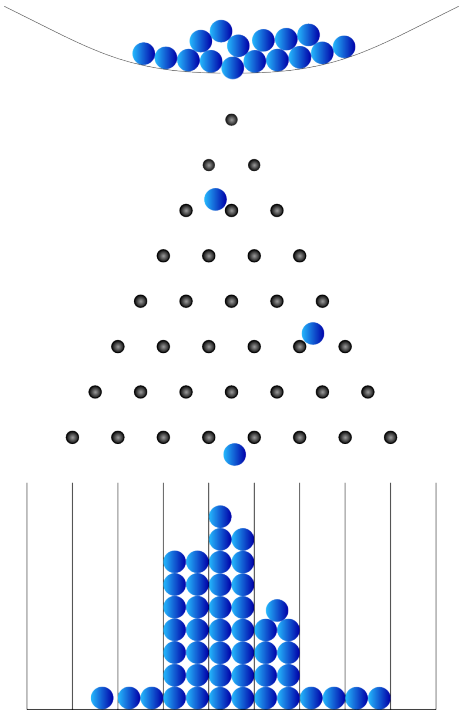
UNIVERSITÉ DE GENÈVE

Probabilités et Statistique

Y. Velenik



— Version préliminaire du 26 octobre 2016 —

Dernière version téléchargeable à l'adresse
http://www.unige.ch/math/folks/velenik/cours.html

Table des matières

Table des matières **3**

0 Introduction **5**
0.1 Modélisation des phénomènes aléatoires 6

I Espaces de probabilité discrets **11**

1 Probabilité, indépendance **13**
1.1 Mesures de probabilité 13
1.2 Quelques résultats combinatoires 16
1.3 Probabilité conditionnelle, formule de Bayes 22
1.4 Indépendance 26
1.5 Expériences répétées 29

2 Variables aléatoires discrètes **31**
2.1 Variables aléatoires discrètes et leurs lois 31
2.2 Indépendance de variables aléatoires 36
2.3 Vecteurs aléatoires discrets 37
2.4 Espérance, variance, covariance et moments 39

3 Marche aléatoire simple sur $\mathbb{Z}$ **55**
3.1 Description du processus 55
3.2 Quelques propriétés importantes 57
3.3 Le premier retour au point de départ 58
3.4 La loi de l'arc-sinus pour la dernière visite en 0 61
3.5 La loi de l'arc-sinus pour les temps de séjour 62

4 Fonctions génératrices **65**
4.1 Définition, propriétés 65
4.2 Application aux processus de branchement 69
4.3 Application à la marche aléatoire simple sur $\mathbb{Z}$ 72
4.4 Fonction génératrice conjointe 75

II Espaces de probabilité généraux **77**

5 Approche axiomatique **79**

5.1 Construction d'espaces de probabilité . . . 79
5.2 Variables aléatoires . . . 85
5.3 Indépendance . . . 88
5.4 Espérance . . . 89
5.5 Variables aléatoires à densité . . . 91
5.6 Processus en temps discret. . . . 105

6 Fonctions caractéristiques 109
6.1 Définition et propriétés élémentaires . . . 109
6.2 Théorèmes d'inversion et de continuité . . . 111
6.3 Quelques exemples classiques . . . 113

7 Théorèmes limites 117
7.1 Un point technique . . . 117
7.2 Quelques outils . . . 118
7.3 Modes de convergence . . . 120
7.4 La loi des grands nombres . . . 122
7.5 Le Théorème Central Limite . . . 126
7.6 La loi 0-1 de Kolmogorov . . . 128

8 Retour aux marches aléatoires 131
8.1 Compléments concernant la marche sur $\mathbb{Z}$ . . . 131
8.2 Marche aléatoire simple sur $\mathbb{Z}^d$ . . . 134

9 Les chaînes de Markov 141
9.1 Définition et exemples . . . 141
9.2 Chaînes de Markov absorbantes . . . 146
9.3 Chaînes de Markov irréductibles . . . 150

10 Modèle de percolation 159
10.1 Définition . . . 159
10.2 Transition de phase . . . 159

11 Le processus de Poisson 165
11.1 Définition et propriétés élémentaires . . . 165
11.2 Autres propriétés . . . 172

12 Introduction à la statistique 183
12.1 Estimateurs . . . 183
12.2 Intervalles de confiance . . . 189
12.3 Tests d'hypothèses . . . 193

Index 201

Chapitre 0

Introduction

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, par exemple en fiabilité, ou dans les assurances et la finance. 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, elle a acquis une place importante en mathématiques de par son intérêt intrinsèque, et, de par sa versatilité, possède un des spectres les plus larges en mathématiques, 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 SMS envoyés la nuit du 31 décembre ; 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 mathématiques permettant de modéliser de tels phénomènes aléatoires a occupé les scientifiques depuis plusieurs siècles. Motivés 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. Les premières publications sur le sujet remontent à G. Cardano [1] avec son livre *Liber De Ludo Aleæ* (publié en 1663, mais probablement achevé

1. Girolamo Cardano (1501, Pavie - 1576, Rome), parfois connu sous le nom de Jérôme Cardan, mathématicien, philosophe et médecin italien. Féru d'astrologie, on dit qu'il avait prévu le jour de sa mort, mais que celle-ci ne semblant pas vouloir se produire d'elle-même, il se suicida afin de rendre sa prédiction correcte.

en 1563), ainsi qu'à Kepler [2] et Galilée [3]. Toutefois, il est généralement admis que la théorie des probabilités débute réellement avec les travaux de Pascal [4] et de Fermat [5]. La théorie fut ensuite développée par de nombreuses personnes, dont Huygens [6], J. Bernoulli [7], de Moivre [8], D. Bernoulli [9], Euler [10], Gauss [11] et Laplace [12]. La théorie moderne des probabilités est fondée sur l'approche axiomatique de Kolmogorov [13], basée sur la théorie de la mesure de Borel [14] et Lebesgue [15]. Grâce à cette approche, la théorie a alors connu un développement très rapide tout au long du XX$^{\text{ème}}$ 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 de dé : $\Omega = \{(a_k)_{k\geq 1} \,:\, a_k \in \{1,\ldots,6\}, \forall k \geq 1\}$.
5. Taille d'une personne : $\Omega = \mathbb{R}^+$.
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.

2. Johannes Kepler (1571, Weil der Stadt - 1630, Ratisbonne), mathématicien, astronome et astrologue allemand.
3. Galilée ou Galileo Galilei (1564, Pise - 1642, Arcetri), physicien et astronome italien.
4. 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 ces dernières à la fin de 1654, à la suite d'une expérience mystique, et se consacre à la réflexion philosophique et religieuse.
5. Pierre de Fermat (1601, Beaumont-de-Lomagne - 1665, Castres), juriste et mathématicien français.
6. Christiaan Huygens (1629, La Haye — 1695, La Haye), mathématicien, astronome et physicien néerlandais.
7. Jacques ou Jakob Bernoulli (1654, Bâle - 1705, Bâle), mathématicien et physicien suisse.
8. Abraham de Moivre (1667, Vitry-le-François - 1754, Londres), mathématicien français.
9. Daniel Bernoulli (1700, Groningen - 1782, Bâle), médecin, physicien et mathématicien suisse.
10. 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.
11. Johann Carl Friedrich Gauss (1777, Brunswick - 1855, Göttingen), mathématicien, astronome et physicien allemand.
12. Pierre-Simon Laplace (1749, Beaumont-en-Auge - 1827, Paris), mathématicien, astronome et physicien français.
13. Andreï Nikolaïevich Kolmogorov (1903, Tambov - 1987, Moscou), mathématicien russe.
14. Félix Édouard Justin Émile Borel (1871, Saint-Affrique - 1956, Paris), mathématicien et homme politique français.
15. Henri Léon Lebesgue (1875, Beauvais - 1941, Paris), mathématicien français.

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. Par exemple, dans le troisième exemple, on pourrait tout aussi bien prendre $\Omega = \mathbb{N}^*$, en ne retenant que la durée de la partie ; dans le cinquième, on pourrait limiter, par exemple, Ω à $[0,3]$ (mètres), voire à $\{1,2,\dots,3000\}$ (millimètres), sans perte de généralité.

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 caractérisé par 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 » :

$$\{(m,n) \in \Omega \,:\, 2|(m+n)\}\,.$$

Exemple 0.3. On effectue une suite de lancers de dé :

$$\Omega = \{(a_k)_{k\geq 1} \,:\, a_k \in \{1,\dots,6\}, \forall k \geq 1\}\,.$$

L'événement « le 6 est sorti avant le 1 » correspond à

$$\{(a_k)_{k\geq 1} \in \Omega \,:\, \min\{n \geq 1 \,:\, a_n = 6\} < \min\{n \geq 1 \,:\, a_n = 1\}\}.$$

(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 $\min \varnothing = +\infty$.)

Introduisons un peu de terminologie.

Définition 0.1. *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 de ces derniers une probabilité, qui représente le degré de confiance que l'on a en sa réalisation [16]. 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 0.1. *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, supposez que l'on puisse donner un sens au tirage au hasard d'un nombre réel dans l'intervalle* $[0,1]$*, de façon uniforme (c'est-à-dire sans privilégier aucun de ces nombres). On verra comment le faire dans la seconde partie de ce cours. 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 !*

Il est important de remarquer à ce point que la détermination de la probabilité à associer à un événement donné ne fait pas partie du modèle que nous cherchons à construire (on pourra cependant parfois la déterminer si l'on nous donne la probabilité d'autres événements). Le but de la théorie des probabilités est de définir un cadre mathématique permettant de décrire des phénomènes aléatoires, mais déterminer les paramètres permettant d'optimiser l'adéquation entre le modèle et l'expérience réelle qu'il tente de décrire n'est pas du ressort de la théorie (c'est une tâche dévolue aux statistiques). En particulier, nous ne nous intéresserons pas aux différentes interprétations de la notion de probabilité. Contentons-nous 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 correspondant à des expériences pouvant être reproduites à l'infini, dans les mêmes conditions et de façon indépendante. On identifie alors la probabilité d'un événement avec la fréquence asymptotique de réalisation de cet événement lorsque l'expérience est répétée 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, et ne permet par exemple pas 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 7,5 sur l'échelle de Richter dans les 30 prochaines années ». Dans de telles affirmations, l'événement en question ne correspond pas à une expérience renouvelable, et la notion de probabilité n'a plus d'interprétation en termes de fréquence, mais en termes de quantification de notre degré de certitude subjectif quant à la réalisation de l'événement en question. En résumé, il existe de nombreuses interprétations du concept de probabilité, dont certaines sont beaucoup moins contraignantes que l'interprétation fréquentiste, mais il s'agit d'un problème épistémologique que nous ne discuterons pas ici

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 mentionnée plus haut. Répétons N fois une expérience, dans les mêmes conditions, et notons

16. Comme on le verra dans le seconde partie de ce cours, il n'est pas toujours possible d'associer une probabilité à tous les sous-ensembles de Ω. On devra alors restreindre la notion d'événements à une classe de « bons » sous-ensembles. Ceci n'aura cependant absolument aucune incidence pratique. En effet, aucun des sous-ensembles exclus n'admet de description explicite !

