]$ pour tout $n \geqslant 0$. Considérons les différentes hypothèses :

(i) En prenant $n = K$, $X_T = X_{T\wedge K} \in \mathscr{L}^1$ et $\mathbb{E}[X_T] = \mathbb{E}[X_{T\wedge K}] \leqslant \mathbb{E}[X_0]$.

(ii) D'une part, $|X_{T\wedge n}(\omega)| \leqslant K$ pour tout n et ω. D'autre part, $\lim_{n\to\infty} X_{T\wedge n} = X_T$ presque sûrement (puisque $T < \infty$ presque sûrement). On a donc $\mathbb{E}[|X_T|] = \lim_{n\to\infty} \mathbb{E}[|X_{T\wedge n}|] < \infty$ et $\mathbb{E}[X_T] = \lim_{n\to\infty} \mathbb{E}[X_{T\wedge n}] \leqslant \mathbb{E}[X_0]$, par le théorème de convergence dominée.

(iii) D'une part, $|X_{T\wedge n}-X_0| = |\sum_{k=1}^{T\wedge n}(X_k-X_{k-1})| \leqslant \sum_{k=1}^{T\wedge n}|X_k-X_{k-1}| \leqslant KT$. D'autre part, $\mathbb{E}[T]<\infty$ et $\lim_{n\to\infty} X_{T\wedge n} = X_T$ presque sûrement. Ainsi, $|X_{T\wedge n}| \leqslant |X_{T\wedge n}-X_0|+|X_0| \leqslant KT+|X_0| \in \mathscr{L}^1$. Il suit donc du théorème de convergence dominée que $\mathbb{E}[|X_T|] = \lim_{n\to\infty}\mathbb{E}[|X_{T\wedge n}|] \leqslant K\mathbb{E}[T] + \mathbb{E}[|X_0|] < \infty$ et $\mathbb{E}[X_T] = \lim_{n\to\infty}\mathbb{E}[X_{T\wedge n}] \leqslant \mathbb{E}[X_0]$.

(iv) À nouveau, $\lim_{n\to\infty} X_{T\wedge n} = X_T$ presque sûrement. On a donc $\mathbb{E}[X_T] = \mathbb{E}[\liminf_{n\to\infty} X_{T\wedge n}] \leqslant \liminf_{n\to\infty}\mathbb{E}[X_{T\wedge n}] \leqslant \mathbb{E}[X_0]$, par le lemme de Fatou.

2. Comme $(X_n)_{n\geqslant 0}$ et $(-X_n)_{n\geqslant 0}$ sont deux sur-martingales, il suit du point précédent, sous les hypothèses (i), (ii) ou (iii), que $\mathbb{E}[X_T] \leqslant \mathbb{E}[X_0]$ et $\mathbb{E}[-X_T] \leqslant \mathbb{E}[-X_0]$ et l'affirmation suit. □

Afin de pouvoir appliquer le Théorème 10.23, il est clairement utile de disposer d'un outil permettant de vérifier que le temps d'arrêt T est dans $\mathscr{L}^1$ (en particulier, presque sûrement fini).

Lemme 10.24. *Soit T un temps d'arrêt tel qu'il existe $N\in\mathbb{N}^*$ et $\epsilon\in(0,1)$ pour lesquels, presque sûrement,*

$$\forall n\geqslant 0, \qquad \mathbb{E}[\mathbf{1}_{\{T\leqslant n+N\}} \,|\, \mathcal{F}_n] \geqslant \epsilon.$$

Alors, $\mathbb{P}(T>kN) \leqslant (1-\epsilon)^k$ pour tout $k\geqslant 1$. En particulier, $T\in\mathscr{L}^1$.

Démonstration. Par hypothèse, $\mathbb{P}(T\leqslant N) = \mathbb{E}[\mathbf{1}_{\{T\leqslant N\}}] = \mathbb{E}\big[\mathbb{E}[\mathbf{1}_{\{T\leqslant N\}}\,|\,\mathcal{F}_0]\big] \geqslant \epsilon$, ce qui établit le cas $k=1$.

Supposons, par induction, que $0<\mathbb{P}(T>kN)\leqslant(1-\epsilon)^k$ (notez que si $\mathbb{P}(T>kN)=0$, alors $\mathbb{P}(T>(k+1)N)=0$). Par hypothèse,

$$\mathbb{E}[\mathbf{1}_{\{T>(k+1)N\}}\,|\,\mathcal{F}_{kN}] = 1-\mathbb{E}[\mathbf{1}_{\{T\leqslant kN+N\}}\,|\,\mathcal{F}_{kN}] \leqslant 1-\epsilon.$$

En particulier, il suit du Lemme 9.5 que $\mathbb{P}(T>(k+1)N\,|\,T>kN) = \mathbb{E}[\mathbf{1}_{\{T>(k+1)N\}}\,|\,T>kN] \leqslant 1-\epsilon$. Il suit donc de notre hypothèse d'induction que

$$\mathbb{P}(T>(k+1)N) = \mathbb{P}(T>(k+1)N\,|\,T>kN)\mathbb{P}(T>kN) \leqslant (1-\epsilon)\mathbb{P}(T>kN) \leqslant (1-\epsilon)^{k+1}. \qquad \square$$

Exemple 10.25. Soit $(X_k)_{k\geqslant 0}$ une chaîne de Markov absorbante sur S fini, de matrice de transition P. Soit $\tau_{\mathcal{A}}$ le temps d'absorption de la chaîne. On a vu dans l'Exemple 10.8 que $(f(X_k))_{k\geqslant 0}$ est une martingale pour toute fonction P-harmonique f. Par le Lemme 10.24 et la définition des chaînes absorbantes, $\tau_{\mathcal{A}}<\infty$ presque sûrement. Il suit donc du Théorème 10.23 que, pour tout $i\in S$,

$$\mathbb{E}_i[f(X_{\tau_{\mathcal{A}}})] = \mathbb{E}_i[f(X_0)] = f(i)$$

et on retrouve le Théorème 7.20. ⌟

Lemme 10.26. *Soit $(X_k)_{k\geqslant 0}$ une sous-martingale, $m\in\mathbb{N}$ et T un temps d'arrêt tel que $\mathbb{P}(T<m)=1$. Alors,*

$$\mathbb{E}[X_0] \leqslant \mathbb{E}[X_T] \leqslant \mathbb{E}[X_m].$$

Démonstration. La première inégalité est une conséquence immédiate du point (i) du Théorème 10.23, puisque $(-X_k)_{k\geqslant 0}$ est une sur-martingale. Pour la seconde, posons $C_k := \mathbf{1}_{\{T\leqslant k-1\}}$. Le processus $C\bullet X$ est une sous-martingale et $(C\bullet X)_k = (X_k-X_T)\mathbf{1}_{\{T<k\}} = X_k - X_{T\wedge k}$. Par conséquent,

$$\mathbb{E}[X_m]-\mathbb{E}[X_T] = \mathbb{E}[X_m - X_{T\wedge m}] = \mathbb{E}[(C\bullet X)_m] \geqslant \mathbb{E}[(C\bullet X)_0] = 0. \qquad \square$$

10.4.1 Quelques applications

Passons à quelques applications du Théorème 10.23.

Ruine du joueur

On considère la marche aléatoire simple sur $\mathbb{Z}$, $S_n \coloneqq a + \sum_{i=1}^n X_i$, où les variables aléatoires $(X_k)_{k\geqslant 1}$ sont i.i.d. et telles que $\mathbb{P}(X_1 = 1) = p$ et $\mathbb{P}(X_1 = -1) = q \coloneqq 1 - p$. On suppose $0 < a < N$.

Notre but est de revisiter le problème de la ruine du joueur (Section 3.3) en déterminant la probabilité que la marche visite 0 avant de visiter N. Introduisons donc le temps d'arrêt $T \coloneqq \inf\{n \geqslant 0 \,|\, S_n \in \{0, N\}\}$. Comme

$$\forall i \in \{1, \dots, N-1\}, \qquad \mathbb{P}_i(T \leqslant N) \geqslant (p \vee q)^N,$$

le Lemme 10.24 implique que $T \in \mathscr{L}^1$.
▷ Si $p = q = 1/2$, alors $(S_n)_{n\geqslant 0}$ est une martingale et le théorème d'arrêt donne

$$a = \mathbb{E}_a[S_0] = \mathbb{E}_a[S_T] = \mathbb{E}_a[S_T \mathbf{1}_{\{S_T=0\}}] + \mathbb{E}_a[S_T \mathbf{1}_{\{S_T=N\}}] = N\mathbb{P}_a(S_T = N).$$

Par conséquent, on retrouve bien $\mathbb{P}_a(S_T = N) = \frac{a}{N}$, c'est-à-dire $\mathbb{P}(S_T = 0) = 1 - \frac{a}{N}$.
▷ Si $p \neq q$, il nous faut trouver une martingale appropriée. Considérons le processus stochastique $(M_n)_{n\geqslant 0}$, avec $M_n \coloneqq (q/p)^{S_n}$. Il s'agit bien d'une martingale (relativement à la filtration naturelle associée à la marche), puisque

$$\begin{aligned}\mathbb{E}_a[M_{n+1} \,|\, \mathcal{F}_n] = \mathbb{E}_a[(q/p)^{S_{n+1}} \,|\, \mathcal{F}_n] &= \mathbb{E}_a[(q/p)^{S_n}(q/p)^{X_{n+1}} \,|\, \mathcal{F}_n] \\ &= (q/p)^{S_n}\mathbb{E}[(q/p)^{X_{n+1}}] = M_n\big((q/p)p + (p/q)q\big) = M_n,\end{aligned}$$

presque sûrement. En outre, les accroissements $M_{n\wedge T} - M_{(n-1)\wedge T}$ sont évidemment bornés uniformément en $n \geqslant 0$ et ω (mais pas en N). Le théorème d'arrêt implique donc que

$$\mathbb{E}_a[M_T] = \mathbb{E}_a[M_{T\wedge T}] = \mathbb{E}_a[M_{0\wedge T}] = (q/p)^a.$$

Or, $\mathbb{E}_a[M_T] = \mathbb{P}_a(S_T = 0) + (q/p)^N\mathbb{P}_a(S_T = N)$. On retrouve donc bien

$$\mathbb{P}_a(S_T = 0) = \frac{(q/p)^a - (q/p)^N}{1 - (q/p)^N}.$$

Problème du deuxième cœur

On part d'un paquet bien mélangé de 52 cartes à jouer (les 52! ordres possibles sont équiprobables). On retourne les cartes une à une jusqu'à l'apparition du premier cœur. Quelle est la probabilité que la carte suivante soit également un cœur ?

Notons $X_0, \dots, X_{51}$ la suite (finie) dans laquelle X_k dénote le nombre de cœurs restant dans le paquet après le tirage de la k-ième carte. Soit $(\mathcal{F}_k)_{k=0}^{51}$ la filtration naturelle. Observons que $X_{k+1} - X_k$ est égal à -1 si la $(k+1)$-ième carte tirée est un cœur et est égale à 0 sinon. On a donc, presque sûrement,

$$\mathbb{E}[X_{k+1} - X_k \,|\, \mathcal{F}_k] = -\frac{X_k}{52 - k}.$$

En particulier, $\mathbb{E}[X_{k+1} \,|\, \mathcal{F}_k] = \frac{52-(k+1)}{52-k}X_k$ presque sûrement. Il s'ensuit que le processus $Y_k \coloneqq \frac{X_k}{52-k}$ est une martingale relativement à $(\mathcal{F}_k)_{k=0}^{51}$. Soit $T \coloneqq \inf\{k \geqslant 0 \,|\, X_k < 13\}$ le temps où le premier cœur est tiré. Évidemment, $T < 52$. On peut donc appliquer le théorème d'arrêt, ce qui nous donne

$$\mathbb{E}[Y_T] = \mathbb{E}[Y_0] = \frac{13}{52} = \frac{1}{4}.$$

La conclusion suit en observant que $\mathbb{E}[Y_T] = \mathbb{P}((T+1)\text{-ième carte tirée est un cœur})$.

Problème de la secrétaire

On considère la procédure de recrutement suivante :

- ▷ N candidates pour un unique poste se présentent successivement à leur entretien d'embauche.
- ▷ La qualité des candidates selon les critères de l'employeur est représentée par un nombre dans l'intervalle $[0, 1]$.
- ▷ La k-ième candidate est interrogée au temps k et sa qualité est déterminée exactement lors de l'entretien. Immédiatement après celui-ci, l'employeur décide soit de l'engager, soit de rejeter sa candidature ; la décision est irréversible.

Quelle est la stratégie optimale afin de maximiser la qualité moyenne de la candidate retenue ?

On modélise ce problème de la façon suivante : la qualité de la k-ième candidate est donnée par une variable aléatoire $X_k \sim \mathsf{U}(0,1)$. Ces variables aléatoires sont supposées i.i.d..

Déterminer la meilleure stratégie consiste à trouver un temps d'arrêt T tel que $\mathbb{E}[X_T]$ soit maximal.

Remarque 10.27. *Afin de motiver l'analyse ci-dessous, commençons par un argument heuristique. Une idée naturelle consiste à retenir la k-ième candidate si X_k dépasse l'espérance du résultat que l'on obtiendrait en la rejetant. Notons donc α_k l'espérance de la qualité de la candidate retenue si l'on rejette la k-ième candidate (en appliquant la même stratégie à toutes les candidates suivantes). On a donc $\alpha_N = 0$ (on ne rejette jamais la N-ième candidate). Afin de déterminer les autres valeurs de α_k, notons Q_k la qualité de la candidate retenue si la k-ième est rejetée. On a donc*

$$Q_{k-1} = X_k \mathbf{1}_{\{X_k > \alpha_k\}} + Q_k \mathbf{1}_{\{X_k \leqslant \alpha_k\}}.$$

En prenant l'espérance, on obtient alors

$$\begin{aligned}\alpha_{k-1} = \mathbb{E}[Q_{k-1}] &= \mathbb{E}[X_k \mathbf{1}_{\{X_k > \alpha_k\}}] + \mathbb{E}[Q_k \mathbf{1}_{\{X_k \leqslant \alpha_k\}}] \\ &= \mathbb{E}[X_k \,|\, X_k > \alpha_k]\mathbb{P}(X_k > \alpha_k) + \alpha_k \mathbb{P}(X_k \leqslant \alpha_k) = \frac{1+\alpha_k}{2}(1-\alpha_k) + \alpha_k^2 = \frac{1}{2} + \frac{1}{2}\alpha_k^2.\end{aligned}$$

Cela suggère d'engager la première candidate telle que $X_k > \alpha_k$, où les α_k sont solutions du système ci-dessus.

Le lemme suivant montre que l'argument heuristique esquissé ci-dessus conduit bien à la stratégie optimale.

Lemme 10.28. *Soit $\alpha_N := 0$ et, pour $k \in \{1, \dots, N\}$, $\alpha_{k-1} := \frac{1}{2} + \frac{1}{2}\alpha_k^2$. Soit $\mathcal{F}_k := \sigma(X_1, \dots, X_k)$. Soit $T^* := \inf\{k \geqslant 1 \,|\, X_k > \alpha_k\}$. Alors,*

$$\mathbb{E}[X_T] \leqslant \mathbb{E}[X_{T^*}],$$

pour tout temps d'arrêt T (relativement à $(\mathcal{F}_k)_{k=1}^N$).

Démonstration. ▷ Étant donné un temps d'arrêt arbitraire T relativement à la filtration $(\mathcal{F}_k)_{k=0}^N$, on définit le processus stochastique $(Y_k^T)_{k=0}^N$ par $Y_0^T := \alpha_0$ et

$$\forall k \in \{1, \dots, N\}, \qquad Y_k^T := X_{T \wedge k} \vee \alpha_k.$$

Le processus $(Y_k^T)_{k=0}^N$ est une sur-martingale relativement à $(\mathcal{F}_k)_{k=1}^N$. En effet, on a d'une part, lorsque l'événement $T \leqslant k-1$ a lieu,

$$\mathbb{E}[Y_k^T \,|\, \mathcal{F}_{k-1}] = \mathbb{E}[X_T \vee \alpha_k \,|\, \mathcal{F}_{k-1}] = X_T \vee \alpha_k \leqslant X_T \vee \alpha_{k-1} = Y_{k-1}^T,$$

presque sûrement, puisque la suite (α_k) est décroissante ($\alpha_{k-1} - \alpha_k = \frac{1}{2} + \frac{1}{2}\alpha_k^2 - \alpha_k = \frac{1}{2}(1-\alpha_k)^2 \geqslant 0$).