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

Exemple 0.4. On jette deux dés non pipés. Il est alors naturel de prendre $\Omega = \{(n,m) \in \{1,2,3,4,5,6\}^2\}$. Les dés étant supposés bien 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. En fait, comme on le verra dans la seconde partie de ce cours, il sera très utile (et plutôt naturel !) d'imposer une condition plus forte que 3., à savoir

3'. Si $A_1, A_2, \ldots$ sont des événements deux-à-deux disjoints, alors

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

Ceci ne nous concernera pas pour la première partie de ce cours, dans laquelle nous supposons l'univers Ω fini ou dénombrable : dans ce cas, on verra que l'on peut associer à chaque événement élémentaire sa probabilité, et en déduire la probabilité des événements composites. Les propriétés 1, 2 et 3' deviennent alors, dans ce cadre-là, des *conséquences* de cette construction.

Première partie

Espaces de probabilité discrets

Résumé

Dans cette partie du cours, nous nous restreindrons au cas où l'univers associé à l'expérience aléatoire est fini ou dénombrable. On parle alors d'espaces de probabilité discrets. La formulation mathématique de la théorie est beaucoup plus simple dans ce cas, mais permet déjà d'étudier de nombreux problèmes d'intérêt.

Chapitre 1

Probabilité, probabilité conditionnelle et indépendance

1.1 Mesures de probabilité

On considère une expérience aléatoire dont l'univers Ω est fini ou dénombrable. Dans ce cas, on peut associer à chaque résultat possible de l'expérience sa probabilité. Ceci définit une application de Ω dans $[0,1]$, appelée la fonction de masse.

Définition 1.1. *Une* ***fonction de masse*** *sur Ω est une application $f : \Omega \to [0,1]$ telle que*

$$\sum_{\omega \in \Omega} f(\omega) = 1.$$

Remarque 1.1. *Il convient de faire quelques commentaires sur l'écriture utilisée ci-dessus. Une expression du type*

$$\sum_{a \in A} g(a)$$

a un sens évident lorsque A est un ensemble fini et non-vide, puisque dans ce cas on a affaire à une somme finie. Lorsque A est un ensemble infini dénombrable, il convient d'être plus prudent. Dans le cas où la fonction $g : A \to \mathbb{R}$ est positive, comme c'est le cas dans la définition précédente, il n'y a pas de problème : A étant dénombrable, il est possible de numéroter ses éléments, disons $A = \{a_1, a_2, \ldots\}$. On pose alors

$$\sum_{a \in A} g(a) = \sum_{i=1}^{\infty} g(a_i).$$

Il est important d'observer que cette définition ne dépend pas *de l'ordre choisi pour les éléments de A : la série apparaissant dans le membre de droite étant à termes positifs, la somme est inchangée lorsque l'ordre des termes est modifié.*

On utilisera occasionnellement la notation également lorsque $A = \varnothing$. Dans ce cas, la somme est définie comme étant égale à 0.

Exemple 1.1.
- ▷ Pour un dé non pipé, on prend $\Omega = \{1,2,3,4,5,6\}$ et $f(i) = \frac{1}{6}$, $i = 1, \ldots, 6$.
- ▷ Pour un dé pipé, on pourra avoir par exemple $f(1) = \frac{1}{6}$, $f(2) = f(3) = f(4) = f(5) = \frac{1}{8}$ et $f(6) = \frac{1}{3}$.
- ▷ Pour 5 lancers d'une pièces bien équilibrée, on prendra $f(\omega) = 2^{-5}$, pour tout $\omega \in \Omega = \{P,F\}^5$.

▷ On lance un dé jusqu'à la première apparition d'un 6. Si l'on s'intéresse au nombre de lancers nécessaires, il est naturel de prendre $\Omega = \mathbb{N}^\star \cup \{+\infty\}$, la valeur $+\infty$ correspondant à une expérience où le 6 ne sort jamais. Si le dé est équilibré, on verra que $f(n) = \frac{1}{6}(\frac{5}{6})^{n-1}$, pour tout $n \in \mathbb{N}^\star$. En particulier,

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

et la probabilité de ne jamais voir de 6 est nulle.

♢

Une fois en possession d'une fonction de masse, on peut définir la probabilité d'un événement arbitraire $A \subset \Omega$.

Définition 1.2. *Soit Ω un ensemble fini ou dénombrable et f une fonction de masse sur Ω. La probabilité de l'événement $A \subset \Omega$ est définie par*

$$\mathbb{P}(A) = \sum_{\omega \in A} f(\omega).$$

L'application $\mathbb{P} : \mathcal{P}(\Omega) \to [0,1]$ est la **mesure de probabilité** *sur Ω associée à la fonction de masse f. La paire $(\Omega,\mathbb{P})$ définit un* **espace de probabilité discret**.

Évidemment, étant donnée une mesure de probabilité $\mathbb{P}$, on peut immédiatement retrouver la fonction de masse correspondante : $f(\omega) = \mathbb{P}(\{\omega\})$.

Exemple 1.2. On lance un dé équilibré. Soit A l'événement « le résultat est pair ». Alors,

$$\mathbb{P}(A) = \mathbb{P}(\{2,4,6\}) = f(2) + f(4) + f(6) = \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{2}\,.$$

♢

Énonçons à présent quelques propriétés élémentaires, mais extrêmement importantes de telles mesures de probabilité.

Théorème 1.1. *Toute mesure de probabilité $\mathbb{P}$ sur un ensemble Ω fini ou dénombrable possède les propriétés suivantes.*

1. $\mathbb{P}(\Omega) = 1$.
2. *(σ-additivité) Soit $(A_k)_{k\geq 1}$ une collection d'événements 2 à 2 disjoints (c'est-à-dire tels que $A_i \cap A_j = \varnothing$ pour tout $i \neq j$). Alors,*

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

Démonstration. La première affirmation suit immédiatement de la définition. Pour la seconde, il suffit d'observer que

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{\omega \in \bigcup_{i=1}^{\infty} A_i} f(\omega) = \sum_{i=1}^{\infty} \sum_{\omega \in A_i} f(\omega) = \sum_{i=1}^{\infty} \mathbb{P}(A_i),$$

puisque chaque $\omega \in \bigcup_i A_i$ appartient à exactement un des ensemble A_i. □

Corollaire 1.1. *Toute application $\mathbb{P} : \mathscr{P}(\Omega) \to [0,1]$ possédant les propriétés 1. et 2. du théorème possède également les propriétés suivantes.*

1. $\mathbb{P}(\varnothing) = 0$.
2. *Pour tout $A \subset \Omega$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.*
3. *(Monotonicité) Pour tout $A \subset B \subset \Omega$,*
$$\mathbb{P}(A) \leq \mathbb{P}(B).$$
4. *(Additivité finie) Soit $A_1, \ldots, A_n$ une collection finie d'événements 2 à 2 disjoints. Alors,*
$$\mathbb{P}(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$
5. *(Sous-σ-additivité) Soit I un ensemble fini ou dénombrable et $(A_i)_{i\in I}$ une collection d'événements. Alors,*
$$\mathbb{P}(\bigcup_{i\in I} A_i) \leq \sum_{i\in I} \mathbb{P}(A_i).$$
6. *Pour tout $A, B \subset \Omega$,*
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$
7. *Plus généralement, pour tout collection finie $A_1, A_2, \ldots, A_n \subset \Omega$,*
$$\mathbb{P}(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{1\leq i<j\leq n} \mathbb{P}(A_i \cap A_j) + \sum_{1\leq i<j<k\leq n} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1}\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n).$$

   *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**[1]).*

Démonstration. *1.* La collection d'événements $\varnothing, \varnothing, \varnothing, \ldots$ est 2-à-2 disjointe. Il suit donc de la σ-additivité que
$$\mathbb{P}(\varnothing) = \mathbb{P}(\varnothing \cup \varnothing \cup \varnothing \cup \cdots) = \mathbb{P}(\varnothing) + \mathbb{P}(\varnothing) + \mathbb{P}(\varnothing) + \cdots,$$
ce qui n'est possible que si $\mathbb{P}(\varnothing) = 0$.
4. Il suffit d'appliquer la propriété de σ-additivité à la collection $(B_k)_{k\geq 1}$ avec $B_k = A_k$ pour $1 \leq k \leq n$, et $B_k = \varnothing$ pour $k > n$, et de conclure en utilisant le fait que $\mathbb{P}(\varnothing) = 0$.
2. $1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$.
3. $\mathbb{P}(B) = \mathbb{P}(A \cup (B \setminus A)) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$.
5. Il suffit de considérer le cas dénombrable $(A_k)_{k\geq 1}$ (sinon on complète avec une infinité de copies de l'ensemble vide). Introduisons $B_1 = A_1$ et, pour $k \geq 2$, $B_k = A_k \setminus \bigcup_{i=1}^{k-1} B_i$. 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, $\mathbb{P}(\bigcup_{k=1}^{\infty} A_k) = \mathbb{P}(\bigcup_{k=1}^{\infty} B_k) = \sum_{k=1}^{\infty} \mathbb{P}(B_k) \leq \sum_{k=1}^{\infty} \mathbb{P}(A_k)$.
6. Comme $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus (A \cap B))$, l'affirmation suit de $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus (A \cap B))$.
7. Sera fait en exercice. □

1. Carlo Emilio Bonferroni (1892, Bergame – 1960, Florence), mathématicien italien, spécialiste en théorie des probabilités.

Un cas particulièrement important est celui où la même probabilité est associée à chaque résultat possible de l'expérience. Bien entendu, ceci n'est possible que si Ω est fini (pourquoi ?).

Définition 1.3. *On appelle* ***mesure de probabilité uniforme*** *sur un ensemble Ω fini, la mesure de probabilité associée à la fonction de masse $f(\omega) = 1/|\Omega|$, pour tout $\omega \in \Omega$. On dit dans ce cas qu'il y a* ***équiprobabilité****.*

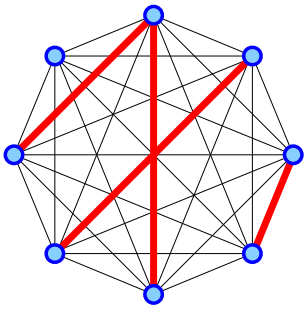
Manifestement, lorsqu'il y a équiprobabilité, la probabilité d'un événement A est simplement donnée par

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

En d'autres termes, la probabilité de A est alors donnée par le quotient du « nombre de cas favorables » par le « nombre de cas total ».