D'autre part, lorsque l'événement $T > k-1$ a lieu, on a presque sûrement

$$\begin{aligned}\mathbb{E}[Y_k^T \,|\, \mathcal{F}_{k-1}] = \mathbb{E}[X_k \vee \alpha_k \,|\, \mathcal{F}_{k-1}] = \mathbb{E}[X_k \vee \alpha_k] &= \mathbb{E}[X_k \mathbf{1}_{\{X_k > \alpha_k\}}] + \mathbb{E}[\alpha_k \mathbf{1}_{\{X_k \leqslant \alpha_k\}}] \\ &= \frac{1}{2}(1-\alpha_k^2) + \alpha_k^2 = \alpha_{k-1} \leqslant Y_{k-1}^T,\end{aligned}$$

où l'on a utilisé le fait que X_k est indépendant de $\mathcal{F}_{k-1}$ pour la seconde égalité.
▷ Considérons à présent le cas du temps d'arrêt T^*. L'observation essentielle est que, dans ce cas, le processus $(Y_k^{T^*})_{k=0}^N$ est une martingale. En effet, on a d'une part, lorsque l'événement $T^* \leqslant k-1$ se produit, presque sûrement,

$$\mathbb{E}[Y_k^{T^*} \mid \mathcal{F}_{k-1}] = X_{T^*} \vee \alpha_k.$$

Or, par définition de T^*, $X_{T^*} > \alpha_{T^*}$. De plus, $\alpha_{T^*} \geqslant \alpha_{k-1} \geqslant \alpha_k$. Par conséquent, presque sûrement,

$$\mathbb{E}[Y_k^{T^*} \mid \mathcal{F}_{k-1}] = X_{T^*} \vee \alpha_k = X_{T^*} = X_{T^*} \vee \alpha_{k-1} = Y_{k-1}^{T^*}.$$

D'autre part, lorsque l'événement $T^* > k-1$ a lieu, on a presque sûrement

$$\mathbb{E}[Y_k^{T^*} \mid \mathcal{F}_{k-1}] = \alpha_{k-1} = Y_{k-1}^{T^*}.$$

La conclusion suit donc du théorème d'arrêt, puisque

$$\mathbb{E}[X_T] \leqslant \mathbb{E}[X_T \vee \alpha_T] = \mathbb{E}[Y_T^T] \leqslant \mathbb{E}[Y_0^T] = \alpha_0 = \mathbb{E}[Y_0^{T^*}] = \mathbb{E}[Y_{T^*}^{T^*}] = \mathbb{E}[X_{T^*} \vee \alpha_{T^*}] = \mathbb{E}[X_{T^*}]. \qquad \square$$

10.4.2 Inégalités maximales

Lemme 10.29 (Inégalité maximale de Doob). *Soit $(X_k)_{k\geqslant 0}$ une sous-martingale. Alors,*

$$\forall c > 0, \forall n \geqslant 0, \qquad \mathbb{P}(\max_{0\leqslant k\leqslant n} X_k \geqslant c) \leqslant \frac{\mathbb{E}[X_n^+]}{c}.$$

Démonstration. Soit $T \coloneqq \inf\{k \geqslant 0 \mid X_k \geqslant c\} \wedge n$ et $A \coloneqq \{\max_{0\leqslant k\leqslant n} X_k \geqslant c\}$. On a

$$\mathbb{E}[X_n \mathbf{1}_A] - \mathbb{E}[X_T \mathbf{1}_A] = \mathbb{E}[X_n] - \mathbb{E}[X_T] \geqslant 0,$$

l'égalité suivant de l'observation que $X_n \mathbf{1}_{A^c} = X_T \mathbf{1}_{A^c}$, et l'inégalité du Lemme 10.26. Par conséquent,

$$\mathbb{E}[X_n^+] \geqslant \mathbb{E}[X_n \mathbf{1}_A] \geqslant \mathbb{E}[X_T \mathbf{1}_A] \geqslant c\,\mathbb{P}(A) = c\,\mathbb{P}(\max_{0\leqslant k\leqslant n} X_k \geqslant c). \qquad \square$$

Lemme 10.30 (Inégalité de Doob–Kolmogorov). *Soit $(X_k)_{k\geqslant 0}$ une $\mathscr{L}^2$-martingale. Alors,*

$$\forall c > 0, \forall n \geqslant 0, \qquad \mathbb{P}(\max_{0\leqslant k\leqslant n} |X_k| \geqslant c) \leqslant \frac{\mathbb{E}[X_n^2]}{c^2}.$$

Démonstration. L'affirmation suit du Lemme 10.29 et du fait que $(X_n^2)_{n\geqslant 0}$ est une sous-martingale (voir l'Exemple 10.11) :

$$\mathbb{P}(\max_{0\leqslant k\leqslant n} |X_k| \geqslant c) = \mathbb{P}(\max_{0\leqslant k\leqslant n} X_k^2 \geqslant c^2) \leqslant \frac{\mathbb{E}[X_n^2]}{c^2}. \qquad \square$$

10.5 Convergence des martingales

Dans cette section, nous allons nous intéresser à la convergence d'une martingale $(X_n)_{n\geqslant 0}$ lorsque n tend vers l'infini. Ce type de résultats joue un rôle très important en théorie des probabilités. Nous discuterons les cas des martingales bornées dans $\mathscr{L}^1$ et $\mathscr{L}^2$.

10.5.1 Convergence des martingales bornées dans $\mathscr{L}^1$

Soit $X = (X_n)_{n\geqslant 0}$ un processus stochastique dont les accroissements $X_n - X_{n-1}$ représentent le gain par unité de mise au n-ième jeu. Considérons la stratégie de mise $(C_n)_{n\geqslant 1}$ suivante :

1. On fixe $a < b$.
2. On attend, sans miser, que X descende au-dessous de a.
3. On mise une unité à chaque jeu jusqu'à ce que X monte au-dessus de b.
4. On retourne au point 2.

Plus formellement, on pose $C_1 := \mathbf{1}_{\{X_0<a\}}$ et

$$\forall n \geqslant 2, \qquad C_n := \mathbf{1}_{\{C_{n-1}=1,\, X_{n-1}\leqslant b\}} + \mathbf{1}_{\{C_{n-1}=0,\, X_{n-1}<a\}}.$$

Le processus $(C_n)_{n\geqslant 1}$ est manifestement prévisible relativement à la filtration naturelle.

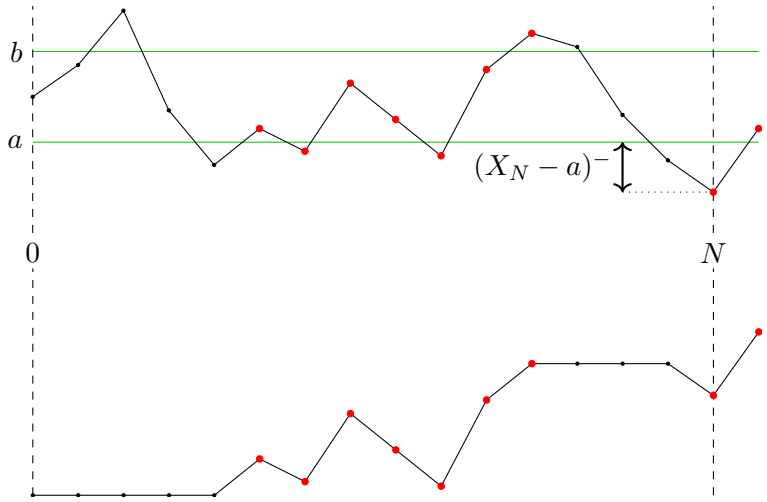


FIGURE 10.1: Haut : une réalisation du processus $(X_k)_{k\geqslant 0}$. Les points rouges indiquent les jeux au cours desquels une mise a été déposée. Bas : le processus $((C\bullet X)_n)_{n\geqslant 0}$ des gains.

Définition 10.31. *On note $U_N[a,b](\omega)$ le nombre de **traversées montantes** de $[a,b]$ par $n \mapsto X_n(\omega)$ jusqu'au temps N, c'est-à-dire le plus grand entier $k \in \mathbb{N}$ tel que*

$$\exists 0 \leqslant s_1 < t_1 < s_2 < t_2 < \cdots < s_k < t_k \leqslant N, \quad \forall i \in \{1,\dots,k\}, \quad X_{s_i}(\omega) < a,\ X_{t_i}(\omega) > b.$$

Il suit de la définition précédente que la stratégie ci-dessus rapport au moins, jusqu'au temps N,

$$(C\bullet X)_N(\omega) \geqslant (b-a)U_N[a,b](\omega) - \big(X_N(\omega) - a\big)^-, \tag{10.3}$$

où le premier terme du membre de droite correspond au gain dû aux traversées montantes et le second terme à la perte éventuelle lors du dernier intervalle.

Lemme 10.32. *Soit $(X_k)_{k\geqslant 0}$ une sur-martingale. Alors, pour tout $a < b$,*

$$\mathbb{E}\big[U_N[a,b]\big] \leqslant \frac{\mathbb{E}[(X_N - a)^-]}{b-a}.$$

Démonstration. $(C_n)_{n\geqslant 1}$ est prévisible, positif et borné. Par conséquent, il suit du Lemme 10.15 que $C\bullet X$ est une sur-martingale ; en particulier, $\mathbb{E}[(C\bullet X)_N] \leqslant 0$. La conclusion suit donc de (10.3). □

Corollaire 10.33. *Soit $(X_k)_{k\geqslant 0}$ une sur-martingale bornée dans $\mathscr{L}^1$ et $a < b$ deux réels. Alors, si l'on note $U_\infty[a,b] \coloneqq \lim_{N\to\infty} U_N[a,b]$, on a*

$$(b-a)\mathbb{E}\big[U_\infty[a,b]\big] \leqslant |a| + \sup_{n\geqslant 0}\mathbb{E}[|X_n|] < \infty.$$

En particulier,

$$\mathbb{P}(U_\infty[a,b] = \infty) = 0.$$

Démonstration. Il suit du Lemme 10.32 que

$$(b-a)\mathbb{E}\big[U_N[a,b]\big] \leqslant |a| + \mathbb{E}[|X_N|] \leqslant |a| + \sup_{n\geqslant 0}\mathbb{E}[|X_n|].$$

On conclut en prenant la limite $N \to \infty$ à l'aide du théorème de convergence monotone. □

Théorème 10.34 (Théorème de convergence de Doob)**.** *Soit $(X_k)_{k\geqslant 0}$ une sur-martingale bornée dans $\mathscr{L}^1$. Alors, il existe $X_\infty \in \mathscr{L}^1$, $\mathcal{F}_\infty$-mesurable et telle que*

$$X_n \xrightarrow{\text{p.s.}} X_\infty.$$

Démonstration. Soit

$$\begin{aligned}\Lambda \coloneqq \{\omega \mid X_n(\omega) \text{ ne converge pas dans } [-\infty,\infty]\} &= \{\omega \mid \liminf_{n\to\infty} X_n(\omega) < \limsup_{n\to\infty} X_n(\omega)\} \\ &= \bigcup_{\substack{a,b\in\mathbb{Q}\\ a<b}} \underbrace{\{\omega \mid \liminf_{n\to\infty} X_n(\omega) < a < b < \limsup_{n\to\infty} X_n(\omega)\}}_{=:\Lambda_{a,b}}.\end{aligned}$$

Or, comme $\Lambda_{a,b} \subset \{\omega \mid U_\infty[a,b](\omega) = \infty\}$, le Corollaire 10.33 implique que $\mathbb{P}(\Lambda_{a,b}) = 0$ pour tout $a < b$. On en conclut que $\mathbb{P}(\Lambda) = 0$. La variable aléatoire $X_\infty(\omega) \coloneqq \liminf_{n\to\infty} X_n(\omega)$ est $\mathcal{F}_\infty$-mesurable et $X_n \xrightarrow{\text{p.s.}} X_\infty$. Finalement, $X_\infty \in \mathscr{L}^1$, puisqu'il suit du lemme de Fatou que

$$\mathbb{E}[|X_\infty|] = \mathbb{E}[\liminf_{n\to\infty}|X_n|] \leqslant \liminf_{n\to\infty}\mathbb{E}[|X_n|] \leqslant \sup_{n\geqslant 1}\mathbb{E}[|X_n|] < \infty.$$ □

Corollaire 10.35. *Soit $(X_k)_{k\geqslant 0}$ une sur-martingale positive. Alors, il existe $X_\infty \in \mathscr{L}^1$, $\mathcal{F}_\infty$-mesurable et telle que $X_n \xrightarrow{\text{p.s.}} X_\infty$.*

Démonstration. Comme $\mathbb{E}[|X_n|] = \mathbb{E}[X_n] \leqslant \mathbb{E}[X_0]$, $(X_k)_{k\geqslant 0}$ est bornée dans $\mathscr{L}^1$ et le Théorème 10.34 s'applique. □

Exemple 10.36. Soit $(S_n)_{n\geqslant 0}$ la marche aléatoire simple symétrique sur $\mathbb{Z}$ partant de 1. On considère le temps d'arrêt $T \coloneqq \inf\{n \geqslant 0 \mid S_n = 0\}$ et le processus arrêté $X_n \coloneqq S_{n\wedge T}$.

$(X_n)_{n\geqslant 0}$ étant une martingale positive, le corollaire implique que $X_n \xrightarrow{\text{p.s.}} X_\infty$. De plus, $(S_n)_{n\geqslant 0}$ étant récurrente, on a $X_\infty = 0$ (puisque $\lim_{n\to\infty} S_{T\wedge n} = S_T = 0$ presque sûrement). En particulier, $\mathbb{E}[X_\infty] = 0$.

Mais, pour tout $n \geqslant 0$, $\mathbb{E}[X_n] = \mathbb{E}[X_0] = 1$. Par conséquent, $X_n \overset{1}{\not\to} X_\infty$. ⌟

L'exemple précédent montre que la convergence dans $\mathscr{L}^1$ n'est pas garantie pour une martingale bornée dans $\mathscr{L}^1$ et requiert des hypothèses supplémentaires. Nous allons analyser cette question dans les deux sections suivantes.

10.5.2 Convergence des martingales bornées dans $\mathscr{L}^2$

L'observation cruciale qui sous-tend l'analyse dans cette section est la propriété suivante : soit $m \leqslant n$, $(X_k)_{k\geqslant 0}$ une $\mathscr{L}^2$-martingale et Y une variable aléatoire $\mathcal{F}_m$-mesurable telle que $Y \in \mathscr{L}^2$. Alors,

$$\mathbb{E}[Y(X_n - X_m)] = \mathbb{E}\big[Y\mathbb{E}[X_n - X_m \,|\, \mathcal{F}_m]\big] = 0, \tag{10.4}$$

puisque $\mathbb{E}[X_n - X_m \,|\, \mathcal{F}_m] = \mathbb{E}[X_n \,|\, \mathcal{F}_m] - X_m = X_m - X_m = 0$ presque sûrement. En particulier, pour tout $0 \leqslant n < m$,

$$\begin{aligned}\mathbb{E}\big[(X_m - X_n)^2\big] &= \mathbb{E}\Big[\Big(\sum_{k=n+1}^{m}(X_k - X_{k-1})\Big)^2\Big] \\ &= 2\sum_{n+1\leqslant k<\ell<m}\mathbb{E}\big[(X_k - X_{k-1})(X_\ell - X_{\ell-1})\big] + \sum_{k=n+1}^{m}\mathbb{E}\big[(X_k - X_{k-1})^2\big] \\ &= \sum_{k=n+1}^{m}\mathbb{E}\big[(X_k - X_{k-1})^2\big], \end{aligned} \tag{10.5}$$

puisque $\mathbb{E}\big[(X_k - X_{k-1})(X_\ell - X_{\ell-1})\big] = 0$ par (10.4).

Lemme 10.37. *Une $\mathscr{L}^2$-martingale $(X_k)_{k\geqslant 0}$ est bornée dans $\mathscr{L}^2$ si et seulement si*

$$\sum_{k\geqslant 1}\mathbb{E}[(X_k - X_{k-1})^2] < \infty.$$

Démonstration. Comme $X_n = X_0 + \sum_{k=1}^{n}(X_k - X_{k-1})$,

$$X_n^2 = X_0^2 + 2\sum_{k=1}^{n}X_0(X_k - X_{k-1}) + \Big(\sum_{k=1}^{n}(X_k - X_{k-1})\Big)^2.$$

Par (10.4), $\mathbb{E}[X_0(X_k - X_{k-1})] = 0$ pour tout $1 \leqslant k \leqslant n$. Il suit donc de (10.5) que

$$\mathbb{E}[X_n^2] = \mathbb{E}[X_0^2] + \sum_{k=1}^{n}\mathbb{E}[(X_k - X_{k-1})^2]. \qquad \square$$

Théorème 10.38. *Soit $(X_k, \mathcal{F}_n)_k \geqslant 0$ une martingale bornée dans $\mathscr{L}^2$. Alors, il existe une variable aléatoire X telle que*

$$X_n \xrightarrow{\text{p.s.}} X \qquad \textit{et} \qquad X_n \xrightarrow{2} X.$$

Démonstration. Une martingale bornée dans $\mathscr{L}^2$ est nécessairement bornée dans $\mathscr{L}^1$ (se souvenir de la Remarque 2.61). Par conséquent, la convergence presque sûre suit du Théorème 10.34.

Par Fatou et (10.5),

$$\begin{aligned}\mathbb{E}[(X - X_n)^2] = \mathbb{E}[\liminf_{m\to\infty}(X_{n+m} - X_n)^2] &\leqslant \liminf_{m\to\infty}\mathbb{E}[(X_{n+m} - X_n)^2] \\ &= \liminf_{m\to\infty}\sum_{k=n+1}^{n+m}\mathbb{E}[(X_k - X_{k-1})^2] = \sum_{k=n+1}^{\infty}\mathbb{E}[(X_k - X_{k-1})^2].\end{aligned}$$

Par le Lemme 10.37, le terme de droite tend vers 0 lorsque $n \to \infty$, ce qui établit la convergence dans $\mathscr{L}^2$. $\square$

Exemple 10.39. Soit $(X_k)_{k\geqslant 1}$ une suite de variables aléatoires indépendantes telles que $\mathbb{E}[X_k] = 0$ et $\mathrm{Var}[X_k] < \infty$ pour tout $k \geqslant 1$. Nous allons montrer que

$$\sum_{k=1}^{\infty} \mathrm{Var}[X_k] < \infty \implies \sum_{k=1}^{\infty} X_k \text{ converge presque sûrement et dans } L^2.$$

Le processus $M_n := X_1 + \cdots + X_n$ est une martingale relativement à la filtration $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$. Comme

$$\mathbb{E}[(M_k - M_{k-1})^2] = \mathbb{E}[X_k^2] = \mathrm{Var}[X_k] \qquad \text{et} \qquad \sum_k \mathrm{Var}[X_k] < \infty,$$

$(M_n)_{n\geqslant 0}$ est bornée dans $\mathscr{L}^2$ par le Lemme 10.37. Il suit par conséquent du théorème 10.38 que $\sum_{k=1}^{\infty} X_k = \lim_{n\to\infty} M_n$ existe presque sûrement et la convergence a également lieu dans L^2.

Comme application concrète, considérons la série harmonique aléatoire. On sait que $\sum_{n\geqslant 1} \frac{1}{n} = \infty$, mais que $\sum_{n\geqslant 1} \frac{(-1)^n}{n} = \log 2 < \infty$. Que peut-on dire des séries $\sum_{n\geqslant 1} \frac{\epsilon_n}{n}$ avec $\epsilon_k \in \{-1, 1\}$ pour tout $k \geqslant 1$? Si les $(\epsilon_k)_{k\geqslant 1}$ sont des variables aléatoires i.i.d. telles que $\mathbb{P}(\epsilon_k = 1) = \mathbb{P}(\epsilon_k = -1) = \frac{1}{2}$, alors $\sum_{k\geqslant 1} \mathrm{Var}[\epsilon_k/k] = \sum_{k\geqslant 1} \frac{1}{k^2} < \infty$. Par conséquent, $\sum_{n\geqslant 1} \frac{\epsilon_n}{n}$ converge presque sûrement. ⌟

10.5.3 Martingales uniformément intégrables

Dans cette section, nous allons voir quelle est la condition minimale à introduire pour garantir la convergence dans $\mathscr{L}^1$.

Lemme 10.40. *Soit $X \in \mathscr{L}^1$. Pour tout $\epsilon > 0$, il existe $\delta > 0$ tel que, pour tout $A \in \mathcal{F}$ satisfaisant $\mathbb{P}(A) \leqslant \delta$, on a*

$$\mathbb{E}[|X|\mathbf{1}_A] \leqslant \epsilon.$$

Démonstration. Supposons le contraire. Alors, il existe $\epsilon > 0$ et $(A_n)_{n\geqslant 1} \subset \mathcal{F}$ tels que

$$\forall n \geqslant 1, \qquad \mathbb{P}(A_n) \leqslant 2^{-n} \quad \text{et} \quad \mathbb{E}[|X|\mathbf{1}_{A_n}] > \epsilon.$$

Soit $A := \limsup_{n\to\infty} A_n$. Alors, $\mathbb{P}(A) = 0$ par Borel–Cantelli. On a donc, par Fatou,

$$0 = \mathbb{E}[|X|\mathbf{1}_A] = \mathbb{E}[|X|] - \mathbb{E}[|X|\mathbf{1}_{A^c}] \geqslant \mathbb{E}[|X|] - \liminf_{n\to\infty} \mathbb{E}[|X|\mathbf{1}_{A_n^c}] = \limsup_{n\to\infty} \mathbb{E}[|X|\mathbf{1}_{A_n}] \geqslant \epsilon,$$

ce qui est absurde. □

Définition 10.41. *Une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ est **uniformément intégrable** (UI) si*

$$\lim_{K\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|\mathbf{1}_{\{|X_n|\geqslant K\}}] = 0.$$

Observons que la condition d'uniforme intégrabilité est plus forte que celle d'être borné dans $\mathscr{L}^1$: si $(X_n)_{n\in\mathbb{N}}$ est UI, on a, pour tout K suffisamment grand, $\sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|\mathbf{1}_{\{|X_n|\geqslant K\}}] \leqslant 1$. Par conséquent,

$$\sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|] \leqslant \sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|\mathbf{1}_{\{|X_n|<K\}}] + \sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|\mathbf{1}_{\{|X_n|\geqslant K\}}] \leqslant K + 1 < \infty.$$

Lemme 10.42. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires bornée dans $\mathscr{L}^1$. Alors, $(X_n)_{n\in\mathbb{N}}$ est UI si et seulement si pour tout $\epsilon > 0$, il existe $\delta > 0$ tel que, pour tout $A \in \mathcal{F}$ satisfaisant $\mathbb{P}(A) \leqslant \delta$,*

$$\forall n \in \mathbb{N}, \quad \mathbb{E}[|X_n|\mathbf{1}_A] \leqslant \epsilon.$$

Démonstration. $\boxed{\Rightarrow}$ Soit $\epsilon > 0$. On prend K tel que $\sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|\mathbf{1}_{\{|X_n|\geqslant K\}}] \leqslant \epsilon/2$. On pose $\delta := \epsilon/2K$. Pour tout $A \in \mathcal{F}$ tel que $\mathbb{P}(A) \leqslant \delta$,

$$\forall n \in \mathbb{N}, \qquad \mathbb{E}[|X_n|\mathbf{1}_A] = \mathbb{E}[|X_n|\mathbf{1}_{A\cap\{|X_n|\geqslant K\}}] + \mathbb{E}[|X_n|\mathbf{1}_{A\cap\{|X_n|<K\}}] \leqslant \epsilon.$$

$\boxed{\Leftarrow}$ Soit $\epsilon > 0$. On prend $\delta > 0$ tel que, pour tout $A \in \mathcal{F}$ satisfaisant $\mathbb{P}(A) \leqslant \delta$, on a

$$\forall n \in \mathbb{N}, \qquad \mathbb{E}[|X_n|\mathbf{1}_A] \leqslant \epsilon. \tag{10.6}$$

On pose $M := \sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|]$ et on considère K tel que $M/K \leqslant \delta$. Alors, par l'inégalité de Markov,

$$\forall n \in \mathbb{N}, \qquad \mathbb{P}(|X_n| \geqslant K) \leqslant \frac{\mathbb{E}[|X_n|]}{K} \leqslant \delta.$$

On peut donc appliquer (10.6) avec $A = \{|X_n| \geqslant K\}$, ce qui donne

$$\forall n \in \mathbb{N}, \qquad \mathbb{E}[|X_n|\mathbf{1}_{\{|X_n|\geqslant K\}}] \leqslant \epsilon. \qquad \square$$

L'intégrabilité uniforme garantit la convergence dans $\mathscr{L}^1$ de toute suite convergeant en probabilité.

Lemme 10.43. *Soit $(X_n)_{n\in\mathbb{N}} \subset \mathscr{L}^1$ telles que $X_n \xrightarrow{\mathbb{P}} X$. Si $(X_n)_{n\in\mathbb{N}}$ est UI, alors $X_n \xrightarrow{1} X$.*

Remarque 10.44. *La condition ci-dessus est optimale dans le sens suivant : si X et $(X_n)_{n\in\mathbb{N}}$ appartiennent toutes à $\mathscr{L}^1$, alors on peut montrer que*

$$X_n \xrightarrow{1} X \iff X_n \xrightarrow{\mathbb{P}} X \quad \textit{et} \quad (X_n)_{n\in\mathbb{N}} \textit{ est UI.}$$

Démonstration. Soit $\epsilon > 0$. On a

$$\mathbb{E}[|X_n - X_m|] \leqslant \epsilon + \mathbb{E}[|X_n - X_m|\mathbf{1}_{\{|X_n - X_m|>\epsilon\}}].$$

On a vu que l'uniforme intégrabilité de la suite implique que celle-ci est bornée dans $\mathscr{L}^1$. Il suit donc du Lemme 10.42 qu'il existe $\delta > 0$ tel que, pour tout $A \in \mathcal{F}$ satisfaisant $\mathbb{P}(A) \leqslant \delta$, on a

$$\forall k \in \mathbb{N}, \quad \mathbb{E}[|X_k|\mathbf{1}_A] \leqslant \epsilon. \tag{10.7}$$

Prenons n, m suffisamment grands pour que

$$\mathbb{P}(|X_n - X_m| \geqslant \epsilon) \leqslant \mathbb{P}(|X_n - X| \geqslant \epsilon/2) + \mathbb{P}(|X_m - X| \geqslant \epsilon/2) \leqslant \delta.$$

Il suit donc de (10.7) que

$$\mathbb{E}[|X_n - X_m|\mathbf{1}_{\{|X_n - X_m|>\epsilon\}}] \leqslant \mathbb{E}[|X_n|\mathbf{1}_{\{|X_n - X_m|>\epsilon\}}] + \mathbb{E}[|X_m|\mathbf{1}_{\{|X_n - X_m|>\epsilon\}}] \leqslant 2\epsilon.$$

On a donc montré que, pour tout m, n suffisamment grands, $\mathbb{E}[|X_n - X_m|] \leqslant 3\epsilon$, ce qui implique que la suite $(X_n)_{n\in\mathbb{N}}$ est une suite de Cauchy dans $\mathscr{L}^1$. Par complétude de L^1, elle est donc convergente : $\lim_{n\to\infty} \mathbb{E}[|X_n - X|] = 0$. $\square$

Appliquons le résultat précédent aux martingales.

Théorème 10.45. *Soit $(X_n)_{n\in\mathbb{N}}$ une sur-martingale UI. Alors, il existe $X_\infty \in \mathscr{L}^1$ $\mathcal{F}_\infty$-mesurable telle que*

$$X_n \xrightarrow{\text{p.s.}} X_\infty \qquad \textit{et} \qquad X_n \xrightarrow{1} X_\infty.$$

De plus, si $(X_n)_{n\in\mathbb{N}}$ est une martingale, alors $X_n = \mathbb{E}[X_\infty \,|\, \mathcal{F}_n]$ presque sûrement.

Démonstration. La suite $(X_n)_{n\in\mathbb{N}}$ étant UI, elle est bornée dans $\mathscr{L}^1$. Par le Théorème 10.34, il existe donc $X_\infty \in \mathscr{L}^1$ $\mathcal{F}_\infty$-mesurable telle que $X_n \xrightarrow{\text{p.s.}} X_\infty$. Par le Théorème 6.10, ceci implique que $X_n \xrightarrow{\mathbb{P}} X_\infty$. Finalement, le Lemme 10.43 garantit que $X_n \xrightarrow{1} X_\infty$.

Supposons à présent que $(X_n)_{n\in\mathbb{N}}$ soit une martingale. Fixons $n \in \mathbb{N}$. Nous devons montrer que $X_n = \mathbb{E}[X_\infty \,|\, \mathcal{F}_n]$ presque sûrement, c'est-à-dire que

$$\forall A \in \mathcal{F}_n, \qquad \mathbb{E}[X_n \mathbf{1}_A] = \mathbb{E}[X_\infty \mathbf{1}_A].$$

Soit $A \in \mathcal{F}_n$. On a

$$|\mathbb{E}[X_\infty \mathbf{1}_A] - \mathbb{E}[X_m \mathbf{1}_A]| \leqslant \mathbb{E}[|X_\infty - X_m| \mathbf{1}_A] \leqslant \mathbb{E}[|X_\infty - X_m|]$$

et le membre de droite tend vers 0 lorsque $m \to \infty$, puisque $X_m \xrightarrow{1} X_\infty$. On a donc

$$\mathbb{E}[X_\infty \mathbf{1}_A] = \lim_{m\to\infty} \mathbb{E}[X_m \mathbf{1}_A] = \lim_{m\to\infty} \mathbb{E}\big[\mathbb{E}[X_m \,|\, \mathcal{F}_n] \mathbf{1}_A\big] = \mathbb{E}[X_n \mathbf{1}_A]. \qquad \square$$

Rappelez-vous (Exemple 10.7) qu'une martingale $(X_n)_{n\in\mathbb{N}}$ est fermée s'il existe une variable aléatoire $X \in \mathscr{L}^1$ telle que, presque sûrement, $X_n = \mathbb{E}[X \,|\, \mathcal{F}_n]$ pour tout $n \in \mathbb{N}$.

Théorème 10.46. *Une martingale est fermée si et seulement si elle est UI.*

Démonstration.

$\boxed{\Leftarrow}$ Si $(X_n)_{n\in\mathbb{N}}$ est une martingale UI, le Théorème 10.45 implique l'existence de $X_\infty \in \mathscr{L}^1$ tel que, presque sûrement, $X_n = \mathbb{E}[X_\infty \,|\, \mathcal{F}_n]$ pour tout $n \in \mathbb{N}$.

$\boxed{\Rightarrow}$ Soit $X \in \mathscr{L}^1$ et $(\mathcal{F}_n)_{n\in\mathbb{N}}$ une filtration telle que, presque sûrement, $X_n = \mathbb{E}[X \,|\, \mathcal{F}_n]$ pour tout $n \in \mathbb{N}$. Alors, presque sûrement, $|X_n| = |\mathbb{E}[X \,|\, \mathcal{F}_n]| \leqslant \mathbb{E}[|X| \,|\, \mathcal{F}_n]$. On a donc

$$\forall A \in \mathcal{F}_n, \qquad \mathbb{E}[|X_n| \mathbf{1}_A] \leqslant \mathbb{E}\big[\mathbb{E}[|X| \,|\, \mathcal{F}_n] \mathbf{1}_A\big] = \mathbb{E}\big[|X| \mathbf{1}_A\big]. \tag{10.8}$$

En particulier, $\mathbb{E}[|X_n|] \leqslant \mathbb{E}[|X|]$, ce qui implique, par l'inégalité de Markov,

$$\mathbb{P}(|X_n| \geqslant K) \leqslant \frac{\mathbb{E}[|X_n|]}{K} \leqslant \frac{\mathbb{E}[|X|]}{K}.$$

Fixons $\epsilon > 0$. Par le Lemme 10.40, on peut trouver $\delta > 0$ tel que, pour tout $A \in \mathcal{F}$ tel que $\mathbb{P}(A) \leqslant \delta$, on a $\mathbb{E}[|X| \mathbf{1}_A] \leqslant \epsilon$. On choisit K suffisamment grand pour que $\mathbb{P}(|X_n| \geqslant K) \leqslant \delta$. Il suit donc de (10.8) et de notre choix de δ que

$$\mathbb{E}[|X_n| \mathbf{1}_{\{|X_n| \geqslant K\}}] \leqslant \mathbb{E}[|X| \mathbf{1}_{\{|X_n| \geqslant K\}}] \leqslant \epsilon,$$

ce qui montre que $(X_n)_{n\in\mathbb{N}}$ est UI. $\square$

11 Modèle de percolation

Si l'on place dans la mer un vase de cire modelé à cet usage, en en bouchant l'ouverture avec des matières que la mer ne puisse pénétrer, ce qui passe au travers des cloisons de la cire est de l'eau potable.

Aristote

Dans ce chapitre, nous allons introduire un autre processus très important en théorie des probabilités : le **modèle de percolation** (de Bernoulli). Contrairement aux chaînes de Markov ou aux martingales, il ne s'agit plus d'une famille de variables aléatoires indicées par un paramètre que l'on peut interpréter comme le temps, mais d'une famille de variables aléatoires indicées par un paramètre spatial ; on parle dans ce cas de *champ aléatoire*. Ce modèle peut être défini en dimension quelconque (et en fait, sur un graphe quelconque), mais nous nous contenterons de discuter le cas du réseau triangulaire.

11.1 Définition

Soit $p \in [0,1]$, et soit $(X_i)_{i \in V}$ une famille de variables aléatoires indépendantes suivant une loi de Bernoulli de paramètre p, indicées par les faces du réseau hexagonal. On note $\mathbb{P}_p$ la loi de ce champ. L'hexagone indicé par i est colorié en noir si $X_i = 1$ et en blanc sinon. Des réalisations de ce processus sont représentées sur la Figure 11.1 pour différentes valeurs du paramètre p.

Deux hexagones distincts, indicés respectivement par i et j, sont **voisins** s'ils partagent un côté ; on notera cette relation $i \sim j$.

Deux hexagones $i, j \in V$ appartiennent au même **amas**, noté $i \leftrightarrow j$, s'il existe $i_1, \dots, i_n \in V$ tels que (i) $i_1 = i$ et $i_n = j$, (ii) $i_k \sim i_{k+1}$ pour tout $1 \leqslant k < n$ et (iii) $X_{i_k} = 1$ pour tout $1 \leqslant k \leqslant n$. On notera $C(i) \coloneqq \{j \in V \mid i \leftrightarrow j\}$ l'amas contenant l'hexagone i (avec $C(i) \coloneqq \varnothing$ si $X_i = 0$).

L'interprétation originelle de ce processus est celle d'un modèle de matériau poreux. Un tel matériau contient un grand nombre de trous microscopiques. La question de base que l'on se pose alors est si cette porosité locale induit une porosité globale : si l'on plonge une pierre poreuse dans de l'eau, quelle est la probabilité que le centre de la pierre soit mouillé ? Dans le modèle de percolation, les trous correspondent aux hexagones noirs. La question de base peut alors se reformuler dans les termes suivants : existe-t-il un chemin infini d'hexagone noir contenant l'origine (l'eau pourrait alors se propager infiniment loin à travers ce dernier) ? Il y a de nombreuses autres interprétations bien entendu : épidémie, feu de forêt, etc. Ce modèle est devenu l'exemple classique pour modéliser des milieux aléatoires.