Exemple 1.3. Nous allons à présent introduire un exemple non-trivial d'espace de probabilité fini : le graphe aléatoire d'Erdős[2]–Rényi[3]. Soient $m \geq 0$ et $n \geq 1$ deux entiers. Le **graphe aléatoire** $\mathcal{G}(n,m)$ est l'espace de probabilité sur l'ensemble des graphes $G = (S,A)$ avec ensemble de sommets $S = \{1,\ldots,n\}$ et ensemble d'arêtes $A \subset \{\{i,j\} : 1 \leq i < j \leq n\}$ satisfaisant $|A| = m$. La mesure de probabilité sur cet ensemble est la mesure uniforme.

À titre d'exemple, voici une réalisation du graphe aléatoire $\mathcal{G}(8,4)$ (les arêtes présentes sont indiquées en rouge) :



♢

1.2 Quelques résultats combinatoires

Nous allons à présent rappeler certains résultats élémentaires de combinatoire qui sont régulièrement utilisés par la suite. On utilisera la notation suivante : pour $n \geq r \geq 1$, le **symbole de Pochhammer**[4] $(n)_r$ est défini par

$$(n)_r = n(n-1)(n-2)\cdots(n-r+1).$$

On posera également $(n)_0 = 1$.

1.2.1 Échantillons ordonnés

Considérons un ensemble de n éléments $\{a_1,\ldots,a_n\}$. Un **échantillon ordonné de taille** r est une suite ordonnée de r éléments de l'ensemble. On distingue deux procédures :

2. Pál Erdős (1913, Budapest – 1996, Varsovie), également ortographié Paul Erdős, Paul Erdös ou Paul Erdos, mathématicien hongrois.
3. Alfréd Rényi (1921, Budapest – 1970, Budapest), mathématicien hongrois.
4. Leo August Pochhammer (1841, Stendal – 1920, Kiel), mathématicien prusse.

- ▷ le **tirage avec remise**, durant lequel chaque élément de l'ensemble peut être choisi à plusieurs reprises ;
- ▷ le **tirage sans remise**, durant lequel chaque élément de l'ensemble ne peut être choisi qu'au plus une fois (dans ce cas, on doit évidemment avoir $r \leq n$).

Exemple 1.4. ▷ On lance un dé 10 fois en notant la suite de résultats obtenus. On obtient ainsi un échantillon de taille 10 correspondant à un tirage avec remise à partir de l'ensemble $\{1,\ldots,6\}$.

▷ En Suisse, le résultat d'un tirage au loto correspond à extraire un échantillon de taille 6 par tirage sans remise à partir de l'ensemble $\{1,\ldots,42\}$ [5].

♢

Lemme 1.1. *On considère un ensemble A à $n \geq 1$ éléments, et $r \in \mathbb{N}$.*

1. *Le nombre d'échantillons de taille r distincts que l'on peut obtenir par tirage avec remise d'éléments de A est égal à n^r.*
2. *Pour $r \leq n$, le nombre d'échantillons de taille r distincts que l'on peut obtenir par tirage sans remise d'éléments de A est égal à $(n)_r$.*
3. *Le nombre de façons d'ordonner l'ensemble est égal à $n!$.*

Démonstration. *1.* Dans le cas du tirage avec remise, chacun des r éléments peut être choisi de n façons différentes. Par conséquent, le nombre total d'échantillons possibles est égal à n^r.

2. Dans le cas sans remise, le premier élément est choisi parmi n, le second parmi $n-1$ (celui choisi lors du premier tirage ne pouvant pas être choisi à nouveau), le troisième parmi $n-2$, etc. On a donc un nombre total d'échantillons possibles égal à $(n)_r$.

3. Suit de *2.* puisque cela revient à faire n tirages sans remise et que $(n)_n = n!$. □

Jusqu'à présent, il n'a pas été fait mention de probabilité. Lorsque nous parlerons d'**échantillon aléatoire** de taille r, l'adjectif « aléatoire » signifiera que l'on a muni l'ensemble de tous les échantillons possibles d'une mesure de probabilité. *Sauf mention explicite du contraire, on considérera la mesure uniforme.*

Considérons à présent un échantillon aléatoire avec remise de taille r. On s'intéresse à l'événement « aucun élément n'a été choisi plus d'une fois ». Le Lemme 1.1 montre que, parmi les n^r échantillons possibles, $(n)_r$ satisfont cette contrainte. Par conséquent, la probabilité que notre échantillon ne contienne pas de répétition est donnée par $(n)_r/n^r$. Ce résultat a des conséquences qui peuvent sembler surprenantes.

Exemple 1.5. Supposons que, dans une certaine ville, il y ait 7 accidents par semaine. Alors, durant la quasi-totalité des semaines, certains jours verront plusieurs accidents. En posant $n = r = 7$, on voit en effet que la probabilité d'avoir exactement un accident chaque jour de la semaine est seulement de $0{,}00612\ldots$; cela signifie qu'un tel événement n'aura lieu en moyenne qu'environ une fois tous les trois ans ! ♢

Exemple 1.6. Supposons que 23 personnes se trouvent dans la même salle. Quelle est la probabilité qu'au moins deux d'entre elles aient leur anniversaire le même jour ? On peut modéliser cette situation, en première approximation, par un tirage aléatoire avec remise à partir de l'ensemble $\{1,\ldots,365\}$, avec la mesure uniforme ; un modèle plus réaliste devrait prendre en compte les années bissextiles, ainsi que les variations saisonnières du taux de natalité (sous nos latitudes, le nombre de naissances est plus élevé en été qu'en hiver [6], par exemple), etc. Pour le modèle

5. Notons toutefois que l'ordre ne joue par contre aucun rôle pour déterminer si une grille est gagnante

6. Ceci dit, considérer une répartition inhomogène des naissances ne peut qu'augmenter la probabilité d'avoir plusieurs personnes avec la même date d'anniversaire...

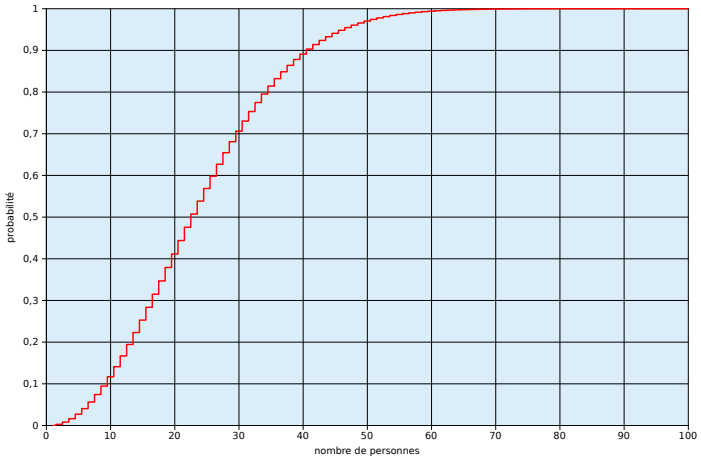


FIGURE 1.1: La probabilité qu'au moins deux personnes aient leur anniversaire à la même date, en fonction de la taille du groupe.

précédent, il suit de la discussion ci-dessus que la probabilité qu'au moins deux des 23 personnes aient leur anniversaire le même jour est donnée par $1 - (365)_{23}/365^{23} = 0{,}507\ldots$: il y a plus d'une chance sur deux que ça ait lieu !

Cette probabilité est de 97% s'il y a 50 personnes, et de 99,99996% pour 100 personnes ; voir la figure 1.1. ♢

1.2.2 Échantillons non ordonnés

Considérons à présent le problème d'extraire un échantillon de taille r d'une population de taille n sans tenir compte de l'ordre. En d'autres termes, étant donnée une population de taille n, nous cherchons à déterminer le nombre de sous-populations de taille r.

Lemme 1.2. *Une population de taille n possède $\binom{n}{r}$ différentes sous-populations de taille $r \leq n$.*

Démonstration. Chaque sous-population de taille r peut être ordonnée de $r!$ façons différentes. Puisque le nombre total d'échantillons ordonnés de taille r obtenus sans remise est égal à $(n)_r$, on en déduit que le nombre d'échantillons non-ordonnés de taille r doit être égal à $(n)_r/r! = \binom{n}{r}$. □

Exemple 1.7. Au poker, chaque joueur reçoit 5 cartes parmi 52. Le nombre de mains possibles est donc de $\binom{52}{5} = 2\,598\,960$. Calculons alors la probabilité d'avoir 5 cartes de valeurs différentes. On peut choisir ces valeurs de $\binom{13}{5}$ façons différentes. Il faut ensuite associer à chacune une couleur, ce qui donne un facteur additionnel 4^5. Par conséquent, la probabilité en question est donnée par $4^5 \cdot \binom{13}{5}/\binom{52}{5} = 0{,}5071\ldots$. ♢

Exemple 1.8. Considérons la distribution aléatoire de r boules dans n urnes. Quelle est la probabilité qu'une urne donnée contienne exactement k boules ? On peut choisir les k boules de $\binom{r}{k}$ façons. Les autres $r-k$ boules doivent être réparties parmi les $n-1$ urnes restantes, ce qui peut se faire de $(n-1)^{r-k}$ façons. Il s'ensuit que la probabilité en question est donnée par

$$\frac{1}{n^r} \cdot \binom{r}{k} \cdot (n-1)^{r-k} = \binom{r}{k} \cdot \frac{1}{n^k} \cdot \left(1 - \frac{1}{n}\right)^{r-k}.$$

Il s'agit d'un cas particulier de la **distribution binomiale**, que nous reverrons plus tard. ♢

Exemple 1.9. Retournons au graphe aléatoire de l'Exemple 1.3. On a clairement

$$\left|\{\{i,j\} : 1 \leq i < j \leq n\}\right| = \binom{n}{2} \equiv N.$$

Par conséquent, le nombre total de graphes dans $\mathcal{G}(n,m)$ est donné par $\binom{N}{m}$. La fonction de masse associée à ce modèle est donc donnée par

$$f(G) = \binom{N}{m}^{-1}, \qquad \forall G \in \mathcal{G}(n,m).$$

(On fait ici un léger abus de notation en utilisant la même écriture pour l'espace de probabilité et pour l'univers.) ♢

Exemple 1.10. On offre à 100 condamnés à mort la possibilité d'être graciés s'ils parviennent à gagner à un « jeu ». On les conduit donc tous dans une salle A et on leur décrit la procédure à laquelle ils vont être soumis :

- ▷ Chaque prisonnier, à tour de rôle, sera conduit dans une salle B.
- ▷ La salle B contient 100 coffres, fermés, numérotés de 1 jusqu'à 100. Chacun des coffres contient le nom d'un unique prisonnier, et le nom de chacun des prisonniers est contenu dans un des coffres.
- ▷ Le prisonnier pourra alors ouvrir au plus 50 de ces 100 coffres, choisis comme il le désire.
- ▷ Si son nom se trouve dans un des coffres ouverts, on le conduira dans une salle C. Les coffres seront alors refermés et on passera au prisonnier suivant.
- ▷ Si au moins un des prisonniers ne trouve son nom dans aucun des coffres qu'il aura ouverts, tous les prisonniers seront exécutés. S'ils trouvent tous leur nom, ils seront libérés.