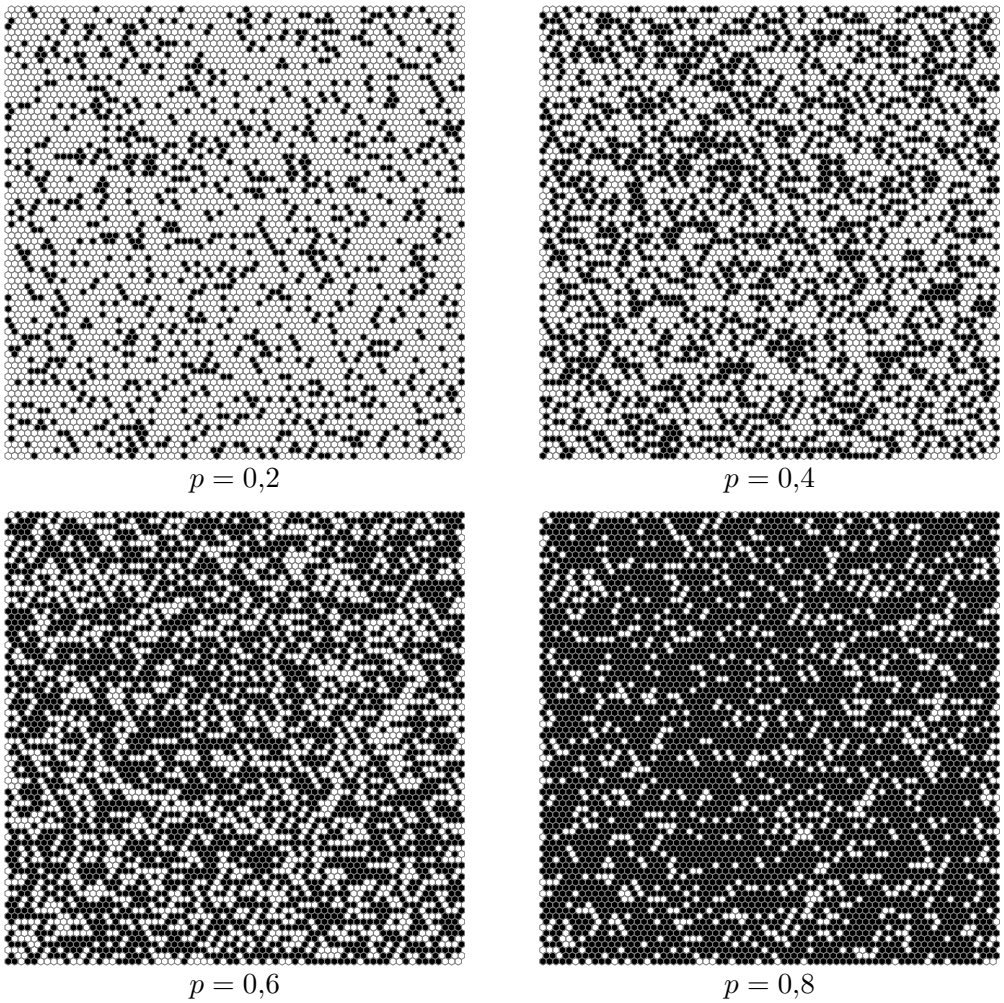


FIGURE 11.1: Quatre réalisations du processus de percolation pour différentes valeurs de p.

11.2 La transition de phase

Soit $\Theta(p) \coloneqq \mathbb{P}_p(|C(0)| = \infty)$ la probabilité qu'il existe un chemin infini d'hexagones noirs partant de 0 (en d'autres termes : la probabilité que de l'eau injectée à l'infini parvienne jusqu'à l'origine). Évidemment, $\Theta(0) = 0$ et $\Theta(1) = 1$. On pourrait alors s'attendre à ce que la fonction $p \mapsto \Theta(p)$ interpole de façon lisse entre ces deux valeurs lorsque p croît de 0 à 1. Or, ce n'est pas du tout le cas : $\Theta(p)$ reste identiquement nulle pour tout $p < p_c \in (0, 1)$, puis devient strictement positive pour $p > p_c$ (voir la Figure 11.2). Un tel type de comportement singulier lorsqu'on varie un paramètre d'un modèle est appelé une **transition de phase**. Le résultat suivant, fondamental, précise ce qui vient d'être dit.

Théorème 11.1. *1. Il existe $0 < p_c < 1$ tel que*

$$\forall p < p_c, \qquad \Theta(p) = 0$$
$$\forall p > p_c, \qquad \Theta(p) > 0.$$

2. La probabilité qu'il existe (au moins) un amas infini est égale à 1 si $\Theta(p) > 0$, et 0 sinon.

Remarque 11.2. *On possède aujourd'hui énormément d'information sur ce modèle. En voici quelques exemples, que l'on ne discutera pas davantage ici (mais il y a régulièrement des cours sur ce modèle en master, où ces résultats, et bien davantage, sont établis) :*

1. *un argument (très général) de théorie ergodique permet de montrer qu'avec probabilité* 1*, il n'y a jamais plus d'un amas infini ;*
2. *pour le réseau considéré ici, des considérations (non triviales) de symétrie permettent de montrer que* $p_{\mathrm{c}} = \frac{1}{2}$ *;*
3. *pour le réseau considéré ici, on peut montrer que* $\Theta(p_{\mathrm{c}}) = 0$*, c'est-à-dire que la fonction* $p \mapsto \Theta(p)$ *est continue.*

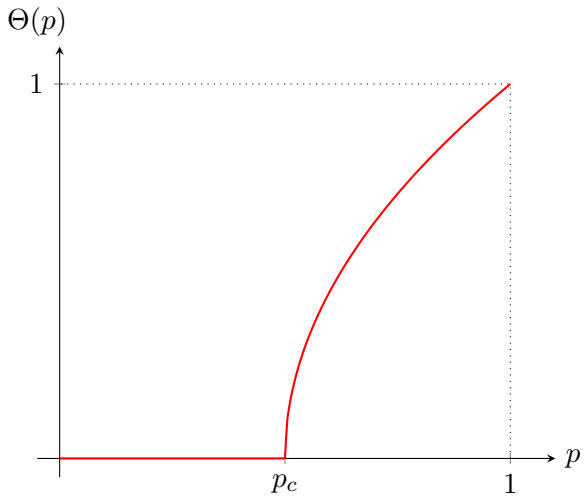


FIGURE 11.2: Comportement qualitatif de la fonction $p \mapsto \Theta(p)$.

Démonstration. On démontre d'abord la seconde affirmation. Clairement, l'existence d'au moins un amas infini est un événement asymptotique, puisque le fait de changer la couleur d'un nombre fini d'hexagones n'a pas d'influence sur sa réalisation. Par conséquent, il suit de la loi zéro-un de Kolmogorov (Théorème 6.21) que la probabilité qu'il existe au moins un amas infini a probabilité 0 ou 1.

Supposons que $\Theta(p) > 0$. Alors,

$$\mathbb{P}_p(\exists i \in V,\ |C(i)| = \infty) \geqslant \mathbb{P}_p(|C(0)| = \infty) > 0,$$

ce qui implique que $\mathbb{P}_p(\exists i \in V,\ |C(i)| = \infty) = 1$.

Supposons que $\Theta(p) = 0$. Alors,

$$\mathbb{P}_p(\exists i \in V,\ |C(i)| = \infty) = \mathbb{P}_p\Big(\bigcup_{i \in V} \{|C(i)| = \infty\}\Big) \leqslant \sum_{i \in V} \mathbb{P}_p(|C(i)| = \infty) = 0,$$

puisque $\mathbb{P}_p(|C(i)| = \infty) = \mathbb{P}_p(|C(0)| = \infty) = \Theta(p) = 0$.

Passons à présent à la première partie du théorème. Celle-ci suit clairement des trois affirmations suivantes :

(i) $\Theta(p) = 0$ pour tout p suffisamment petit ;

(ii) $\Theta(p) > 0$ pour tout p suffisamment proche de 1 ;

(iii) $p \mapsto \Theta(p)$ est une fonction croissante.

(i) On appelle **chemin de longueur n** dans V une suite $\pi = (\pi_1, \dots, \pi_n)$ de sommets de V tous distincts et tels que $\pi_k \sim \pi_{k-1}$ pour tout $k = 2, \dots, n$. Soit $\mathcal{N}(n)$ l'ensemble des chemins de longueur n commençant en $\pi_1 = 0$ et $N(n) := \#\mathcal{N}(n)$ leur nombre. On vérifie facilement que $N(n) \leqslant 6^n$. En effet, lorsque l'on construit un tel chemin hexagone par hexagone, on a au plus 6 choix à chaque étape.

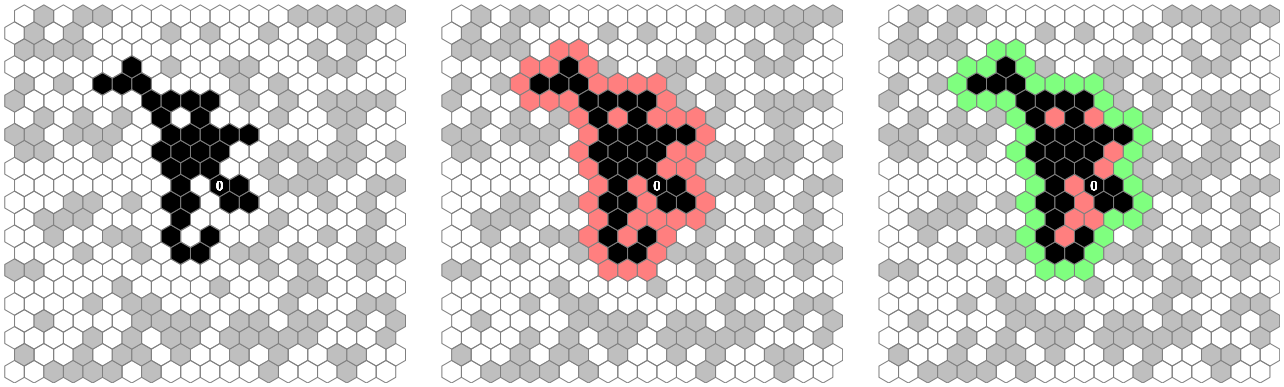


FIGURE 11.3: Lorsque $X_0 = 1$ mais que l'amas $C(0)$ est fini, il y a toujours un circuit d'hexagones blancs entourant l'origine. En effet, tous les hexagones n'appartenant pas à $C(0)$, mais voisins d'au moins un hexagone de $C(0)$ sont nécessairement blancs (les hexagones indiqués en rouge dans la figure du milieu) et l'on peut toujours extraire de ceux-ci un circuit d'hexagones blancs entourant l'origine (les hexagones indiqués en vert dans la figure de droite).

Soit à présent $N_{\hexagon}(n)$ le nombre de chemins de longueur n composés uniquement d'hexagones noirs. Étant donné $\pi = (\pi_1, \dots, \pi_n) \in \mathcal{N}(n)$, la probabilité de l'événement $\{\pi \text{ noir}\} := \{\forall k \in \{1, \dots, n\},\ X_{\pi_k} = 1\}$ est exactement donnée par $\prod_{k=1}^{n} \mathbb{P}_p(X_{\pi_k} = 1) = p^n$. Par conséquent,

$$\mathbb{E}_p[N_{\hexagon}(n)] = \mathbb{E}_p\Big[\sum_{\pi \in \mathcal{N}(n)} \mathbf{1}_{\{\pi \text{ noir}\}}\Big] = \sum_{\pi \in \mathcal{N}(n)} \mathbb{P}_p(\pi \text{ noir}) = p^n N(n) \leqslant (6p)^n.$$

Lorsque l'événement $\{|C(0)| = \infty\}$ est réalisé, il existe de tels chemins d'hexagones noirs de toutes les longueurs. On obtient donc

$$\forall n \geqslant 1, \qquad \mathbb{P}_p(|C(0)| = \infty) \leqslant \mathbb{P}_p(N_{\hexagon}(n) \geqslant 1) \leqslant \mathbb{E}_p[N_{\hexagon}(n)] \leqslant (6p)^n.$$

En laissant $n \to \infty$, on conclut que $\mathbb{P}_p(|C(0)| = \infty) = 0$ dès que $p < \frac{1}{6}$.

(ii) On va utiliser un argument dû à Peierls[1] introduit en 1936 dans l'étude d'un autre champ aléatoire très célèbre : le modèle d'Ising[2] ; cet argument prend une forme particulièrement simple dans le modèle de percolation. On appelle **circuit de longueur n** un chemin $\pi = (\pi_1, \dots, \pi_n)$ de longueur n tels que $\pi_1 \sim \pi_n$. L'observation élémentaire, mais cruciale, est que lorsque l'hexagone en 0 est noir, mais que l'amas $C(0)$ est fini, il existe un circuit composé entièrement d'hexagones blancs entourant l'origine (*cf.* Figure 11.3).

Notons $\mathcal{N}^{\circlearrowleft}(n)$ l'ensemble des circuits de longueur n entourant l'origine. On peut à nouveau facilement borner leur nombre $N^{\circlearrowleft}(n) := \#\mathcal{N}^{\circlearrowleft}(n)$. En effet, le nombre de circuits de longueur n contenant un sommet donné est inférieur à 6^n, puisqu'il est inférieur au nombre de chemins de longueur n contenant ce sommet. De plus, un circuit de longueur n entourant l'origine contient nécessairement un des hexagones se trouvant sur le même rangée que 0, à sa droite et à une distance inférieure à n. On obtient donc que $N^{\circlearrowleft}(n) \leqslant n6^n$.

Soit $\pi = (\pi_1, \dots, \pi_n) \in \mathcal{N}^{\circlearrowleft}(n)$; la probabilité de l'événement $\{\pi \text{ blanc}\} := \{\forall k \in \{1, \dots, n\},\ X_{\pi_k} = 0\}$ est égale à $\prod_{k=1}^{n} \mathbb{P}_p(X_{\pi_k} = 0) = (1-p)^n$. Soit $N_{\hexagon}^{\circlearrowleft}(n)$ le nombre de circuits entourant l'origine entièrement composés d'hexagones blancs. Alors,

$$\mathbb{E}_p[N_{\hexagon}^{\circlearrowleft}(n)] = \sum_{\pi \in \mathcal{N}^{\circlearrowleft}(n)} \mathbb{P}_p(\pi \text{ blanc}) = (1-p)^n N^{\circlearrowleft}(n) \leqslant n(6(1-p))^n.$$

1. Sir Rudolf Ernst Peierls (1907, Berlin – 1995, Oxford), physicien théoricien allemand. Il s'installa en Angleterre en 1933, et fut anobli en 1968.

2. Ernst Ising (1900, Cologne – 1998, Peoria), physicien allemand. On estime à environ 800 le nombre d'articles publiés chaque année portant sur le modèle d'Ising.

Comme on l'a vu, lorsque $X_0 = 1$ et $|C(0)| < \infty$, il existe un entier n tel que $N_{\hexagon}^{\circlearrowleft}(n) \geqslant 1$. On a donc

$$\begin{aligned}\mathbb{P}_p(X_0 = 1, |C(0)| < \infty) \leqslant \mathbb{P}_p\Big(\bigcup_{n\geqslant 6}\{N_{\hexagon}^{\circlearrowleft}(n) \geqslant 1\}\Big) &\leqslant \sum_{n\geqslant 6}\mathbb{P}_p\big(N_{\hexagon}^{\circlearrowleft}(n) \geqslant 1\big)\\ &\leqslant \sum_{n\geqslant 6}\mathbb{E}_p[N_{\hexagon}^{\circlearrowleft}(n)] \leqslant \sum_{n\geqslant 6} n(6(1-p))^n,\end{aligned}$$

où l'on a utilisé le fait que le plus court circuit entourant l'origine est de longueur 6. Comme le membre de droite tend vers 0 lorsque $p \to 1$,

$$\mathbb{P}_p(|C(0)| = \infty) = 1 - (1-p) - \mathbb{P}_p(X_0 = 1, |C(0)| < \infty) > 0$$

pour tout p suffisamment proche de 1.

(iii) Il reste à montrer que $\Theta(p)$ est une fonction croissante de p. Soit $(Y_i)_{i\in V}$ une famille de variables aléatoires i.i.d. de loi uniforme sur $[0,1]$; on note $\widehat{\mathbb{P}}$ leur loi conjointe. Pour $p \in [0,1]$, on définit les variables aléatoires $(X_i^p)_{i\in V}$ par

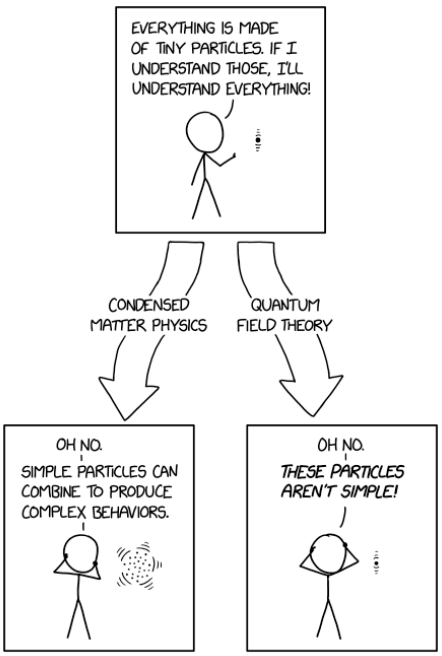
$$X_i^p := \begin{cases} 1 & \text{si } Y_i \leqslant p,\\ 0 & \text{si } Y_i > p.\end{cases}$$

Il est clair que, pour chaque $p \in [0,1]$, les variables aléatoires $(X_i^p)_{i\in V}$ sont i.i.d. et suivent une loi de Bernoulli de paramètre $\widehat{\mathbb{P}}(X_i^p = 1) = \mathbb{P}(Y_i \leqslant p) = p$. La loi conjointe de $(X_i^p)_{i\in V}$ est donc donnée par $\mathbb{P}_p$.