Les prisonniers peuvent se mettre d'accord sur une stratégie commune afin de maximiser leur chance de survie.

La « stratégie » naïve consistant, pour chacun des prisonniers, à ouvrir 50 des coffres au hasard n'est guère prometteuse : chaque prisonnier a une chance sur deux de trouver son nom dans les coffres qu'il ouvre (pourquoi ?), et les prisonniers seront donc exécutés avec probabilité $1 - 2^{-100}$...

Nous allons montrer qu'il existe une bien meilleure stratégie leur donnant plus de 30% de chance de survie !

Cette stratégie (dont on peut montrer qu'elle est optimale) consiste pour les prisonniers à procéder de la façon suivante :

- ▷ Les prisonniers commencent par se numéroter de 1 à 100 au hasard (uniformément).
- ▷ Lorsque le prisonnier auquel a été associé le numéro i est conduit dans la salle B, il ouvre le coffre portant le numéro i et lit le nom qui y est contenu. S'il s'agit de son nom, il s'interrompt et est conduit à la salle C. S'il ne s'agit pas de son nom, alors il s'agit du nom d'un autre prisonnier dont le numéro est j. Il ouvre alors le coffre j, lit le nom inscrit, et continue de la même façon jusqu'à ce qu'il ait soit ouvert 50 coffres, soit trouvé son nom dans un des coffres.

Quelle est la probabilité pour que chaque prisonnier trouve son nom dans un des 50 coffres qu'il ouvre ?

L'observation cruciale est qu'une fois les prisonniers numérotés, les noms contenus dans les coffres définissent une permutation de l'ensemble $\{1,\ldots,100\}$: la permutation associe au numéro inscrit sur le coffre le numéro correspondant au prisonnier dont le nom est contenu dans le coffre. Cette permutation est aléatoire, puisqu'elle dépend de l'ordre dans lequel sont numérotés les prisonniers. De plus, les permutations sont équiprobables, puisqu'il en est de même de l'ordre de numérotation des prisonniers.

Ainsi, le prisonnier numéro i trouvera son nom dans l'un des 50 coffres qu'il ouvre si le cycle de la permutation contenant l'élément i est de longueur au plus 50. En effet, dans ce cas, il va nécessairement trouver un coffre contenant le nom associé au numéro i (ce qui ferme le cycle). Or, c'est son nom qui est associé au numéro i.

On voit donc que pour que tous les prisonniers survivent, il est nécessaire et suffisant que *tous les cycles de la permutation soient de longueur au plus* 50. On est donc conduit au problème combinatoire suivant.

Soit S_{2n} l'ensemble des permutations de $\{1,\dots,2n\}$. On munit cet ensemble de la mesure de probabilité uniforme, c'est-à-dire $f(\pi) = 1/(2n)!$, pour toute permutation $\pi \in S_{2n}$.

On veut déterminer la probabilité de l'événement

$$A = \{\pi \in S_{2n} \,:\, \text{tous les cycles de } \pi \text{ sont de longueur au plus } n\}\,.$$

Il nous faut donc déterminer la cardinalité de A. Il est en fait plus simple de déterminer celle de l'événement complémentaire « il existe exactement un cycle de longueur strictement supérieure à n » (pourquoi peut-on écrire « exactement » ?). Dans ce cas, on peut commencer par fixer la longueur $\ell > n$ du plus grand cycle. Le nombre de façons de choisir les ℓ éléments composant ce dernier est $\binom{2n}{\ell}$. Il convient ensuite de les ordonner afin de former un cycle : ceci peut se faire de $(\ell-1)!$ façons différentes (observez que les ordres 1,5,3,7 et 3,7,1,5, par exemple, décrivent le même cycle !). Finalement, il reste à considérer toutes les permutations possibles des $2n-\ell$ éléments n'appartenant pas au plus grand cycle, ce qui contribue un facteur $(2n-\ell)!$. On obtient donc finalement que le nombre de permutations appartenant à A est égale à

$$(2n)! - \sum_{\ell=n+1}^{2n} \binom{2n}{\ell} (\ell-1)!(2n-\ell)! = (2n)! - (2n)! \sum_{\ell=n+1}^{2n} \frac{1}{\ell},$$

et la probabilité recherchée est donc

$$\mathbb{P}(A) = 1 - \sum_{\ell=n+1}^{2n} \frac{1}{\ell}.$$

En particulier, pour tout n,

$$\mathbb{P}(A) \geq 1 - \int_n^{2n} \frac{1}{x}\,\mathrm{d}x = 1 - \log 2 > 30\%.$$

♢

1.2.3 Partitionnement

Finalement, considérons le nombre de façons de partitionner une population en k sous-populations de tailles données.

Lemme 1.3. *Soit $r_1,\dots,r_k$ des entiers positifs (éventuellement nuls) tels que $r_1+\cdots+r_k = n$. Le nombre de façons de répartir n objets dans k familles, de sorte à ce que la $i^{ème}$ famille contienne r_i éléments, est égal à*

$$\frac{n!}{r_1!r_2!\cdots r_k!}.$$

Démonstration. Pour remplir la première famille, il faut choisir r_1 objets parmi n, ce qui peut se faire de $\binom{n}{r_1}$ façons. Pour remplir la seconde famille, il faut choisir r_2 objets parmi les $n-r_1$

objets restants, soit $\binom{n-r_1}{r_2}$ possibilités. En continuant ainsi, on obtient que le nombre de telles répartitions est de

$$\binom{n}{r_1}\binom{n-r_1}{r_2}\binom{n-r_1-r_2}{r_3}\cdots\binom{n-r_1-\cdots-r_{k-1}}{r_k} = \frac{n!}{r_1!r_2!\cdots r_k!}.$$

□

Exemple 1.11. À une table de bridge, les 52 cartes sont distribuées à 4 joueurs. Quelle est la probabilité que chacun reçoive un as ? Le nombre total de différentes répartitions est de $52!/(13!)^4$. Les 4 as peuvent être ordonnés de 4! façons différentes, et chaque ordre correspond à une façon de les répartir parmi les 4 joueurs. Les 48 cartes restantes peuvent ensuite être réparties de $48!/(12!)^4$ façons. Par conséquent, la probabilité en question est de

$$4!\,\frac{48!}{(12!)^4}\Big/\frac{52!}{(13!)^4} = 0{,}105\ldots$$

♢

1.2.4 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 [7] (pourquoi retrouve-t-on bien la formule usuelle lorsque $\alpha \in \mathbb{N}$?).

Lemme 1.4. *Soient $x,y,\alpha \in \mathbb{R}$. 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* $\alpha \geq 0$;
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. □

1.2.5 Formule de Stirling

Il se révèle souvent très 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 [8].

Lemme 1.5. *Pour tout $n \geq 1$, on a*

$$e^{1/(12n+1)}\,n^n e^{-n}\sqrt{2\pi n} \leq n! \leq e^{1/(12n)}\,n^n e^{-n}\sqrt{2\pi n}.$$

Démonstration. Une version de ce résultat sera démontrée en exercice. □

7. Sir Isaac Newton (1643, Woolsthorpe-by-Colsterworth – 1727, Londres), philosophe, mathématicien, physicien, alchimiste, astronome et théologien anglais.

8. James Stirling (1692, Garden – 1770, Leadhills), mathématicien britannique.

1.3 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 le nombre N_B de tentatives lors desquelles B est réalisé, et le nombre $N_{A\cap B}$ 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 1.4. *Soit $B \subset \Omega$ un événement tel que $\mathbb{P}(B) > 0$. Pour tout $A \subset \Omega$, la* ***probabilité conditionnelle de A sachant B*** *est la quantité*

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

Lemme 1.6. *Soit $B \subset \Omega$ un événement tel que $\mathbb{P}(B) > 0$. Alors l'application $\mathbb{P}(\cdot \mid B) : \mathcal{P}(\Omega) \to \mathbb{R}$ est une mesure de probabilité sur Ω et sur B.*

Démonstration. On vérifie aisément que la fonction

$$f_{|B}(\omega) = \begin{cases} \mathbb{P}(\{\omega\})/\mathbb{P}(B) & \text{si } \omega \in B, \\ 0 & \text{sinon,} \end{cases}$$

est une fonction de masse sur Ω et sur B, et que $\mathbb{P}(A \mid B) = \sum_{\omega\in A} f_{|B}(\omega)$, pour tout $A \subset \Omega$. □

Exemple 1.12. On jette deux dés non pipés. Sachant que le premier jet nous donne 3, quelle est la probabilité que la somme soit supérieure à 6 ? Ici, $B = \{(3,k) : k = 1,\dots,6\}$, $A = \{(a,b) \in \{1,\dots,6\}^2 : a + b > 6\}$, 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 1.13. Considérons les deux problèmes suivants :

- ▹ Vous êtes invité chez une personne dont vous savez qu'elle a exactement deux enfants. Lorsque vous sonnez à sa porte, un garçon vient vous ouvrir. Quelle est la probabilité que l'autre enfant soit également un garçon ?
- ▹ Vous êtes invité chez une personne dont vous savez qu'elle a exactement deux enfants. Lorsque vous sonnez à sa porte, un garçon vient vous ouvrir. Vous entendez un bébé pleurer dans la maison. Quelle est la probabilité que l'autre enfant soit également un garçon ?

Nous allons voir que les réponses à ces deux problèmes ne sont pas les mêmes : dans le premier cas, la probabilité est de $1/3$, alors que dans le second elle est de $1/2$. Afin de vérifier cela, formalisons plus précisément ces deux situations.

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 $1/4$. On désire déterminer la probabilité que les deux enfants soient des garçons (conditionnellement aux informations disponibles dans chacune des deux situations décrites), ce qui correspond à l'événement $A = \{(\mathrm{G},\mathrm{G})\}$.