L'intérêt de cette construction est qu'elle permet de définir *simultanément* tous les processus $(X_i^p)_{i\in V}$ pour $p \in [0,1]$ sur un même espace de probabilité, ce qui permet de les comparer réalisation par réalisation. C'est ce que l'on appelle faire un **couplage** de ces processus. En particulier, on voit que la présence d'un amas infini contenant l'origine pour X^p implique l'existence d'un amas infini pour tous les processus $X^{p'}$ avec $p' \geqslant p$, puisque $Y_i \leqslant p \implies Y_i \leqslant p'$, pour tout $p' \geqslant p$, ce qui implique que chaque hexagone noir dans X^p l'est nécessairement également dans $X^{p'}$. On a donc, pour $0 \leqslant p \leqslant p' \leqslant 1$,

$$\begin{aligned}\Theta(p) = \mathbb{P}_p(|C(0)| = \infty) &= \widehat{\mathbb{P}}(|C(0)| = \infty \text{ dans } X^p)\\ &\leqslant \widehat{\mathbb{P}}(|C(0)| = \infty \text{ dans } X^{p'}) = \mathbb{P}_{p'}(|C(0)| = \infty) = \Theta(p').\end{aligned}$$

□



https://xkcd.com/2933

12 Introduction à la statistique

It is the mark of a truly intelligent person to be moved by statistics.

George Bernard Shaw

If your experiment needs statistics, you ought to have done a better experiment.

Ernest Rutherford

Ce chapitre est consacré à une brève introduction aux méthodes statistiques. Le point de vue est conceptuellement très différent de celui des chapitres précédents, dont la nature est plus probabiliste. Plutôt que de se donner à priori un espace de probabilité (ou une collection de variables aléatoires de loi conjointe donnée) et d'étudier ses propriétés, ici on considère le problème suivant : on se donne une collection $x_1, \dots, x_n$ d'observations résultant de la répétition d'une série d'expériences aléatoires et on cherche à déterminer la loi des variables aléatoires correspondantes.

12.1 Estimateurs

12.1.1 Définition, consistance, biais

Dans ce chapitre, nous supposerons toujours que les observations sont obtenues lors d'expériences indépendantes effectuées dans les mêmes conditions ; ceci conduit à la notion de n-échantillon.

Définition 12.1. *Soit $\mathbb{P}$ une mesure de probabilité sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Un **échantillon de taille n** (ou **n-échantillon**) de loi $\mathbb{P}$ est une famille $X_1, \dots, X_n$ de variables aléatoires i.i.d. de loi $\mathbb{P}$.*

*Une **réalisation** d'un n-échantillon est le résultat de n tirages indépendants selon la loi $\mathbb{P}$; c'est une collection $x_1, \dots, x_n$ de points de $\mathbb{R}^d$.*

Exemple 12.2.

1. Sondage de n individus sur une question binaire. Dans ce cas, on modélise l'échantillon par une collection de n variables aléatoires indépendantes suivant toutes une loi de Bernoulli de paramètre $p \in [0, 1]$.
2. Durée de vie de composants électroniques. Dans ce cas, une possibilité consiste à modéliser les durées de vie par une famille de variables aléatoires i.i.d. de loi exponentielle de paramètre $\lambda > 0$.
3. Répartition de la taille des individus dans une population homogène. Une possibilité consiste à modéliser cette situation par une collection de variables aléatoires i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$. ⌟

Dans chaque cas, les variables aléatoires formant le n-échantillon suivent une loi $\mathbb{P}$ dont la nature est connue, mais dépendant d'un ou plusieurs paramètres en général inconnus ; on notera θ la collection de paramètres, Θ l'ensemble des valeurs que θ peut prendre et $\mathbb{P}_\theta$ la loi correspondante. Pour les exemples précédents :

1. $\theta \coloneqq p \in \Theta \coloneqq [0,1]$.
2. $\theta \coloneqq \lambda \in \Theta \coloneqq \mathbb{R}_+^*$.
3. $\theta \coloneqq (\mu, \sigma^2) \in \Theta \coloneqq \mathbb{R}_+^* \times \mathbb{R}_+^*$.

Le problème fondamental est de prédire (une valeur approchée de) θ à partir des données (c'est-à-dire d'une réalisation du n-échantillon). On parle alors d'**estimation paramétrique**. Plus généralement, on peut être intéressé à estimer $f(\theta)$ pour une certaine fonction f définie sur Θ.

Définition 12.3. *Soit $X_1, \dots, X_n$ un n-échantillon de loi $\mathbb{P}_\theta$. Un **estimateur** de $f(\theta)$ est une fonction mesurable $\hat{f} : \Omega \to f(\Theta)$, $\hat{f}(\omega) = F(X_1(\omega), \dots, X_n(\omega))$, où $F : \mathbb{R}^{nd} \to f(\Theta)$ ne dépend pas de θ.*

Chaque réalisation d'un estimateur $\hat{f}$ de $f(\theta)$ fournit une **estimée** de $f(\theta)$. Comme on ne connaît pas la valeur de θ (c'est précisément ce que l'on cherche à estimer !), l'estimateur $\hat{f}$ ne doit évidemment pas en dépendre.

La raison pour laquelle on doit se contenter d'estimer les paramètres de la loi est que l'on ne dispose que d'échantillons finis. Une propriété essentielle que l'on demande à un estimateur est de donner, dans la limite où la taille de l'échantillon tend vers l'infini, la valeur exacte de la quantité que l'on cherche à estimer.

Définition 12.4. *Soit $X_1, \dots, X_n$ un n-échantillon de loi $\mathbb{P}_\theta$ et $\hat{f}_n$ un estimateur de $f(\theta)$. $\hat{f}_n$ est un estimateur **consistant** (ou **convergent**) s'il converge en probabilité vers $f(\theta)$ lorsque $n \to \infty$:*

$$\forall \epsilon > 0, \forall \theta \in \Theta, \qquad \lim_{n\to\infty} \mathbb{P}_\theta(|\hat{f}_n - f(\theta)| \geqslant \epsilon) = 0.$$

Exemple 12.5. Soit $X_1, \dots, X_n$ un n-échantillon. La moyenne empirique

$$\overline{X}_n \coloneqq \frac{1}{n}(X_1 + \cdots + X_n)$$

est un estimateur de $f(\theta) \coloneqq \mathbb{E}_\theta[X]$. La loi des grands nombres implique que cet estimateur est consistant. ⌟

Une caractéristique classique d'un estimateur est son biais.

Définition 12.6. *Le **biais** d'un estimateur $\hat{f}$ de $f(\theta)$ est défini par $\mathbb{E}_\theta[\hat{f} - f(\theta)] = \mathbb{E}_\theta[\hat{f}] - f(\theta)$. On dit que $\hat{f}$ est un estimateur **sans biais** de $f(\theta)$ si $\forall \theta \in \Theta,\ \mathbb{E}_\theta[\hat{f}] = f(\theta)$. Dans le cas contraire, on dit qu'il est **biaisé**.*

Insister sur l'absence de biais est utile lorsqu'on veut démontrer l'optimalité de certains estimateurs. Dans la pratique, ce n'est pas une condition toujours désirable : il est en effet tout à fait possible qu'un estimateur biaisé soit meilleur qu'un estimateur sans biais. Nous reviendrons sur ce point plus tard.

Définition 12.7. *Une famille $(\hat{f}_n)_{n\geqslant 1}$ d'estimateurs de $f(\theta)$ est **asymptotiquement sans biais** si*

$$\forall \theta \in \Theta, \qquad \lim_{n\to\infty} \mathbb{E}_\theta[\hat{f}_n] = f(\theta).$$

Lemme 12.8. *Si $\hat{f}_n$ est un estimateur de $f(\theta)$ asymptotiquement sans biais dont la variance tend vers 0 lorsque $n \to \infty$, alors $\hat{f}_n$ est un estimateur consistant de $f(\theta)$.*

Démonstration. Soit $\epsilon > 0$. Par le Lemme 6.4,

$$\mathbb{P}_\theta(|\hat{f}_n - f(\theta)| \geqslant \epsilon) = \mathbb{P}_\theta((\hat{f}_n - f(\theta))^2 \geqslant \epsilon^2) \leqslant \epsilon^{-2}\mathbb{E}_\theta[(\hat{f}_n - f(\theta))^2],$$

pour tout $\theta \in \Theta$. L'affirmation suit, puisque $\mathbb{E}_\theta[(\hat{f}_n - f(\theta))^2] = \mathrm{Var}_\theta[\hat{f}_n] + (\mathbb{E}_\theta[\hat{f}_n - f(\theta)])^2$ et que chacun de ces deux termes tend vers 0 par hypothèse. □

12.1.2 Quelques exemples

Moyenne empirique

Soit $X_1, \dots, X_n$ un n-échantillon de loi $\mathbb{P}_\theta$. On cherche à estimer l'espérance $f(\theta) \coloneqq \mathbb{E}_\theta[X_1]$. Un estimateur naturel est la moyenne empirique de l'échantillon :

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n).$$

Comme mentionné plus haut, sa consistance suit de la loi des grands nombres. D'autre part,

$$\mathbb{E}_\theta[\overline{X}_n] = \frac{1}{n}(\mathbb{E}_\theta[X_1] + \cdots + \mathbb{E}_\theta[X_n]) = \mathbb{E}_\theta[X_1] = f(\theta).$$

Il s'agit donc d'un estimateur sans biais de $f(\theta)$.

Variance empirique

On désire à présent estimer la variance $f(\theta) \coloneqq \mathrm{Var}_\theta[X_1]$. Un estimateur naturel est

$$\tilde{\sigma}_n^2 \coloneqq \frac{1}{n}\big(X_1^2 + \cdots + X_n^2\big) - \big(\frac{1}{n}(X_1 + \cdots + X_n)\big)^2.$$

La loi des grands nombres implique sa consistance, puisque le premier terme converge vers $\mathbb{E}_\theta[X_1^2]$ et le second vers $\mathbb{E}_\theta[X_1]^2$. Calculons le biais de cet estimateur. On a

$$\mathbb{E}_\theta\Big[\frac{1}{n}(X_1^2 + \cdots + X_n^2)\Big] = \mathbb{E}_\theta[X_1^2],$$
$$\mathbb{E}_\theta\Big[\big(\frac{1}{n}(X_1 + \cdots + X_n)\big)^2\Big] = \frac{1}{n}\mathbb{E}_\theta[X_1^2] + \frac{n-1}{n}\mathbb{E}_\theta[X_1]^2.$$

Par conséquent,

$$\mathbb{E}_\theta[\tilde{\sigma}_n^2] = \frac{n-1}{n}\big(\mathbb{E}_\theta[X_1^2] - \mathbb{E}_\theta[X_1]^2\big) = \frac{n-1}{n}\sigma^2.$$

Cet estimateur est donc biaisé. On peut facilement en déduire une variante à la fois consistante et sans biais :

$$V_n^2 \coloneqq \frac{n}{n-1}\tilde{\sigma}_n^2.$$

Covariance empirique

On considère un n-échantillon $(X_1, Y_1), \dots, (X_n, Y_n)$ de loi $\mathbb{P}_\theta$ et on cherche à estimer la covariance $f(\theta) \coloneqq \mathrm{Cov}_\theta[X, Y]$. Des considérations tout à fait similaires à celles faites ci-dessus pour la variance montrent que l'estimateur naturel

$$\tilde{\tau}_n \coloneqq \frac{1}{n}\big(X_1Y_1 + \cdots + X_nY_n\big) - \big(\frac{1}{n}(X_1 + \cdots + X_n)\big)\big(\frac{1}{n}(Y_1 + \cdots + Y_n)\big)$$

est consistant et biaisé, mais que l'estimateur

$$\hat{\tau}_n \coloneqq \frac{n}{n-1}\tilde{\tau}_n$$

est consistant et sans biais.

Méthode de Monte-Carlo.

On cherche à approximer numériquement

$$I \coloneqq \int_a^b h(x)\,\mathrm{d}x,$$

avec $h : [a,b] \to \mathbb{R}$. Une approche consiste à interpréter I comme une espérance :

$$I = (b-a)\int_{\mathbb{R}} h(x)\frac{\mathbf{1}_{[a,b]}(x)}{b-a}\,\mathrm{d}x = (b-a)\,\mathbb{E}[h(X)],$$

où $X \sim \mathsf{U}(a,b)$. On peut ainsi estimer I à l'aide de l'estimateur

$$\hat{I} \coloneqq (b-a)\frac{1}{n}\big(h(U_1)+\cdots+h(U_n)\big),$$

où $U_1,\dots,U_n$ est un n-échantillon de loi uniforme sur $[a,b]$. $\hat{I}$ est un estimateur sans biais et consistant de I. Notons toutefois que cette méthode est particulièrement inefficace comparativement aux autres méthodes d'analyse numérique habituellement utilisées pour résoudre ce problème (au moins lorsque la fonction f a un peu de régularité). Elle devient cependant beaucoup plus performante que ces dernières lorsque l'on remplace l'intervalle par un domaine de $\mathbb{R}^d$ avec d grand.

12.1.3 Construction d'estimateurs

Nous allons à présent discuter deux méthodes permettant de construire des estimateurs de $f(\theta)$: la méthode des moments et celle du maximum de vraisemblance.

Méthode des moments

Soit $X_1,\dots,X_n$ un n-échantillon de loi $\mathbb{P}_\theta$. Supposons que l'on puisse trouver une fonction g telle que $\theta = \mathbb{E}_\theta[g(X_1)]$. Alors, on peut estimer θ à l'aide de l'estimateur naturel

$$\hat{\theta} = \frac{1}{n}\big(g(X_1)+\cdots+g(X_n)\big)$$

et on vérifie immédiatement que ce dernier est consistant et sans biais.

Exemple 12.9. Si $X_1,\dots,X_n$ est un n-échantillon de loi uniforme sur $[0,\theta]$, $\theta > 0$, alors

$$\mathbb{E}_\theta[X_1] = \tfrac{1}{2}\theta$$

et on peut utiliser $\hat{\theta} \coloneqq 2\overline{X}_n$ pour estimer θ sans biais. ⌟

Plus généralement, on peut chercher une relation de la forme $\theta = h(\mathbb{E}_\theta[g(X_1)])$. Un choix classique, qui donne son nom à la méthode, correspond alors à considérer $g(x) = x^r$, ce qui permet d'estimer θ lorsque ce dernier peut s'exprimer en termes d'un moment $\mathbb{E}_\theta[X^r]$: $\theta = h(\mathbb{E}_\theta[X^r])$. On considère alors l'estimateur, en général biaisé,

$$\hat{\theta} \coloneqq h\big(\frac{1}{n}(X_1^r+\cdots+X_n^r)\big).$$

Exemple 12.10. Si $X_1,\dots,X_n$ est un n-échantillon de loi exponentielle de paramètre θ, alors puisque

$$\mathbb{E}_\theta[X_1] = 1/\theta,$$

on peut utiliser $\hat{\theta} \coloneqq 1/\overline{X}_n$ pour estimer θ. ⌟

Estimateur du maximum de vraisemblance

On considère un n-échantillon $X_1, \dots, X_n$ de loi $\mathbb{P}_\theta$. Étant en possession d'une réalisation $x_1, \dots, x_n$ du n-échantillon, une approche naturelle au problème de l'estimation consiste à déterminer, parmi toutes les valeurs possibles de θ, celle sous laquelle il était le plus probable d'avoir observé les valeurs $x_1, \dots, x_n$. En d'autres termes, on cherche la valeur de θ qui explique le mieux les valeurs obtenues. Nous allons à présent construire un estimateur basé sur cette idée. On suppose, pour commencer, les variables aléatoires $X_1, \dots, X_n$ discrètes.

Définition 12.11. *La **vraisemblance** (ou **fonction de vraisemblance**), notée $L_{x_1,\dots,x_n}(\theta)$, d'un modèle en $x_1, \dots, x_n$ est la probabilité de l'événement $\{X_1 = x_1, \dots, X_n = x_n\}$ lorsque le paramètre est θ.*

Remarque 12.12. *Insistons sur le fait que la variable est θ ; $x_1, \dots, x_n$ sont des paramètres.*

Par indépendance des observations, on peut écrire

$$L_{x_1,\dots,x_n}(\theta) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_i = x_i).$$

La définition ci-dessus n'a de sens que pour des variables aléatoires discrètes. Dans le cas de variables à densité, on travaille avec les densités de probabilité :

$$L_{x_1,\dots,x_n}(\theta) = \prod_{i=1}^{n} f_\theta(x_i),$$

où f_θ est la densité de probabilité associée à la loi $\mathbb{P}_\theta$.

Définition 12.13. *On appelle **estimateur du maximum de vraisemblance** de θ la variable aléatoire correspondant à la valeur $\hat{\theta} \coloneqq \operatorname{argmax} L_{X_1,\dots,X_n}(\theta)$ en laquelle la fonction de vraisemblance atteint son maximum.*

Exemple 12.14. On considère une réalisation $x_1, \dots, x_n \in \mathbb{R}_+^*$ d'un n-échantillon de loi $\mathsf{exp}(\lambda)$. La fonction de vraisemblance est

$$L_{x_1,\dots,x_n}(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda(x_1+\dots+x_n)}.$$

Pour trouver le maximum, on considère la **log-vraisemblance**,

$$\log L_{x_1,\dots,x_n}(\lambda) = n \log \lambda - \lambda(x_1 + \dots + x_n).$$

La dérivée de cette dernière s'annule en $\lambda = n/(x_1 + \dots + x_n)$, et on vérifie qu'il s'agit d'un maximum. L'estimateur du maximum de vraisemblance de λ est donc

$$\hat{\lambda} = \frac{n}{X_1 + \dots + X_n}. \qquad \lrcorner$$

Exemple 12.15. On considère un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$. On veut estimer les deux paramètres à présent, c'est-à-dire $\theta = (\mu, \sigma^2)$. Le calcul est similaire (mais on travaille avec une fonction de deux variables à présent) et est laissé en exercice. On trouve que l'estimateur du maximum de vraisemblance est $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$ où

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu})^2. \qquad \lrcorner$$

Exemple 12.16. On considère un n-échantillon de loi $\mathsf{U}(0,\theta)$, avec $\theta > 0$. La fonction de vraisemblance prend la forme

$$L_{x_1,\dots,x_n}(\theta) = \frac{1}{\theta^n}\prod_{i=1}^{n}\mathbf{1}_{\{x_i<\theta\}} = \frac{1}{\theta^n}\mathbf{1}_{\{\max_i x_i\leqslant\theta\}}.$$

La fonction de vraisemblance est nulle si $\theta < \max_i x_i$. Supposons donc que $\theta \geqslant \max_i x_i$. Dans ce cas, $L_{x_1,\dots,x_n}(\theta) = \theta^{-n}$, qui est une fonction décroissante de θ. Le maximum est donc atteint en $\theta = \max_i x_i$. L'estimateur du maximum de vraisemblance est donc donné par

$$\hat{\theta} = \max\{X_1,\dots,X_n\}.$$ ⌟

12.1.4 Comparaison d'estimateurs

Étant donné qu'il est possible de définir une multitude d'estimateurs différents pour la même quantité, il est important d'avoir un moyen de les comparer. Une façon de procéder consiste à considérer la dispersion de la loi de l'estimateur, puisque celle-ci représente l'erreur typique que l'on commet lors d'une application.

Définition 12.17. *Soit $\hat{f}$ un estimateur de $f(\theta)$. Le **risque quadratique** de l'estimateur $\hat{f}$ est défini par*

$$\mathcal{R}_{\hat{f}}(\theta) \coloneqq \mathbb{E}_\theta[(\hat{f} - f(\theta))^2].$$

Définition 12.18. *Si $\hat{f}_1$ et $\hat{f}_2$ sont deux estimateurs de $f(\theta)$, on dira que $\hat{f}_1$ est meilleur que $\hat{f}_2$ si*

$$\forall\theta\in\Theta, \qquad \mathcal{R}_{\hat{f}_1}(\theta) < \mathcal{R}_{\hat{f}_2}(\theta).$$

Lemme 12.19. *Soit $\hat{f}$ un estimateur de $f(\theta)$. Alors*

$$\mathcal{R}_{\hat{f}}(\theta) = \mathrm{Var}_\theta[\hat{f}] + \big(\mathbb{E}_\theta[\hat{f} - f(\theta)]\big)^2.$$

En particulier, si $\hat{f}$ est sans biais, alors

$$\mathcal{R}_{\hat{f}}(\theta) = \mathrm{Var}_\theta[\hat{f}].$$

Démonstration. Exercice élémentaire. □

Observez que cette décomposition montre qu'afin de minimiser le risque, il peut être favorable d'avoir un biais, si cela permet de faire décroître la variance.

Exemple 12.20. On considère un n-échantillon distribué uniformément sur $[0,\theta]$, $\theta > 0$. Le risque associé à l'estimateur

$$\bar{\theta} = \frac{2}{n}(X_1 + \cdots + X_n)$$

de θ vaut

$$\mathcal{R}_{\bar{\theta}} = \frac{4}{n}\mathrm{Var}_\theta[X_1] = \frac{\theta^2}{3n}.$$

Considérons à présent l'estimateur du maximum de vraisemblance de θ,

$$\tilde{\theta} = \max\{X_1,\dots,X_n\}.$$

Manifestement, cet estimateur est biaisé, puisqu'on a toujours $\mathbb{E}_\theta[\tilde{\theta}] < \theta$. Commençons par déterminer la loi de $\tilde{\theta}$: pour tout $x\in[0,\theta]$,

$$\mathbb{P}_\theta(\tilde{\theta}\leqslant x) = \mathbb{P}_\theta(X_1\leqslant x,\dots,X_n\leqslant x) = \big(\mathbb{P}_\theta(X_1\leqslant x)\big)^n = \Big(\frac{x}{\theta}\Big)^n,$$

et donc la densité de $\tilde{\theta}$ est donnée par

$$f_{\tilde{\theta}}(x) = \frac{n}{\theta^n} x^{n-1} \mathbf{1}_{[0,\theta]}(x).$$

Par conséquent,

$$\mathbb{E}_\theta[\tilde{\theta}] = \frac{n}{n+1}\theta$$

et $\tilde{\theta}$ est asymptotiquement sans biais. On peut maintenant calculer son risque quadratique :

$$\mathcal{R}_{\tilde{\theta}}(\theta) = \frac{2\theta^2}{(n+1)(n+2)}.$$

Comparons les 2 estimateurs ci-dessus : on voit que $\mathcal{R}_{\tilde{\theta}}(\theta) \geqslant \mathcal{R}_{\hat{\theta}}(\theta)$ pour tout $\theta > 0$ et tout $n \geqslant 1$, l'inégalité étant stricte dès que $n \geqslant 3$. L'estimateur $\tilde{\theta}$ est donc plus performant, malgré son biais. Remarquons qu'on peut facilement corriger le biais en considérant l'estimateur

$$\frac{n+1}{n}\tilde{\theta}.$$

⌟

12.2 Intervalles de confiance

12.2.1 Définition et exemples

Lorsque l'on cherche à estimer un paramètre, il est souvent plus utile de donner un renseignement du type $a \leqslant \theta \leqslant b$ accompagné du degré de confiance que l'on peut avoir en cette affirmation, plutôt qu'une valeur numérique précise sans information quant à sa précision. On dit alors qu'on fournit une **estimation par intervalle** de θ.

On considère comme toujours un n-échantillon de loi $\mathbb{P}_\theta$.

Définition 12.21. *Soit $\alpha \in (0,1)$. Un intervalle $I = I(X_1, \dots, X_n)$ (aléatoire, ne dépendant pas de θ) est appelé **intervalle de confiance pour $f(\theta)$ au niveau $1-\alpha$** si*

$$\forall \theta \in \Theta, \qquad \mathbb{P}_\theta(I \ni f(\theta)) = 1 - \alpha.$$

*$1-\alpha$ est appelé **niveau de confiance** de l'estimation.*

Exemple 12.22. On considère un n-échantillon avec loi $\mathcal{N}(\mu, 1)$. On a vu que la moyenne empirique $\overline{X}_n$ est un estimateur sans biais de μ. On veut construire un intervalle $[T_1, T_2]$, avec $T_1 := \overline{X}_n - a$ et $T_2 := \overline{X}_n + a$ (intervalle symétrique autour de la moyenne empirique). Il suit de l'Exemple 5.15 que $\overline{X}_n \sim \mathcal{N}(\mu, \frac{1}{n})$. Par conséquent, $Z := \sqrt{n}(\overline{X}_n - \mu) \sim \mathcal{N}(0,1)$. On a donc

$$\mathbb{P}_\mu(I \ni \mu) = 1 - \alpha \quad \Leftrightarrow \quad \mathbb{P}_\mu(|\overline{X}_n - \mu| \leqslant a) = \mathbb{P}(|Z| \leqslant a\sqrt{n}) = 1 - \alpha.$$

Pour $\alpha = 10\%$, on trouve que cette dernière identité est satisfaite si $a\sqrt{n} \simeq 1{,}64$. Par conséquent, l'intervalle

$$I = \left[\overline{X}_n - \frac{1{,}64}{\sqrt{n}}, \overline{X}_n + \frac{1{,}64}{\sqrt{n}}\right]$$

est un intervalle de confiance à 90% pour μ. ⌟

Exemple 12.23. On considère un n-échantillon distribué uniformément sur $[0, \theta]$, $\theta > 0$. Manifestement, l'estimateur du maximum de vraisemblance $\hat{\theta} = \max\{X_1, \dots, X_n\}$ satisfait toujours $\hat{\theta} \leqslant \theta$. On peut donc prendre $T_1 = \hat{\theta}$. On cherche T_2 de la forme $C\hat{\theta}$ avec $\mathbb{P}_\theta(C\hat{\theta} \geqslant \theta) = 1 - \alpha$. Dans ce cas,

$$I = [\hat{\theta}, C\hat{\theta}]$$

sera un intervalle de confiance au niveau $1-\alpha$. On a déjà vu que

$$\mathbb{P}_\theta(\hat{\theta} \leqslant x) = \left(\frac{x}{\theta}\right)^n.$$

On a donc

$$\mathbb{P}_\theta(C\hat{\theta} \geqslant \theta) = 1 - \mathbb{P}_\theta(C\hat{\theta} < \theta) = 1 - \left(\frac{1}{C}\right)^n.$$

L'intervalle recherché est donc

$$I = \left[\hat{\theta}, \alpha^{-1/n}\hat{\theta}\right].$$ ⌟

À nouveau, il n'y a pas en général unicité de l'intervalle de confiance à un niveau donné. Dans ce cas, à niveaux de confiance égaux, l'intervalle le plus petit sera considéré le meilleur, puisqu'il donne l'estimation la plus précise.

12.2.2 Intervalles de confiance par excès et asymptotiques

En général, il est suffisant de borner inférieurement la confiance que l'on a dans l'estimation.

Définition 12.24. *Un intervalle $I = I(X_1, \dots, X_n)$ (indépendant de θ) est un intervalle de confiance pour θ au niveau $1-\alpha$* ***par excès*** *si*

$$\forall \theta \in \Theta, \qquad \mathbb{P}_\theta(I \ni \theta) \geqslant 1 - \alpha.$$

Exemple 12.25. Soit $X_1, \dots, X_n$ un n-échantillon. On suppose la variance $\mathrm{Var}[X_1] = \sigma^2$ connue, et on cherche à estimer par intervalle $f(\theta) = \mathbb{E}_\theta[X_1]$. Notant $\overline{X}_n$ la moyenne empirique, il suit de l'inégalité de Bienaymé–Tchebychev (Lemme 2.65) que

$$\mathbb{P}_\theta(|\overline{X}_n - f(\theta)| < \delta) \geqslant 1 - \frac{\sigma^2}{n\delta^2}.$$

On en conclut que

$$I = \left[\overline{X}_n - \frac{\sigma}{\sqrt{n\alpha}}, \overline{X}_n + \frac{\sigma}{\sqrt{n\alpha}}\right]$$

est un intervalle de confiance par excès au niveau $1-\alpha$. ⌟

Une façon efficace de déterminer des intervalles de confiance valables asymptotiquement est d'approximer, via le théorème central limite, la loi de la moyenne empirique par une loi normale.

Définition 12.26. *Pour un n-échantillon $X_1, \dots, X_n$, un intervalle de confiance* ***asymptotique*** *pour θ au niveau $1-\alpha$ est un intervalle $I_n = I_n(X_1, \dots, X_n)$ tel que*

$$\forall \theta \in \Theta, \qquad \lim_{n\to\infty} \mathbb{P}_\theta(I_n \ni \theta) = 1 - \alpha.$$

Un intervalle de confiance ***asymptotique par excès*** *pour θ au niveau $1-\alpha$ est un intervalle $I_n = I_n(X_1, \dots, X_n)$ tel que*

$$\forall \theta \in \Theta, \qquad \lim_{n\to\infty} \mathbb{P}_\theta(I_n \ni \theta) \geqslant 1 - \alpha.$$

Exemple 12.27. On considère un n-échantillon, dont la variance $\sigma^2 = \mathrm{Var}_\theta[X_1]$ est connue. On désire estimer la moyenne $\mu = \mathbb{E}_\theta(X_1)$. On considère la moyenne empirique. Par le théorème central limite,

$$\mathbb{P}_\theta\left(\overline{X}_n \in \left[\mu - \frac{a\sigma}{\sqrt{n}}, \mu + \frac{a\sigma}{\sqrt{n}}\right]\right) \xrightarrow{n\to\infty} \mathbb{P}(Z \in [-a, a]),$$

où $Z \sim \mathcal{N}(0,1)$. Si l'on choisit a tel que $\mathbb{P}(Z \in [-a,a]) = 1-\alpha$, l'intervalle

$$I_n = \left[\overline{X}_n - \frac{a\sigma}{\sqrt{n}}, \overline{X}_n + \frac{a\sigma}{\sqrt{n}}\right]$$

est un intervalle de confiance asymptotique pour μ au niveau $1-\alpha$.

Comme application, considérons la situation suivante : on mesure une grandeur μ. L'incertitude moyenne vaut $0{,}73$. Combien faut-il de mesures pour déterminer μ avec une précision de 10^{-1} ? L'échantillon est formé de n mesures $X_1, \dots, X_n$. On a pour l'espérance $\mathbb{E}_\theta(X_i) = \mu$ et pour l'écart-type $\sigma = 0{,}73$. En prenant comme estimateur la moyenne empirique et un niveau de confiance de 99%, on trouve $a \simeq 2{,}58$ et donc l'intervalle

$$I_n = \Big[\overline{X}_n - \frac{1{,}88}{\sqrt{n}}, \overline{X}_n + \frac{1{,}88}{\sqrt{n}}\Big].$$

On choisit à présent le plus petit n tel que $1{,}88/\sqrt{n} \leqslant 0{,}1$, c'est-à-dire $n \geqslant 355$. ⌟

12.3 Tests d'hypothèses

12.3.1 Un exemple

La garantie d'un constructeur pour ses composants électroniques est de 2 ans. Il peut accepter au plus un taux de 10% de pièces tombant en panne pendant cette période et désire donc s'assurer que $\mathbb{P}_\theta(T \geqslant 2) \geqslant 0{,}9$, où T est le temps de vie de ces composants, de loi supposée exponentielle de paramètre $1/\theta$. Ceci revient à s'assurer que $\theta \geqslant -2/\log(0{,}9) = \theta^\star \simeq 19$. On veut donc déterminer si l'hypothèse $\theta < \theta^\star$ est réaliste, auquel cas il sera nécessaire de revoir la chaîne de fabrication.

À partir d'un n-échantillon, on obtient une estimation $\hat{\theta}_n$ de θ. En se basant sur cette estimation, le constructeur doit prendre sa décision : soit accepter le taux de défaillance actuel, soit remplacer la chaîne de fabrication. Supposons qu'un taux de défaillance supérieur à 10% mette l'entreprise en péril, alors le constructeur acceptera d'investir dans une nouvelle chaîne de fabrication au moindre soupçon que $\theta < \theta^\star$. Il convient donc de minimiser le risque de prédire, à partir de l'échantillon, que $\theta \geqslant \theta^\star$, alors qu'en réalité $\theta < \theta^\star$. Ceci introduit une asymétrie entre l'hypothèse $\theta < \theta^\star$ et son complémentaire. Dans une telle situation, on appelle l'hypothèse cruciale $\theta < \theta^\star$, l'**hypothèse nulle**.

- ▷ L'**erreur de 1re espèce** consiste à rejeter l'hypothèse nulle alors qu'elle est vraie.
- ▷ L'**erreur de 2nde espèce** consiste à ne pas rejeter l'hypothèse nulle alors qu'elle est fausse.

Idéalement, on aimerait minimiser ces deux erreurs, mais ceci n'est pas possible, car elles sont antagonistes : diminuer l'une fait augmenter l'autre.

L'erreur de première espèce est le risque que le constructeur cherche avant tout à minimiser (elle peut mettre son entreprise en danger). Il se fixe donc une probabilité d'erreur α, appelée le **seuil**, correspondant au risque maximal qu'il est prêt à prendre ; on choisit par exemple $\alpha = 5\%$. Supposons qu'il existe z_0 tel que

$$\forall \theta \in [0, \theta^\star), \qquad \mathbb{P}_\theta(\hat{\theta}_n \geqslant z_0) \leqslant 5\%.$$

Dans ce cas, si l'on observe $\hat{\theta}_n \geqslant z_0$, il ne sera pas raisonnable de supposer que $\theta \in (0, \theta^\star]$, puisque l'événement $\hat{\theta}_n \geqslant z_0$ ne se produit alors que dans au plus 5% des cas. Le fabricant rejettera donc l'hypothèse $\theta < \theta^\star$ et aura raison dans au moins 95% des cas. Il estimera donc, avec une **confiance** de 95%, que le pourcentage de pièces qui tomberont en panne avant deux ans est inférieur à 10%.