Considérons à présent la première situation. L'information que vous obtenez lorsqu'un garçon ouvre la porte est qu'au moins un des deux enfants est un garçon, ce qui correspond à l'événement $B = \{(\mathrm{G},\mathrm{G}), (\mathrm{F},\mathrm{G}), (\mathrm{G},\mathrm{F})\}$. On obtient 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 à la seconde situation décrite. L'information disponible 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 $C = \{(\mathrm{G},\mathrm{G}), (\mathrm{G},\mathrm{F})\}$. On a donc

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

♢

Définition 1.5. *Une famille $(B_i)_{i\in I}$, 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 1.2. *Soit $(B_i)_{i\in I}$ une partition de Ω telle que $\mathbb{P}(B_i) > 0$, pour tout $i \in I$, et soit $A \subset \Omega$.*

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) Si $\mathbb{P}(A) > 0$,*

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

Démonstration. Par σ-additivité,

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

La seconde relation suit de l'observation que

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

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

Remarque 1.2. *Dans la terminologie statistique, on appelle* $\mathbb{P}(B_i)$ *la probabilité* ***à priori*** *de* B_i *et* $\mathbb{P}(B_i \mid A)$ *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 1.14. On se donne deux urnes. La première contient deux balles rouges et trois balles bleues ; la seconde trois rouges et quatre bleues. Une balle est tirée au hasard de la première urne et placée dans la seconde. On tire ensuite au hasard une balle de la seconde urne : quelle est la probabilité qu'elle soit bleue ?

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

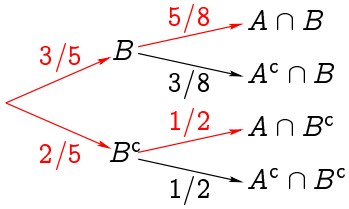
$$\mathbb{P}(A) = \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c).$$

À présent,

$$\mathbb{P}(A \mid B) = \mathbb{P}(A \mid \text{la } 2^{\text{ème}} \text{ urne contient trois balles rouges et cinq bleues}) = \tfrac{5}{8} \ ;$$
$$\mathbb{P}(A \mid B^c) = \mathbb{P}(A \mid \text{la } 2^{\text{ème}} \text{ urne contient quatre balles rouges et quatre bleues}) = \tfrac{1}{2}.$$

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

On représente souvent des situations très simples de ce type de la façon suivante :



♢

Exemple 1.15. Le test de dépistage d'un certain virus n'est pas infaillible :

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

Il est donc important de répondre aux questions suivantes :

1. Étant donné que son test est positif, quelle est la probabilité qu'un individu ne soit pas porteur du virus ?
2. 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 une information supplémentaire : dans la population totale, la fraction de porteurs est approximativement de $1/1000$.

Formalisons tout cela. On introduit les événements suivants :

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

On a donc les informations suivantes :

$$\mathbb{P}(T \mid V^c) = \tfrac{1}{100}, \qquad \mathbb{P}(T^c \mid V) = \tfrac{2}{100}, \qquad \mathbb{P}(V) = \tfrac{1}{1000},$$

et on veut calculer

$$1.\ \mathbb{P}(V^c \mid T), \qquad 2.\ \mathbb{P}(V \mid T^c).$$

La formule de Bayes nous dit que

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

Nous connaissons toutes les valeurs correspondant aux quantités du membre de droite (observez que $\mathbb{P}(T \mid V) = 1 - \mathbb{P}(T^c \mid V) = 98/100$). On obtient donc

$$\mathbb{P}(V^c \mid T) = \frac{\frac{1}{100} \cdot \frac{999}{1000}}{\frac{1}{100} \cdot \frac{999}{1000} + \frac{98}{100} \cdot \frac{1}{1000}} = 0{,}91\ldots$$

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

$$\mathbb{P}(V \mid T^c) = 0{,}00002...$$

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 risques, la probabilité à priori d'être porteur, $\mathbb{P}(V)$, peut devenir proche de 1 et non pas 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 minuscule, tandis que la probabilité d'être porteur alors que le test est négatif est très importante. ♢

L'usage des probabilités conditionnelles peut se révéler très délicat, et l'intuition peut parfois jouer des tours, comme le montrent les exemples suivants.

Exemple 1.16. Un bienfaiteur vous propose le jeu suivant. Il va vous présenter 3 enveloppes fermées ; 2 d'entre elles contiennent du papier journal, la dernière un chèque de 1 000 000 CHF. Vous devrez choisir une enveloppe, sans l'ouvrir. Il ouvrira ensuite une des deux enveloppes restantes et vous montrera qu'elle contient du papier journal. Vous aurez alors le choix entre conserver l'enveloppe choisie initialement, ou bien changer pour celle qui reste. Quelle est la meilleure stratégie ? (Réponse : vous avez deux fois plus de chances de gagner si vous changez ; pourquoi ?) ♢

Exemple 1.17. (Paradoxe du prisonnier) Trois hommes se sont faits 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 2 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. 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. Quelle est la probabilité que A soit exécuté ? ♢

Remarque 1.3. *Dans les 2 exemples précédents, le problème est partiellement mal posé, car la stratégie employée par votre bienfaiteur, ou par le garde, lorsqu'ils ont à prendre une décision n'est pas indiquée. Dans une telle situation, supposez qu'il prend sa décision de façon uniforme (après tout, vous n'avez aucune information sur le sujet, et tout autre choix serait difficile à justifier).*

Si les exemples précédents sont très artificiels et se règlent 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.

Exemple 1.18. (Paradoxe de Simpson[9]) 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	1010
Échec	1801	1190

Le traitement A ayant été administré à 2020 personnes, et 219 d'entre elles ayant été guéries, son taux de succès est donc de $219/2020$, ce qui est très inférieur au taux correspondant pour le traitement B qui est de $1010/2200$. Par conséquent, le traitement B est plus efficace que le traitement A.

Après qu'il ait 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 :

	Femmes		Hommes	
	Traitement A	Traitement B	Traitement A	Traitement B
Succès	200	10	19	1000
Échec	1800	190	1	1000

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 dans les 2 cas. Par conséquent, le traitement A est plus efficace que le traitement B.

Bien entendu, c'est l'assistant qui a raison : quel que soit le sexe du patient, ses chances de guérir sont supérieures avec le traitement A.

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 paramètres que le sexe (l'âge, le poids, ...) dont on n'aurait pas tenu compte et qui modifierait une fois de plus la conclusion ?

Un cas réel célèbre s'est produit lorsque l'université de Berkeley a été poursuivie pour discrimination sexuelle en 1973 : les chiffres des admissions montraient que les hommes ayant posé 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%). Cependant, après avoir analysé séparément les différents départements, on a découvert qu'aucun département n'était significativement biaisé en faveur des hommes ; en fait, la plupart des départements avaient 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. ♢

1.4 Indépendance

En général, l'information 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 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

9. Edward Hugh Simpson. 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.

d'autre part, les deux événements ne jouent pas un rôle symétrique. La notion de probabilité conditionnelle conduit donc à la définition suivante.

Définition 1.6. *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 $(A_i)_{i\in I}$ est ***indépendante*** *sous $\mathbb{P}$ si*

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

pour tous les sous-ensembles finis J de I.

Exemple 1.19. 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 collection finie d'événements. Par exemple, si l'on jette 2 dés équilibrés de couleur différentes, $\Omega = \{(i,j) : 1 \leq i,j \leq 6\}$, et que l'on considère les événements

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

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

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

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

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

♢

Proposition 1.1. *Soient A,B deux événements indépendants. Alors A et B^c sont indépendants, et A^c et B^c sont indépendants.*

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

$$B_1, \ldots, B_n,$$

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

Démonstration. Laissée en exercice. □

Remarque 1.4. *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 2 à 2* ***indépendante****. L'indépendance 2 à 2* n'implique pas l'indépendance*, comme le montre l'exemple suivant.*

Exemple 1.20. 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

$$\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) = \tfrac{1}{2},$$
$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_2 \cap A_3) = \tfrac{1}{4},$$

les événements A_1, A_2 et A_3 sont 2 à 2 indépendants. D'un autre côté,

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = 0 \quad \text{et} \quad \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3) = \tfrac{1}{8},$$

ce qui montre que ces trois événements ne sont pas indépendants. ♢

Exemple 1.21. Retournons au graphe aléatoire $\mathcal{G}(n,m)$; on suppose $n \geq 3$ et $m \geq 2$. 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 encore disponibles.

D'autre part, soient 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. ♢

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 1.22. 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 1.23. 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 2 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 1.7. *Soit C un événement avec $\mathbb{P}(C) > 0$. Deux événements A et B sont* ***indépendants conditionnellement à C*** *sous $\mathbb{P}$ si*

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

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

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

pour tous les sous-ensembles finis J de I.

1.5 Expériences répétées

Dans cette section, nous allons nous intéresser à la description mathématique d'une expérience aléatoire répétée dans les mêmes conditions, de façon indépendante.

Afin de rester concret, nous illustrerons la construction avec le cas particulier du lancer répété d'une pièce de monnaie, un exemple déjà discuté à plusieurs reprises précédemment.

Notons Ω_1 l'univers correspondant à une expérience, et f_1 et $\mathbb{P}_1$ la fonction de masse et la mesure de probabilité associées.

Exemple 1.24. Dans le cas d'un jet d'une pièce de monnaie, $\Omega_1 = \{\mathsf{P},\mathsf{F}\}$, et $f_1(\mathsf{P}) = p$, $f_1(\mathsf{F}) = 1-p \equiv q$, où $p \in [0,1]$ est la probabilité que la pièce tombe sur pile. En particulier, $p = \frac{1}{2}$ dans le cas d'une pièce équilibrée. ♢

Nous allons à présent construire l'espace de probabilité correspondant à 2 répétitions de l'expérience. Évidemment, l'univers associé est donné par le produit cartésien de 2 copies de Ω_1 : $\Omega_2 = \Omega_1 \times \Omega_1 = \{(\omega_1,\omega_2) : \omega_i \in \Omega_1\}$. Nous devons à présent définir la mesure de probabilité $\mathbb{P}_2$ sur Ω_2. L'indépendance des expériences successives implique que les deux événements « le résultat de la première expérience est ω_1 » et « le résultat de la deuxième expérience est ω_2 » doivent être indépendants. De plus, la probabilité d'observer ω_1 lors de la première expérience est donnée par $f_1(\omega_1)$, et similairement pour la deuxième. Ceci implique que

$$f_2(\omega_1,\omega_2) = f_1(\omega_1)f_1(\omega_2), \qquad \forall(\omega_1,\omega_2) \in \Omega_2.$$

Soient $A,B \subset \Omega_1$. L'événement « A a lieu lors de la première expérience et B a lieu lors de la seconde » correspond à $A \times B$. On a alors, comme on le souhaitait,

$$\mathbb{P}_2(A \times B) = \sum_{\substack{\omega_1 \in A \\ \omega_2 \in B}} f_2(\omega_1,\omega_2) = \sum_{\substack{\omega_1 \in A \\ \omega_2 \in B}} f_1(\omega_1)f_1(\omega_2) = \mathbb{P}_1(A)\,\mathbb{P}_1(B).$$

Exemple 1.25. Pour deux jets d'une pièce de monnaie, on obtient

$$\Omega_2 = \{\mathsf{PP},\mathsf{PF},\mathsf{FP},\mathsf{FF}\},$$

et f_2 est déterminée par $f_2(\mathsf{PP}) = p^2$, $f_2(\mathsf{PF}) = f_2(\mathsf{FP}) = pq$ et $f_2(\mathsf{FF}) = q^2$. ♢

On peut aisément itérer la construction ci-dessus de façon à décrire la répétition d'un nombre fini quelconque N d'expériences identiques et indépendantes. On obtient alors l'univers $\Omega_N = \Omega_1 \times \cdots \times \Omega_1$ (n fois), et la fonction de masse $f_n(\omega_1,\ldots,\omega_n) = f_1(\omega_1)\cdots f_1(\omega_n)$.

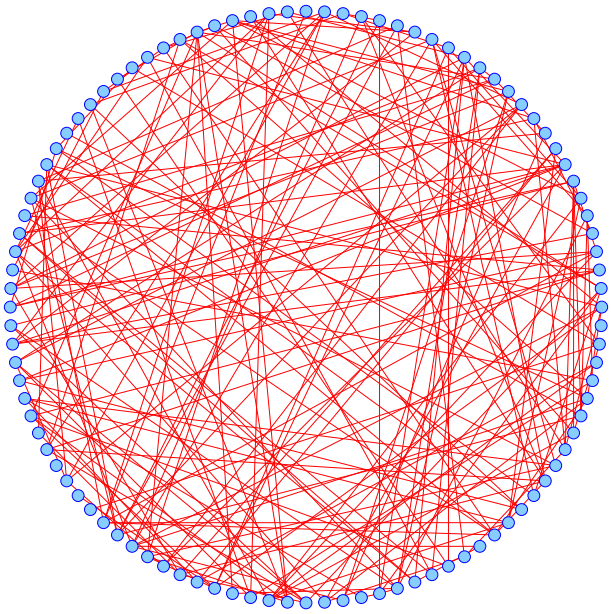
Comme on le verra, il sera souvent pratique de considérer la répétition d'un nombre *infini* d'expériences identiques et indépendantes. L'univers correspondant n'est alors plus dénombrable et une construction plus sophistiquée est nécessaire. Nous y reviendrons plus tard.

Chapitre 2

Variables aléatoires discrètes

2.1 Variables aléatoires discrètes et leurs lois

Il est souvent plus pratique d'associer une 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 1.3 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, de telles valeurs numériques sont des fonctions $X : \Omega \to \mathbb{R}$ associant à un résultat de l'expérience une valeur dans $\mathbb{R}$. Une telle fonction est appelée variable aléatoire.

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

Remarque 2.1. *Il est parfois naturel d'autoriser des variables aléatoires à prendre des valeurs infinies. Bien sûr, ceci n'a d'influence que si la probabilité d'obtenir une valeur infinie est strictement positive. Une variable aléatoire* X *telle que* $\mathbb{P}(X = \infty) > 0$ *est*

dite ***défective****. Dans la suite, les variables aléatoires seront supposées non-défectives, sauf mention explicite du contraire.*

Exemple 2.1. 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. Dans la réalisation de $\mathcal{G}(8,4)$ représentée dans l'Exemple 1.3, on a $N_0 = 1$, $N_1 = 6$, $N_2 = 1$, et $N_k = 0$ pour les autres valeurs de k. ♢

Soit $(\Omega,\mathbb{P})$ un espace de probabilité discret et $X : \Omega \to \mathbb{R}$ une variable aléatoire. Les probabilités qui vont nous intéresser prennent la forme

$$\mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}) = \mathbb{P}(X^{-1}(A)) \equiv \mathbb{P}(X \in A),$$

pour certains sous-ensembles $A \subseteq \mathbb{R}$. La mesure de probabilité $\mathbb{P}$ sur Ω et la variable aléatoire X induisent une mesure de probabilité $\mathbb{P}_X$ sur $\mathbb{R}$ en posant, pour $A \subseteq \mathbb{R}$,

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A).$$

Évidemment, $\mathbb{R}$ n'est pas un ensemble dénombrable. Toutefois, la mesure de probabilité $\mathbb{P}_X$ n'associe une probabilité non-nulle qu'aux éléments du sous-ensemble dénombrable ou fini $X(\Omega)$. On peut donc, en faisant un léger abus de langage, considérer le couple $(\mathbb{R},\mathbb{P}_X)$ comme un espace de probabilité discret.

Définition 2.2. *La mesure de probabilité $\mathbb{P}_X$ sur $\mathbb{R}$ définie par*

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A), \quad \forall A \subset \mathbb{R}$$

est appelée la ***loi*** *de X. La* ***fonction de masse*** *de X est la fonction $f_X : \mathbb{R} \to [0,1]$ donnée par $f_X(x) = \mathbb{P}(X = x)$.*

La fonction de masse satisfait donc $f_X(x) = 0$ pour tout $x \notin X(\Omega)$, et on a, pour tout $A \subset \mathbb{R}$,

$$\mathbb{P}_X(A) = \sum_{x \in A \cap X(\Omega)} f_X(x).$$

Exemple 2.2. 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

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

♢

La mesure de probabilité $\mathbb{P}_X$ contient toute l'information nécessaire pour étudier les propriétés statistiques de la variable aléatoire X ; en particulier, si l'on n'est intéressé que par cette variable aléatoire, l'espace de probabilité de départ $(\Omega,\mathbb{P})$ peut être complétement ignoré, et souvent n'est même pas spécifié, l'espace de probabilité pertinent étant $(\mathbb{R},\mathbb{P}_X)$ (ou, de façon équivalente, $(X(\Omega),\mathbb{P}_X)$). Bien entendu, lorsque Ω n'est plus explicitement mentionné, la variable aléatoire est dite discrète s'il existe un sous-ensemble $K \subset \mathbb{R}$, au plus dénombrable, tel que $\mathbb{P}_X(K) = 1$.

2.1.1 Exemples importants de variables aléatoires discrètes

On présente ici quelques-unes des lois discrètes les plus importantes. Elles sont introduites à partir de leur fonction de masse, et on laisse comme exercice la vérification que celles-ci sont proprement normalisées (c'est-à-dire de somme 1).

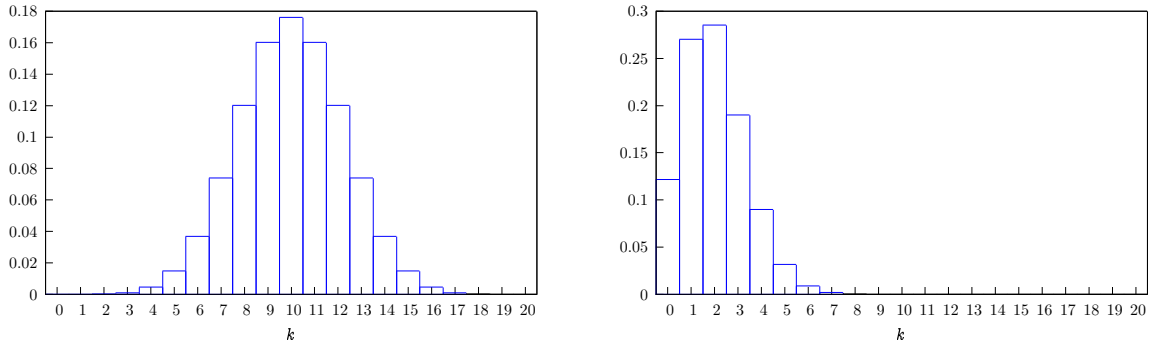


FIGURE 2.1: Loi binomiale pour $n = 20, p = 0{,}5$ (gauche) et $n = 20, p = 0{,}1$ (droite).

Variable aléatoire constante

Une variable aléatoire X est **constante** s'il existe $c \in \mathbb{R}$ tel que $\mathbb{P}(X = c) = 1$.

Loi de Bernoulli

La loi d'une variable aléatoire $X : \Omega \to \{0,1\}$, avec $f_X(1) = p$, $f_X(0) = 1 - p$, $p \in [0,1]$, est appelée **loi de Bernoulli** de paramètre p. On écrit $X \sim \text{bernoulli}(p)$.

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

Exemple 2.3.
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 $A \subset \Omega$, 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

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. La loi de X est appelée **loi binomiale** de paramètres n et p ; $X \sim \text{binom}(n,p)$. Puisqu'il y a $\binom{n}{k}$ façons d'obtenir k succès sur n épreuves, on voit que la fonction de masse associée à cette loi est donnée par

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

Loi de Poisson

Une variable aléatoire X suit une **loi de Poisson**[1] de paramètre $\lambda > 0$, $X \sim \text{poisson}(\lambda)$, si elle prend ses valeurs dans $\mathbb{N}$ et possède la fonction de masse

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

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

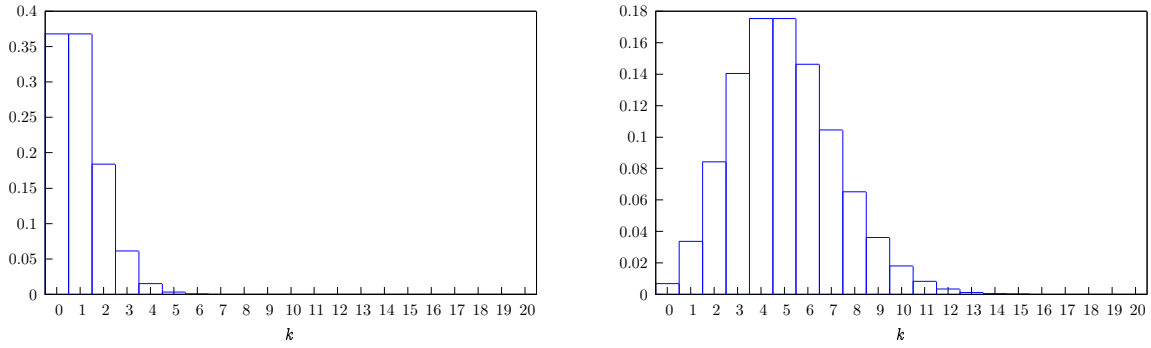


FIGURE 2.2: Loi de Poisson pour $\lambda = 1$ (gauche) et $\lambda = 5$ (droite).