En revanche, si l'on trouve $\hat{\theta}_n < z_0$, alors il existe un risque que $\theta < \theta^\star$. Dans ce cas, le constructeur ne peut pas rejeter l'hypothèse $\theta < \theta^\star$ et doit donc décider d'investir dans une nouvelle chaîne de fabrication plus sûre.

12.3.2 Procédure de test

On se place dans le cadre d'un n-échantillon $X_1, \dots, X_n$ de loi $\mathbb{P}_\theta$ de paramètre $\theta \in \Theta$ inconnu. Étant donné $\Theta_0 \subset \Theta$, $\varnothing \neq \Theta_0 \neq \Theta$, il s'agit de déterminer si θ appartient à Θ_0 ou si θ appartient à son complémentaire $\Theta_1 := \Theta \setminus \Theta_0$. On dit que l'on teste l'**hypothèse nulle** H_0 : « $\theta \in \Theta_0$ » contre l'**hypothèse alternative** H_1 : « $\theta \in \Theta_1$ ».

Définition 12.28. *Une **région de rejet** est un événement $D \in \sigma(X_1, \dots, X_n)$.*

Définition 12.29. *Soit D une région de rejet, H_0 et H_1 deux hypothèses que l'on teste l'une contre l'autre. Une **procédure de test** consiste à*

1. *rejeter H_0 si D se produit ;*
2. *ne pas rejeter H_0 si D ne se produit pas.*

Définition 12.30. *On dit que le test est **au niveau de risque** α, ou **niveau de confiance** $1-\alpha$, si*

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(D) = \alpha.$$

Définition 12.31. *On appelle **puissance** d'un test la valeur*

$$\inf_{\theta \in \Theta_1} \mathbb{P}_\theta(D) = 1 - \beta.$$

À un niveau de confiance donné $1-\alpha$, on cherche donc à maximiser la puissance, ce qui revient à minimiser la probabilité β de commettre une erreur de seconde espèce. Ce critère permet de comparer des tests.

Définition 12.32. *Une hypothèse H est dite **simple** si l'ensemble Θ correspondant est réduit à un seul élément, sinon elle est dite **composite**.*

Exemple 12.33. Supposons que $I = I(X_1, \dots, X_n)$ soit un intervalle de confiance pour θ au niveau de confiance $1-\alpha$. On considère l'hypothèse nulle (simple) H_0 : « $\theta = \theta_0$ » et l'hypothèse alternative (composite) H_1 : « $\theta \neq \theta_0$ ». Alors $D = \{I \not\ni \theta_0\}$ fournit un test de H_0 contre H_1 au niveau de risque α, puisque

$$\mathbb{P}_{\theta_0}(I \not\ni \theta_0) = \alpha. \qquad \lrcorner$$

12.3.3 Cas gaussien

On considère un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$ avec σ^2 connu.

Test de moyenne à variance connue

Test de « $\mu = \mu_0$ » contre « $\mu \neq \mu_0$ ». Soit $\overline{X}_n$ la moyenne empirique (de loi $\mathcal{N}(\mu, \sigma^2/n)$) ; on prend pour région de rejet

$$D \coloneqq \{|\overline{X}_n - \mu_0| \geqslant C\}.$$

On veut un niveau de risque de 5%, c'est-à-dire

$$\mathbb{P}_{\mu_0}(|\overline{X}_n - \mu_0| \geqslant C) = 0{,}05$$

ce qui conduit à prendre $C \simeq 1{,}96\sigma/\sqrt{n}$.

Test de « $\mu \leqslant \mu_0$ » contre « $\mu > \mu_0$ ». Cette fois, on prend pour région de rejet

$$D \coloneqq \{\overline{X}_n > C\}.$$

On veut un niveau de risque de 5%, c'est-à-dire

$$\sup_{\mu \leqslant \mu_0} \mathbb{P}_\mu(D) = \sup_{\mu \leqslant \mu_0} \mathbb{P}\Big(\frac{\sigma}{\sqrt{n}} Z > C - \mu\Big) = 0{,}05,$$

où $Z \sim \mathcal{N}(0,1)$. La borne supérieure est atteinte pour $\mu = \mu_0$, ce qui conduit à choisir $C \simeq \mu_0 + 1{,}64\sigma/\sqrt{n}$.

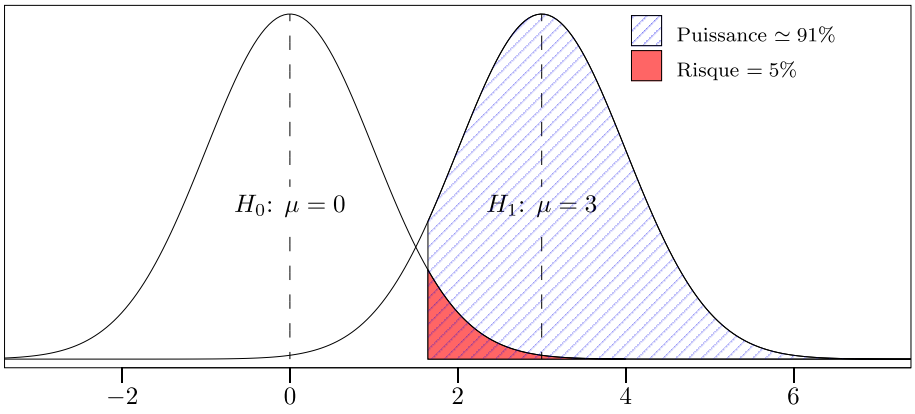


FIGURE 12.1: Test de deux hypothèses simples.

Test d'égalité de moyenne de 2 échantillons de variance connue

On considère un n-échantillon $X_1, \dots, X_n$ de loi $\mathcal{N}(\mu, \sigma^2)$ et un m-échantillon (indépendant du premier) $Y_1, \dots, Y_m$ de loi $\mathcal{N}(\nu, \tau^2)$, avec σ^2, τ^2 connus. On veut tester « $\mu = \nu$ » contre « $\mu \neq \nu$ ».

Ce problème se ramène au premier exemple ci-dessus : on estime $\mu - \nu$ par $\overline{X}_n - \overline{Y}_m \sim \mathcal{N}(\mu - \nu, \frac{\sigma^2}{n} + \frac{\tau^2}{m})$ et on teste « $\mu - \nu = 0$ » contre « $\mu - \nu \neq 0$ ».

12.3.4 Tests d'hypothèses simples

On considère un n-échantillon de loi $\mathbb{P}_\theta$. On va tester H_0 : « $\theta = \theta_0$ » contre H_1 : « $\theta = \theta_1$ ». Nous allons faire cela en comparant les fonctions de vraisemblance $L_{x_1,\dots,x_n}(\theta_0)$ et $L_{x_1,\dots,x_n}(\theta_1)$. C'est ce qu'on appelle le **test de Neyman**[1]**–Pearson**[2]. L'objet central est le **rapport de vraisemblance**,

$$R_{x_1,\dots,x_n}(\theta_0, \theta_1) \coloneqq \frac{L_{x_1,\dots,x_n}(\theta_1)}{L_{x_1,\dots,x_n}(\theta_0)}.$$

On prend pour région de rejet

$$D \coloneqq \{R_{x_1,\dots,x_n}(\theta_0, \theta_1) > C\},$$

où C est une constante à déterminer en fonction du risque choisi. Pour un test avec un risque de 5%, on fixe C de sorte que

$$\mathbb{P}_{\theta_0}(D) = 5\%.$$

Exemple 12.34. Une personne possède deux pièces : l'une est équilibrée, l'autre donne à « face » une probabilité double de celle de « pile ». Elle choisit une de ces deux pièces et effectue 100 lancers. Elle obtient $F = 60$ « face ». Comment déterminer quelle pièce a été utilisée ?

Le modèle est clair : on a un échantillon de taille $n = 100$ suivant une loi de Bernoulli de paramètre p, avec $p \in \{\frac{1}{2}, \frac{2}{3}\}$. On désire tester H_0 : « $p = \frac{1}{2}$ » contre $H_1 : p = \frac{2}{3}$, qui sont deux hypothèses simples.

La fonction de vraisemblance associée à une réalisation de ces n variables aléatoires de Bernoulli avec f succès est

$$p^f (1-p)^{n-f} = (1-p)^n \Big(\frac{p}{1-p}\Big)^f.$$

Le rapport de vraisemblance est donc donné, dans la situation étudiée ici, par

$$R = \left(\frac{1 - \frac{2}{3}}{1 - \frac{1}{2}}\right)^n \left(\frac{\frac{2}{3}/(1 - \frac{2}{3})}{\frac{1}{2}/(1 - \frac{1}{2})}\right)^f = \big(\tfrac{2}{3}\big)^n 2^f.$$

1. Jerzy Neyman (1894, Bendery – 1981, Berkeley), statisticien polonais ; un des fondateurs de la statistique moderne.
2. Egon Sharpe Pearson (1895, Hampstead – 1980, London), statisticien anglais. Fils du célèbre statisticien Karl Pearson.

Il s'agit d'une fonction monotone de f, donc prendre une région de rejet de la forme

$$D \coloneqq \{R > C\}$$

revient à prendre une région

$$D' \coloneqq \{F > C'\},$$

avec C' tel que

$$\mathbb{P}_{\frac{1}{2}}(F > C') = 10\%,$$

pour un niveau de risque de 10%. On peut à présent déterminer C' par simple calcul. Plutôt que d'en déterminer la valeur exacte, nous allons utiliser le théorème central limite afin d'approximer $(F - 50)/5$ par une variable aléatoire $Z \sim \mathcal{N}(0, 1)$. On obtient ainsi

$$\mathbb{P}_{\frac{1}{2}}(F > C') \simeq \mathbb{P}(Z > (C' - 50)/5).$$

Par conséquent, on peut prendre $C' \simeq 56{,}4$.

Puisque, pour notre échantillon, $F = 60$, on est conduit à rejeter H_0. (Remarquons que ce test, de par sa nature, privilégie H_0 par rapport à H_1.) ⌟

On peut montrer que lorsque celui-ci est bien défini, aucun test à un niveau de confiance donné n'est plus puissant que le test ci-dessus.

Lemme 12.35 (Lemme de Neyman–Pearson)**.** *On considère deux hypothèses simples H_0 : « $\theta = \theta_0$ » contre H_1 : « $\theta = \theta_1$ » et on suppose que les lois $\mathbb{P}_{\theta_0}$ et $\mathbb{P}_{\theta_1}$ du n-échantillon sous ces deux hypothèses possèdent les densités de probabilité f_{θ_0} et f_{θ_1} respectivement. Soit $\alpha \in (0, 1)$ et*

$$D \coloneqq \big\{(x_1, \dots, x_n) \,\big|\, \prod_{i=1}^{n} f_{\theta_1}(x_i) > C \prod_{i=1}^{n} f_{\theta_0}(x_i)\big\},$$

où C est choisie de sorte que $\mathbb{P}_{\theta_0}(D) = \alpha$. Alors, pour toute autre région de rejet B telle que $\mathbb{P}_{\theta_0}(B) = \alpha$, on a

$$\mathbb{P}_{\theta_1}(B) \leqslant \mathbb{P}_{\theta_1}(D),$$

avec inégalité stricte si $\mathbb{P}_{\theta_1}(D \setminus B) > 0$.

Démonstration. Notons $\boldsymbol{x} = (x_1, \dots, x_n)$, $\mathrm{d}\boldsymbol{x} = \mathrm{d}x_1 \cdots \mathrm{d}x_n$, et $f(\boldsymbol{x}) = f(x_1) \cdots f(x_n)$. On a

$$\int_{D \setminus B} f_{\theta_0}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \alpha - \int_{D \cap B} f_{\theta_0}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{B \setminus D} f_{\theta_0}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$

puisque $\mathbb{P}(D) = \mathbb{P}(B) = \alpha$. D'autre part, puisque $D \setminus B \subseteq D$ et $B \setminus D \subseteq D^{\mathrm{c}}$, on déduit de l'identité précédente que

$$\int_{D \setminus B} f_{\theta_1}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \geqslant C \int_{D \setminus B} f_{\theta_0}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = C \int_{B \setminus D} f_{\theta_0}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \geqslant \int_{B \setminus D} f_{\theta_1}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

(La première inégalité est stricte si $\mathbb{P}_{\theta_1}(D \setminus B) > 0$.) On a donc bien

$$\mathbb{P}_{\theta_1}(D) = \mathbb{P}_{\theta_1}(D \setminus B) + \mathbb{P}_{\theta_1}(D \cap B) \geqslant \mathbb{P}_{\theta_1}(B \setminus D) + \mathbb{P}_{\theta_1}(D \cap B) = \mathbb{P}_{\theta_1}(B). \qquad \square$$

Remarque 12.36. *Dans le cas d'un échantillon de loi discrète, un résultat similaire reste valide. Deux nouvelles difficultés font toutefois leur apparition : d'une part, il n'est pas toujours possible de trouver C de façon à obtenir un niveau α donné, puisque la fonction de répartition est discontinue ; d'autre part, l'ensemble $\{(x_1, \dots, x_n) \,|\, \mathbb{P}_{\theta_1}(X_1 = x_1, \dots, X_n = x_n) = C\mathbb{P}_{\theta_0}(X_1 = x_1, \dots, X_n = x_n)\}$ n'a plus nécessairement probabilité nulle. Une manière de résoudre simultanément ces deux problèmes est d'utiliser la procédure suivante. Soit $R_{x_1,\dots,x_n}(\theta_0, \theta_1)$ le rapport de vraisemblance. Alors : si $R > C$ on rejette H_0 ; si $R < C$, on ne rejette pas H_0 ; si $R = C$, on rejette H_0 avec probabilité ρ. Ici ρ et C sont choisis tels que $\mathbb{P}_{\theta_0}(D > C) + \rho\mathbb{P}_{\theta_0}(D = C) = \alpha$.*

12.3.5 Tests du χ^2

Jusqu'à présent, on a toujours supposé connue la loi de l'échantillon, le problème se réduisant donc à estimer ses paramètres. C'est ce qu'on appelle faire un test **paramétrique**. Nous allons à présent considérer une expérience aléatoire dont la loi n'est pas connue. On parle alors de test **non paramétrique**.

Le test d'adéquation du χ^2

Les **tests d'adéquation**, ou **tests d'ajustement**, ont pour objet de déterminer à partir d'un échantillon si une variable aléatoire suit ou non une certaine loi. Parmi ces tests, nécessairement non paramétriques, l'un des plus connus et des plus utilisés est le test du χ^2 (Khi-deux).

Considérons donc une expérience aléatoire dont les résultats peuvent être répartis en k classes, avec les probabilités $p_1, \dots, p_k$ ($p_1 + \cdots + p_k = 1$). Ayant réalisé n fois cette expérience, on obtient un vecteur aléatoire $(N_n(1), \dots, N_n(k))$, où $N_n(j)$ est le nombre d'occurrences de la classe j. Par définition, ce vecteur suit une **loi multinomiale** de paramètres $(p_1, \dots, p_k, n)$, c'est-à-dire

$$\mathbb{P}(N_n(1) = n_1, \dots, N_n(k) = n_k) = \frac{n!}{n_1! \cdots n_k!}\, p_1^{n_1} \cdots p_k^{n_k}.$$

Soit $q_1, \dots, q_k \in (0,1)$ tels que $\sum_{i=1}^k q_i = 1$.

On veut tester $H_0 : \forall i,\ p_i = q_i$ contre $H_1 : \exists j,\ p_j \neq q_j$.

q nous donne donc les probabilités de chacune des classes sous l'hypothèse nulle et on est amené à comparer ces dernières avec les fréquences empiriques $N_n(j)/n$. On a ainsi transformé un test non-paramétrique en un test paramétrique portant sur les paramètres d'une loi multinomiale.

Afin de construire notre région de rejet, on introduit la statistique

$$Z_n := \sum_{j=1}^k \frac{(N_n(j) - nq_j)^2}{nq_j} = n \sum_{j=1}^k \frac{\left(\frac{N_n(j)}{n} - q_j\right)^2}{q_j}.$$

Z_n mesure donc les écarts entre les fréquences empiriques et les fréquences théoriques, proprement normalisés. Le test repose sur le résultat suivant, que nous admettrons.

Lemme 12.37. *Soit $(N_1, \dots, N_k)$ un vecteur aléatoire suivant une loi multinomiale de paramètres $(p_1, \dots, p_k, n)$. Alors la variable aléatoire*

$$\sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i}$$

suit asymptotiquement la loi du χ^2 à $k-1$ degrés de liberté, χ^2_{k-1}, dont la densité est

$$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}\, \mathbf{1}_{[0,\infty)}(x).$$

Remarque 12.38. *La raison pour laquelle on a $k-1$ degrés de liberté et non k est qu'une contrainte lie les N_i : $N_1 + \cdots + N_k = n$.*

On déduit du Lemme 12.37 que, sous H_0, Z_n suit asymptotiquement une loi χ^2_{k-1}.

D'autre part, sous H_1, il existe $j \in \{1, \dots, k\}$ tel que

$$\lim_{n\to\infty} \left(\frac{N_n(j)}{n} - q_j\right)^2 = (p_j - q_j)^2 > 0,$$

ce qui implique que $Z_n \xrightarrow{\text{p.s.}} +\infty$.