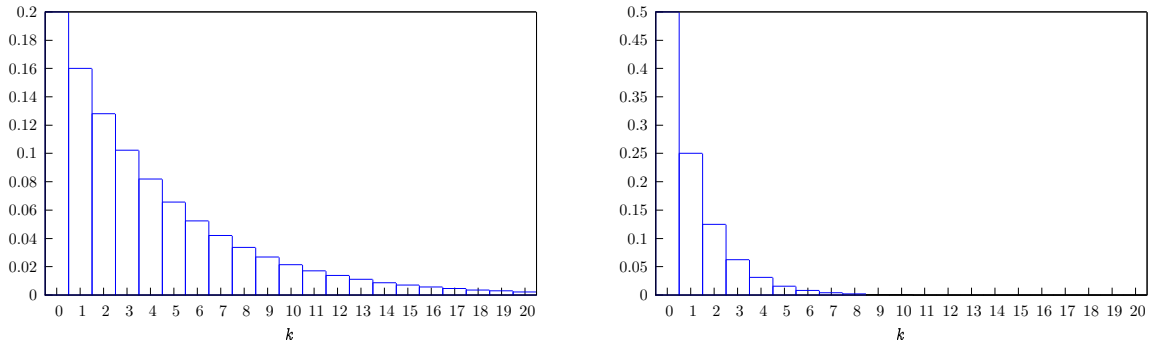


FIGURE 2.3: Loi géométrique pour $p = 0{,}2$ (gauche) et $p = 0{,}5$ (droite).

Considérons une variable aléatoire X suivant une loi binomiale de paramètres n et 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 pendant la transmission). Alors X suit approximativement une loi de Poisson de paramètre $\lambda = np$ (c'est ce qu'on appelle parfois la **loi des petits nombres**). Plus précisément,

$$\begin{aligned} f_X(k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{1}{k!} \frac{n}{n} \frac{n-1}{n} \frac{n-2}{n} \cdots \frac{n-k+1}{n} (np)^k (1-p)^{n-k}. \end{aligned}$$

À présent, en prenant, à k fixé, les limites $n \to \infty$ et $p \to 0$ de telle sorte que $np \to \lambda$, on voit que chacun des rapports $(n-i)/n$ converge vers 1, que $(np)^k$ converge vers λ^k, que $(1-p)^n$ converge vers $e^{-\lambda}$, et que $(1-p)^{-k}$ tend vers 1. Par conséquent,

$$\lim_{\substack{n\to\infty \\ p\to 0 \\ np\to\lambda}} f_X(k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

pour chaque $k \in \mathbb{N}$.

Loi géométrique

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 temps du premier succès

suit la **loi géométrique** de paramètre p ; $X \sim \text{geom}(p)$. La fonction de masse associée est donc donnée par

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

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

Lemme 2.1. *Soit X une variable aléatoire suivant une loi géométrique. Alors, pour tout $k \geq 1$,*

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

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 $(1-p)^n \sum_{m>0} p(1-p)^{m-1} = (1-p)^n$. □

Cette propriété dit par exemple que même si le numéro 53 (sur 90 numéros possibles) n'est pas sorti pendant 178 tirages consécutifs à la loterie, cela ne rend pas sa prochaine apparition plus probable [2].

Loi hypergéométrique

Une urne contient N boules, dont b sont bleues et $r = N - b$ sont rouges. Un échantillon de $n \leq N$ 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, $B \sim \text{hypergeom}(N, b, n)$, dont la fonction de masse est [3]

$$f_B(k) = \binom{b}{k}\binom{N-b}{n-k} \Big/ \binom{N}{n}, \qquad k \in \{(n-r) \vee 0, \ldots, b \wedge n\}.$$

Lemme 2.2. *Pour tout $0 \leq k \leq n$,*

$$\lim_{\substack{N,b\to\infty \\ b/N \to p}} f_B(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Démonstration. Exercice. □

Ce lemme montre qu'il est possible de remplacer la loi hypergéométrique de paramètres N,b et n par une loi binomiale de paramètres n et $p = b/N$ 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 le sondage. 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) $< N$(population sondée)$/10$, cette approximation est légitime.

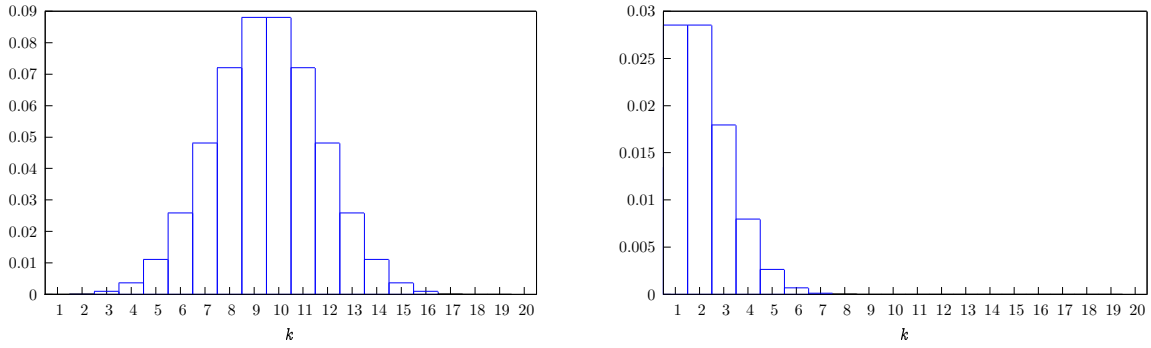


FIGURE 2.4: Loi de Pascal dans le cas $k + r = 20$ pour $p = 0{,}5$ (gauche) et $p = 0{,}1$ (droite).

Loi de Pascal

Si X représente le nombre d'échecs avant le $r^{\text{ème}}$ succès d'une suite d'épreuves de Bernoulli, alors X suit la **loi de Pascal** de paramètres r et p, $X \sim \text{pascal}(r,p)$, dont la fonction de masse est (pourquoi ?)

$$f_X(k) = \binom{k+r-1}{k} p^r (1-p)^k, \qquad k = 0,1,\ldots$$

On parle également de **loi binomiale négative** ou de **loi de Pólya** [4].

Dans certaines applications, il est utile d'autoriser le paramètre r à prendre des valeurs réelles positives pas nécessairement entières.

2.2 Indépendance de variables aléatoires

Rappelons que deux événements A et B sont indépendants si l'occurrence de A n'a pas d'influence sur 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 2.3. *Deux variables aléatoires X et Y sur un espace de probabilité $(\Omega,\mathbb{P})$ sont **indépendantes** si et seulement si les événements*

$$\{X \in A\} \text{ et } \{Y \in B\}$$

*sont indépendants pour tout $A,B \subset \mathbb{R}$. Plus généralement, une famille de variables aléatoires $(X_i)_{i\in I}$ est **indépendante** si les événements*

$$\{X_i \in A_i\}, i \in J,$$

sont indépendants pour tout $A_i \subset \mathbb{R}$, $i \in J$, et tout $J \subset I$ fini.

Le résultat suivant montre qu'il est suffisant de vérifier l'indépendance pour les singletons.

2. Cela s'est produit en 2005 en Italie. De très, très nombreux Italiens ont misé de grosses sommes, certains tout ce qu'ils possédaient. Le total des mises s'est élevé à 4 milliards d'euros, et cette histoire s'est terminée par de nombreuses ruines et même des suicides...

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

4. George Pólya (1887, Budapest – 1985, Palo Alto), mathématicien hongrois.

Lemme 2.3. *La famille $(X_i)_{i\in I}$ de variables aléatoires est indépendante si et seulement si les événements*

$$\{X_i = x_i\}, i \in J,$$

sont indépendants pour tout $x_i \in \mathbb{R}$, $i \in J$, et tout $J \subset I$ fini.

Démonstration. On considère le cas de deux événements A_1 et A_2 ; le cas général se traite de la même manière. On peut supposer, sans perte de généralité, que $A_1 \subset X_1(\Omega)$ et $A_2 \subset X_2(\Omega)$. Par σ-additivité,

$$\begin{aligned}
\mathbb{P}(X_1 \in A_1, X_2 \in A_2) &= \mathbb{P}\Big(\bigcup_{\substack{x_1\in A_1\\ x_2\in A_2}} \{X_1 = x_1, X_2 = x_2\}\Big) \\
&= \sum_{\substack{x_1\in A_1\\ x_2\in A_2}} \mathbb{P}(X_1 = x_1, X_2 = x_2) \\
&= \sum_{\substack{x_1\in A_1\\ x_2\in A_2}} \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2) \\
&= \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2).
\end{aligned}$$

□

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

Lemme 2.4. *Soient $(X_i)_{i\in I}$ une famille de variables aléatoires indépendantes, et $(\varphi_i)_{i\in I}$ une famille de fonctions de $\mathbb{R} \to \mathbb{R}$. 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

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

□

Définition 2.4. *Une famille de variables aléatoires $(X_i)_{i\in I}$ est dite* ***i.i.d.*** *($\equiv$ indépendantes et identiquement distribuées) si elle est indépendante et tous les X_i ont la même loi.*

2.3 Vecteurs aléatoires discrets

Soient X et Y deux variables aléatoires sur un même espace de probabilité $(\Omega,\mathbb{P})$. Les fonctions de masse f_X et f_Y encodent 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.

Exemple 2.4. On demande à deux élèves de faire deux jets à pile ou face chacun, 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, et en particulier $f_{X_1} = f_{X_2} = f_{Y_1} = f_{Y_2}$. Or ces couples ont des propriétés satistiques 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) prenant ses valeurs dans $\mathbb{R}^2$.

Exemple 2.5. Dans le cas de l'exemple précédent, on a alors

$$\mathbb{P}((X_1,X_2) = (x_1,x_2)) = \tfrac{1}{4}, \quad \forall x_1,x_2 \in \{0,1\},$$
$$\mathbb{P}((Y_1,Y_2) = (y_1,y_2)) = \begin{cases} \frac{1}{2}, & \text{si } y_1 = y_2 \in \{0,1\}, \\ 0 & \text{sinon.} \end{cases}$$

♢

Comme pour les variables aléatoires, un vecteur aléatoire induit naturellement une mesure de probabilité sur $\mathbb{R}^n$.

Définition 2.5. *On appelle* ***loi conjointe*** *du vecteur aléatoire* $\mathbf{X} = (X_1,\dots,X_n)$ *la mesure de probabilité sur* $\mathbb{R}^n$ *définie par*

$$\mathbb{P}_{\mathbf{X}}(A) = \mathbb{P}(\mathbf{X} \in A) \equiv \mathbb{P}(\mathbf{X}^{-1}(A)), \qquad \forall A \subset \mathbb{R}^n.$$

Comme pour les variables aléatoires discrètes, la loi conjointe d'un vecteur aléatoire $\mathbf{X}$ est caractérisée par la fonction de masse conjointe.