On peut donc prendre une région de rejet de la forme

$$D \coloneqq \{Z_n > C\},$$

en choisissant C de sorte que, si $Z \sim \chi^2_{k-1}$,

$$\lim_{n\to\infty} \mathbb{P}_q(Z_n > C) = \mathbb{P}(Z > C) = \alpha.$$

Remarque 12.39. *Il est important de réaliser qu'il s'agit d'une approximation asymptotique. Pour qu'elle soit applicable en pratique, on exige généralement que les effectifs théoriques nq_j soient supérieurs à 5.*

Exemple 12.40. Le 8 février 1865, le moine autrichien Gregor Mendel[3] publie ses « *Expériences sur les plantes hybrides* » où il expose les lois de l'hérédité qui portent aujourd'hui son nom. Ces lois, il les a découvertes en étudiant la transmission des caractères biologiques chez les petits pois. En particulier, il s'est intéressé aux caractères « couleur » et « forme ». Ces caractères sont tous deux codés par un gène avec deux allèles. Le caractère « couleur » est soit C (jaune), dominant, soit c (vert), récessif. Le caractère « forme » est soit R (rond), dominant, soit r (ridé), récessif. En croisant deux individus de génotype $CcRr$, il y a 16 génotypes équiprobables pour les descendants, et les phénotypes devraient être distribués de la façon suivante : pois jaunes et ronds avec une fréquence $9/16$, jaunes et ridés avec une fréquence $3/16$, verts et ronds avec une fréquence $3/16$, et verts et ridés avec une fréquence $1/16$. Le tableau suivant contient les résultats de Mendel :

	Jaune, rond	Jaune, ridé	Vert, rond	Vert, ridé
Effectifs	315	101	108	32
Fréquence empirique	315/556	101/556	108/556	32/556
Fréquence théorique	9/16	3/16	3/16	1/16

On désire alors tester l'hypothèse H_0 : les fréquences d'apparition des différents caractères sont bien données par les prédictions de Mendel, contre l'hypothèse alternative. C'est un exemple typique d'utilisation du test d'adéquation du χ^2. On obtient

$$Z_{556} = \frac{\left(315 - 556 \cdot \frac{9}{16}\right)^2}{556 \cdot \frac{9}{16}} + \frac{\left(101 - 556 \cdot \frac{3}{16}\right)^2}{556 \cdot \frac{3}{16}} + \frac{\left(108 - 556 \cdot \frac{3}{16}\right)^2}{556 \cdot \frac{3}{16}} + \frac{\left(32 - 556 \cdot \frac{1}{16}\right)^2}{556 \cdot \frac{1}{16}} \simeq 0{,}47.$$

Soit $Z \sim \chi^2_3$. Pour un seuil de 5%, on obtient que $\mathbb{P}(Z > C) = 0{,}05$ pour $C \simeq 7{,}82$. Puisque $0{,}47 < 7{,}82$, les observations sont compatibles avec l'hypothèse nulle.

En fait, les résultats sont trop bons et il est généralement admis aujourd'hui que Mendel a dû « améliorer » ses données pour les faire mieux coller aux prédictions. ⌟

Le test d'indépendance du χ^2

Nous allons à présent brièvement décrire comment des idées analogues peuvent être utilisées afin de déterminer si deux propriétés sont indépendantes ou liées. Nous nous contenterons de le faire sur un exemple.

On désire déterminer si la couleur des cheveux et celle des yeux sont indépendantes ou liées. Nous nous baserons sur les données suivantes.

3. Johann Gregor Mendel (1822, Heinzendorf – 1884, Brünn), moine et botaniste Autrichien. Il est communément reconnu comme le père fondateur de la génétique.

	ch. blonds	ch. bruns	ch. roux	ch. noirs	total	fréquence
yeux bleus	25	9	7	3	44	44/124
yeux gris	13	17	7	10	47	47/124
yeux marrons	7	13	5	8	33	33/124
total	45	39	19	21	124	
fréquence	45/124	39/124	19/124	21/124		

On veut donc tester l'hypothèse nulle H_0 : ces deux caractères sont indépendants contre l'hypothèse alternative.

Sous H_0, les fréquences d'observations d'une paire donnée de caractères devraient être données par le produit des fréquences de chacun des caractères. Bien entendu, on ne connaît pas ces fréquences, donc on utilise les fréquences empiriques. Par exemple, la fréquence théorique pour « cheveux bruns, yeux bleus » est de $(44/124)(39/124)$ et doit être comparée avec la fréquence empirique $9/124$. Ce problème est donc tout à fait similaire au précédent. La seule subtilité est que le fait que l'on ait eu à estimer $(3-1)+(4-1)=5$ fréquences empiriquement conduit à considérer une loi du χ^2 à $12-1-5=6$ degrés de liberté.

En procédant comme précédemment, on arrive à la conclusion qu'avec un seuil de 5%, l'hypothèse nulle (d'indépendance) doit être rejetée.

https://xkcd.com/892

A Rappels et compléments

A.1 Quelques rappels de théorie de la mesure

Définition A.1. *Soit E un ensemble. Une **tribu** sur E est une collection $\mathcal{E}$ de parties de E satisfaisant*

(i) $E \in \mathcal{E}$,

(ii) $A \in \mathcal{E} \implies E \setminus A \in \mathcal{E}$ *et*

(iii) pour toute suite $(A_n)_{n\geqslant 1} \subset \mathcal{E}$, $\bigcup_{n\geqslant 1} A_n \in \mathcal{E}$.

*Une partie $A \in \mathcal{E}$ est dite **$\mathcal{E}$-mesurable** et $(E, \mathcal{E})$ est un **espace mesurable**.*

Définition A.2. *Soit $\mathcal{C} \subset \mathscr{P}(E)$. Alors,*

$$\sigma(\mathcal{C}) \coloneqq \bigcap_{\substack{\mathcal{E}\text{ tribu sur } E \\ \mathcal{E}\supset\mathcal{C}}} \mathcal{E}$$

*existe et est une tribu, appelée la **tribu engendrée par $\mathcal{C}$**.*

Définition A.3. *Soit E un espace topologique et $\mathcal{O}$ les ouverts de E. La tribu*

$$\mathcal{B}(E) \coloneqq \sigma(\mathcal{O})$$

*est appelée **la tribu borélienne sur E**.*

Définition A.4. *Soit $(E_1, \mathcal{E}_1)$ et $(E_2, \mathcal{E}_2)$ deux espaces mesurables. La tribu*

$$\mathcal{E}_1 \otimes \mathcal{E}_2 \coloneqq \sigma(\{A_1 \times A_2 \,|\, A_1 \in \mathcal{E}_1,\ A_2 \in \mathcal{E}_2\})$$

*est appelée la **tribu produit** sur $E_1 \times E_2$.*

Définition A.5. *Une **mesure** (positive) sur un espace mesurable $(E, \mathcal{E})$ est une fonction $\mu : \mathcal{E} \to [0, \infty]$ telle que*

(i) $\mu(\varnothing) = 0$ *et*

(ii) pour toute suite $(A_n)_{n\geqslant 1} \subset \mathcal{E}$ *de parties deux à deux disjointes de* $\mathcal{E}$, $\mu(\bigcup_{n\geqslant 1} A_n) = \sum_{n\geqslant 1} \mu(A_n)$.

Définition A.6. *Soit E un ensemble fini ou dénombrable, muni de la tribu $\mathcal{E} := \mathscr{P}(E)$. La mesure*

$$\mu(A) := |A|,$$

où $|A|$ dénote le nombre d'éléments de A, est appelée la ***mesure de comptage*** *sur E.*

Définition A.7. *Soit $(E, \mathcal{E})$ un espace mesurable et $x \in E$. La mesure*

$$\delta_x(A) := \begin{cases} 1 & \text{si } A \ni x \\ 0 & \text{si } A \not\ni x \end{cases}$$

est appelée la ***mesure de Dirac*** *en x.*

Définition A.8. *La* ***mesure de Lebesgue*** *sur $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ est l'unique mesure λ_n telle que*

$$\lambda((a_1, b_1) \times \cdots \times (a_n, b_n)) = \prod_{i=1}^{n} (b_i - a_i),$$

pour tout réels $a_k < b_k$, $k = 1, \ldots, n$.

Définition A.9. *Soit $(E, \mathcal{E})$ et $(F, \mathcal{F})$ deux espaces mesurables. Une fonction $f : E \to F$ est $(\mathcal{E}, \mathcal{F})$-**mesurable** si*

$$\forall A \in \mathcal{F}, \qquad f^{-1}(A) \in \mathcal{E}.$$

Définition A.10. *Soit $(E, \mathcal{E})$ et $(F, \mathcal{F})$ deux espaces mesurables, $f : E \to F$ une fonction $(\mathcal{E}, \mathcal{F})$-**mesurable** et μ une mesure sur $(E, \mathcal{E})$. La* ***mesure image*** *$f_*\mu$ est la mesure sur $(F, \mathcal{F})$ définie par*

$$\forall A \in \mathcal{F}, \qquad f_*\mu(A) := \mu(f^{-1}(A)).$$

Définition A.11. *Soit μ une mesure sur $(E, \mathcal{E})$ et $f : E \to [0, +\infty]$. On note*

$$\int f \, \mathrm{d}\mu \equiv \int f(x) \, \mathrm{d}\mu(x) \in [0, +\infty]$$

*l'**intégrale de Lebesgue de f par rapport à μ**.*

Définition A.12. *La fonction mesurable $f : E \to \mathbb{R}$ est* ***intégrable*** *si $\int |f| \, \mathrm{d}\mu < \infty$. Dans ce cas,*

$$\int f \, \mathrm{d}\mu = \int f_+ \, \mathrm{d}\mu - \int f_- \, \mathrm{d}\mu,$$

où $f_+(x) := \max\{f(x), 0\}$ et $f_-(x) := \max\{-f(x), 0\}$.

Définition A.13. *Soit μ, ν deux mesures σ-finies sur $(E_1, \mathcal{E}_1)$ et $(E_2, \mathcal{E}_2)$ respectivement. La* ***mesure produit*** *$\mu_1 \otimes \mu_2$ est l'unique mesure sur $(E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2)$ telle que*

$$\forall A_1 \in \mathcal{E}_1, A_2 \in \mathcal{E}_2, \qquad \mu_1 \otimes \mu_2(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2).$$

Théorème A.14 (Fubini[1]–Tonelli[2])**.** *Soit $(E_1, \mathcal{E}_1, \mu_1)$ et $(E_2, \mathcal{E}_2, \mu_2)$ deux espaces de mesure et $f : E \times F \to [0, +\infty]$ mesurable. Alors,*

$$\int_{E_1 \times E_2} f \, \mathrm{d}(\mu_1 \otimes \mu_2) = \int_{E_1} \left[\int_{E_2} f \, \mathrm{d}\mu_2 \right] \mathrm{d}\mu_1 = \int_{E_2} \left[\int_{E_1} f \, \mathrm{d}\mu_1 \right] \mathrm{d}\mu_2.$$

Théorème A.15 (Convergence monotone). *Soit $(E, \mathcal{E}, \mu)$ un espace de mesure. Pour toute suite croissante $(f_n)_{n\geqslant 1}$ de fonctions mesurables $f_n : E \to [0, +\infty]$, leur limite ponctuelle est mesurable et*

$$\int \lim_{n\to\infty} f_n \,\mathrm{d}\mu = \lim_{n\to\infty} \int f_n \,\mathrm{d}\mu.$$

Théorème A.16 (Convergence dominée). *Soit $(E, \mathcal{E}, \mu)$ un espace de mesure. Soit $(f_n)_{n\geqslant 1}$ une suite de fonctions mesurables $f_n : E \to \mathbb{R}$ telle que*

- ▷ *f_n converge vers f ponctuellement,*
- ▷ *il existe une fonction intégrable g telle que*

$$\forall n \in \mathbb{N}, \forall x \in E, \qquad |f_n(x)| \leqslant g(x).$$

Alors, f est intégrable et

$$\int \lim_{n\to\infty} f_n \,\mathrm{d}\mu = \lim_{n\to\infty} \int f_n \,\mathrm{d}\mu.$$

Théorème A.17 (Lemme de Fatou). *Soit $(E, \mathcal{E}, \mu)$ un espace de mesure. Pour toute suite $(f_n)_{n\geqslant 1}$ de fonctions mesurables $f_n : E \to [0, +\infty]$, la limite inférieure de la suite est mesurable et*

$$\int \liminf_{n\to\infty} f_n \,\mathrm{d}\mu \leqslant \lim_{n\to\infty} \int f_n \,\mathrm{d}\mu.$$

A.2 Compléments divers

A.2.1 Formule du binôme généralisée

Soit $\alpha \in \mathbb{R}$ et $k \in \mathbb{N}$. Le coefficient binomial $\binom{\alpha}{k}$ est défini par

$$\binom{\alpha}{k} := \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!}.$$

On a alors la généralisation suivante de la formule du binôme de Newton[3] (pourquoi retrouve-t-on bien la formule usuelle lorsque $\alpha \in \mathbb{N}$?).

Lemme A.18. *Soit $x, y, \alpha \in \mathbb{R}$. Alors,*

$$(x+y)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^{\alpha-k} y^k,$$

si l'une des conditions suivantes est vérifiée :

1. *$|y/x| < 1$ et $\alpha \in \mathbb{R}$;*
2. *$|y/x| = 1$ et $\alpha \geqslant 0$;*
3. *$y/x = 1$ et $\alpha > -1$.*

Démonstration. En écrivant $(x+y)^\alpha = x^\alpha(1+\frac{y}{x})^\alpha$, on voit qu'il suffit de considérer le cas $x = 1$. Il suffit alors de développer $(1+y)^\alpha$ en série de Taylor autour de $y = 0$ et de vérifier que chacune des conditions données ci-dessus assure la convergence de la série. □

1. Guido Fubini Ghiron (Venise, 1879 – New York, 1943), mathématicien italien.
2. Leonida Tonelli (1885, Gallipoli – 1946, Pise), mathématicien italien
3. Sir Isaac Newton (1643, Woolsthorpe-by-Colsterworth – 1727, Londres), philosophe, mathématicien, physicien, alchimiste, astronome et théologien anglais.

A.2.2 Formule de Stirling

Il se révèle souvent utile, dans de nombreux problèmes de nature combinatoire, d'avoir de bonnes approximations pour $n!$ lorsque n est grand. Le résultat suivant est essentiellement dû à Stirling[4].

Lemme A.19. *Pour tout $n \in \mathbb{N}^*$, on a*

$$e^{1/(12n+1)}\, n^n e^{-n}\sqrt{2\pi n} \leqslant n! \leqslant e^{1/(12n)}\, n^n e^{-n}\sqrt{2\pi n}.$$

Démonstration. Une version de ce résultat sera démontrée en exercice. □

A.2.3 Bornes sur les coefficients binomiaux

On peut extraire le comportement asymptotique du coefficient binomial $\binom{n}{k}$ à l'aide de la formule de Stirling. Il est toutefois souvent utile de disposer de bornes grossières plus simples sur cette quantité.

Lemme A.20. *Soient $n \in \mathbb{N}^*$ et $k \in \{0, \dots, n\}$. Alors,*

$$\left(\frac{n}{k}\right)^k \leqslant \binom{n}{k} \leqslant \frac{n^k}{k!} \leqslant \left(\frac{en}{k}\right)^k.$$

Démonstration. La première inégalité suit immédiatement de

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} = \prod_{i=0}^{k-1} \frac{n-i}{k-i},$$

puisque $(n-i)/(k-i) \geqslant n/k$ pour tout $0 \leqslant i < k$. La seconde inégalité est montrée de façon similaire, en utilisant $n(n-1)\cdots(n-k+1) \leqslant n^k$. Finalement, la dernière suit de

$$\left(\frac{en}{k}\right)^k = \frac{n^k}{k^k} \sum_{\ell=0}^{\infty} \frac{k^\ell}{\ell!} \geqslant \frac{n^k}{k^k}\frac{k^k}{k!} = \frac{n^k}{k!}.$$

□

4. James Stirling (1692, Garden – 1770, Leadhills), mathématicien britannique.

Notations et conventions

Conventions

$A \subset B$ A est un sous-ensemble (pas nécessairement strict) de B

$A \subsetneq B$ A est un sous-ensemble strict de B

$\prod_{i \in \varnothing} a_i \coloneqq 1$

$\sum_{i \in \varnothing} a_i \coloneqq 0$

$\inf \varnothing \coloneqq +\infty$

$\sum_{n \in \mathbb{N}} a_n 0^n \coloneqq a_0$

$\mathbb{P}(A \mid B)\, \mathbb{P}(B) \coloneqq 0$ lorsque $\mathbb{P}(B) = 0$

Notations

$\mathscr{P}(A)$ ensemble des parties de A

$\overline{\mathbb{N}}$ $\mathbb{N} \cup \{+\infty\}$

$\mathcal{B}(\mathbb{R}^d)$ l'ensemble des boréliens de $\mathbb{R}^d$

$\mathcal{B}$ notation alternative pour les boréliens de $\mathbb{R}$

$|A|$ nombre d'éléments de l'ensemble A

λ_d mesure de Lebesgue sur $\mathbb{R}^d$

λ notation alternative pour la mesure de Lebesgue sur $\mathbb{R}$

δ_i masse de Dirac en i

δ_{ij} symbole de Kronecker

$a \vee b$ maximum entre a et b

$a \wedge b$ minimum entre a et b

$\mathbf{1}_A$ l'indicatrice de A : $\mathbf{1}_A(x) = 1$ si $x \in A$ et $\mathbf{1}_A(x) = 0$ sinon

Index

F

G

H

I

L

M

U

V