Définition 2.6. *La* ***fonction de masse conjointe*** *d'un vecteur aléatoire discret* $\mathbf{X} = (X_1,\dots,X_n)$ *est la fonction* $f_{\mathbf{X}} : \mathbb{R}^n \to [0,1]$ *définie par*

$$f_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} = \mathbf{x}), \qquad \forall \mathbf{x} \in \mathbb{R}^n.$$

Définition 2.7. *Étant donnée une fonction de masse conjointe* $f_{(X_1,\dots,X_n)}$*, on appelle* ***fonctions de masse marginales*** *les fonctions de masse* f_{X_i}.

Le lemme suivant montre comment on peut récupérer les fonctions de masse marginales à partir de la fonction de masse conjointe.

Lemme 2.5.

$$f_{X_i}(x_i) = \sum_{x_1,\dots,x_{i-1},x_{i+1},\dots,x_n} f_{(X_1,\dots,X_n)}(x_1,\dots,x_n).$$

Démonstration. Laissée en exercice. □

L'indépendance de la famille $X_1,\dots,X_n$ se formule aisément en termes de la fonction de masse conjointe du vecteur correspondant.

Lemme 2.6. *La famille* $X_1,\dots,X_n$ *de variables aléatoires discrètes est indépendante si et seulement si*

$$f_{(X_1,\dots,X_n)}(x_1,\dots,x_n) = f_{X_1}(x_1)\cdots f_{X_n}(x_n), \qquad \forall (x_1,\dots,x_n) \in \mathbb{R}^n.$$

Démonstration. L'affirmation suit immédiatement des identités

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

et du Lemme 2.3. □

2.4 Espérance, variance, covariance et moments

2.4.1 Espérance

On répète N fois une expérience, 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 l'on a noté E l'ensemble des valeurs possibles (supposé discret) et $N(x)$ le nombre d'expériences ayant donné le nombre x. Supposons qu'on modélise cette expérience par une famille $X_1,\dots,X_n$ de variables aléatoires discrètes indépendantes de même fonction de masse f. On s'attend alors à ce que, pour chaque valeur $x \in E$, la fraction $N(x)/N$ soit proche de la probabilité $f(x)$. Par conséquent, $\sum_{x\in E} x f(x)$ devrait fournir une approximation asymptotiquement correcte de m ; on appelle la quantité correspondante espérance.

Définition 2.8. *Soit X une variable aléatoire discrète et soit f_X sa fonction de masse. On dit que X* ***admet une espérance*** *si*

$$\sum_{x\in X(\Omega)} |x| f_X(x) < \infty.$$

*Dans ce cas on définit l'**espérance** de X par*

$$\mathbb{E}(X) = \sum_{x\in X(\Omega)} x f_X(x).$$

Remarque 2.2. *La condition d'absolue sommabilité est importante : elle garantit que l'espérance ne dépend pas de l'ordre dans lequel les termes sont sommés.*

La seule exception est lorsque la variable aléatoire possède un signe bien défini. Dans ce cas, si cette dernière n'est pas absolument sommable, on définit l'espérance comme étant égale à $+\infty$, resp. $-\infty$, pour une variable aléatoire positive, resp. négative.

Remarque 2.3. *Si l'espace de probabilité sous-jacent est caractérisé par la fonction de masse f, on peut écrire*

$$\mathbb{E}(X) = \sum_{x\in X(\Omega)} x f_X(x) = \sum_{x\in X(\Omega)} x \sum_{\substack{\omega\in\Omega \\ X(\omega)=x}} f(\omega) = \sum_{\omega\in\Omega} X(\omega) f(\omega)\,. \tag{2.1}$$

Remarque 2.4. *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 CHF si le dé indique n. Dans ce cas, le joueur va recevoir en moyenne 3,5 *CHF. 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^{ème} *siècle) : on jette une pièce de monnaie jusqu'à l'apparition du premier « face » ; si cela a lieu au* $T^{ème}$ *lancer, votre gain sera de* 2^T *francs. Quelle serait une mise équitable ? On vérifie facilement que l'espérance 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**.

Le résultat élémentaire suivant est extrêmement utile.

Lemme 2.7. *Soit* $A,B \subset \Omega$. *Alors,* $\mathbb{P}(A) = \mathbb{E}(\mathbf{1}_A)$ *et* $\mathbb{P}(A \cap B) = \mathbb{E}(\mathbf{1}_A \mathbf{1}_B)$.

Démonstration. Laissée en exercice. □

Démontrons à présent quelques propriétés importantes de l'espérance.

Lemme 2.8. *Soient* X, Y *et* X_n, $n \geq 1$, *des variables aléatoires possédant une espérance. Alors,*

1. $X \geq 0 \implies \mathbb{E}(X) \geq 0$.
2. *Si* $\mathbb{P}(X = c) = 1$ *pour un* $c \in \mathbb{R}$, *alors* $\mathbb{E}(X) = c$.
3. $\mathbb{E}(|X|) \geq |\mathbb{E}(X)|$.
4. *(Linéarité) Pour tout* $\alpha,\beta \in \mathbb{R}$, $\alpha X + \beta Y$ *possède une espérance et*
$$\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}(X) + \beta \mathbb{E}(Y).$$
5. *(*σ*-additivité) Supposons que les variables aléatoires* X_n, $n \geq 1$, *soient positives et que* $X = \sum_{n \geq 1} X_n$. *Alors,*
$$\mathbb{E}(X) = \sum_{n \geq 1} \mathbb{E}(X_n).$$
6. *(Convergence monotone) Lorsque les variables aléatoires* X_n, $n \geq 1$, *satisfont* $X_n \nearrow X$ *ponctuellement, on a*
$$\mathbb{E}(X) = \lim_{n \to \infty} \mathbb{E}(X_n).$$

Démonstration. 1. et *2.* sont immédiats.

3. Notons $X_+(\Omega) = \{x \in X(\Omega) : x \geq 0\}$, $X_-(\Omega) = \{x \in X(\Omega) : x < 0\}$. Alors,

$$\begin{aligned}
|\mathbb{E}(X)| = \Big| \sum_{x \in X(\Omega)} x\, f_X(x) \Big| &= \Big| \sum_{x \in X_-(\Omega)} x\, f_X(x) + \sum_{x \in X_+(\Omega)} x\, f_X(x) \Big| \\
&\leq - \sum_{x \in X_-(\Omega)} x\, f_X(x) + \sum_{x \in X_+(\Omega)} x\, f_X(x) \\
&= \sum_{x \in -X_-(\Omega)} x\, f_X(-x) + \sum_{x \in X_+(\Omega)} x\, f_X(x) \\
&= \sum_{x \in |X|(\Omega)} x \left(f_X(x) + f_X(-x) \right) \\
&= \sum_{x \in |X|(\Omega)} x\, f_{|X|}(x) = \mathbb{E}(|X|),
\end{aligned}$$

puisque $|X|(\Omega) = (-X_-(\Omega)) \cup X_+(\Omega)$, et $f_{|X|}(x) = \mathbb{P}(|X| = x) = \mathbb{P}(X = x) + \mathbb{P}(X = -x) = f_X(x) + f_X(-x)$ lorsque $x > 0$.

4. On écrit, avec $E = X(\Omega)$, $F = Y(\Omega)$, $U = \{u = \alpha x + \beta y \,:\, x \in E, y \in F\}$,

$$\begin{aligned}
\sum_{u\in U} |u|\mathbb{P}(\alpha X + \beta Y = u) &= \sum_{u\in U} |u| \sum_{\substack{x\in E, y\in F\\ \alpha x+\beta y = u}} \mathbb{P}(X = x, Y = y)\\
&= \sum_{x\in E, y\in F} \mathbb{P}(X = x, Y = y) \sum_{\substack{u\in U\\ u=\alpha x+\beta y}} |u|\\
&= \sum_{x\in E, y\in F} |\alpha x + \beta y|\mathbb{P}(X = x, Y = y)\\
&\le |\alpha| \sum_{x\in E} |x| \sum_{y\in F} \mathbb{P}(X = x, Y = y) + |\beta| \sum_{y\in F} |y| \sum_{x\in E} \mathbb{P}(X = x, Y = y)\\
&= |\alpha| \sum_{x\in E} |x|\mathbb{P}(X = x) + |\beta| \sum_{y\in F} |y|\mathbb{P}(Y = y)\\
&= |\alpha|\mathbb{E}(|X|) + |\beta|\mathbb{E}(|Y|) < \infty.
\end{aligned}$$

$\alpha X + \beta Y$ possède donc une espérance. En répétant le même calcul sans les valeurs absolues, on obtient le résultat.

5. Notons $S_n = \sum_{i=1}^n X_i$. On observe tout d'abord que $X - S_n \ge 0$, pour tout $n \ge 1$, et donc, par *1* et *4*, $\mathbb{E}(X) - \mathbb{E}(S_n) = \mathbb{E}(X - S_n) \ge 0$, c'est-à-dire

$$\mathbb{E}(X) \ge \mathbb{E}(S_n) = \mathbb{E}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathbb{E}(X_i).$$

En laissant $n \to \infty$, on obtient donc $\mathbb{E}(X) \ge \sum_{i=1}^\infty \mathbb{E}(X_i)$.

Il nous reste à obtenir l'inégalité opposée. Pour ce faire, on fixe $0 < c < 1$ et on introduit la variable aléatoire $\tau = \inf\{n \ge 1 \,:\, S_n \ge cX\}$. Comme $S_n \nearrow X < \infty$, on en conclut que $\tau < \infty$. La somme aléatoire $S_\tau = \sum_{i=1}^\tau X_i$ est une variable aléatoire discrète puisque l'ensemble des valeurs prises par S_τ est inclus dans l'ensemble $S(\Omega) \equiv \bigcup_{n\ge1} S_n(\Omega)$ et que les ensembles $S_n(\Omega)$ sont dénombrables. Par conséquent, il suit de *1.* et de la σ-additivité de $\mathbb{P}$ que

$$c\mathbb{E}(X) \le \mathbb{E}(S_\tau) = \sum_{s\in S(\Omega)} s \sum_{n\ge1} \mathbb{P}(\tau = n, S_n = s) = \sum_{n\ge1} \mathbb{E}(\mathbf{1}_{\{\tau=n\}} S_n),$$

l'interchange des sommes étant justifié puisque tous les termes sont positifs. Similairement, il suit également de *4.* que

$$\begin{aligned}
\sum_{n\ge1} \mathbb{E}(\mathbf{1}_{\{\tau=n\}} S_n) &= \sum_{n\ge1} \sum_{k=1}^n \mathbb{E}(\mathbf{1}_{\{\tau=n\}} X_k) = \sum_{n\ge1} \sum_{k=1}^n \sum_{x\in X_k(\Omega)} x\mathbb{P}(\tau = n, X_k = x)\\
&= \sum_{k\ge1} \sum_{n\ge k} \sum_{x\in X_k(\Omega)} x\mathbb{P}(\tau = n, X_k = x) = \sum_{k\ge1} \sum_{x\in X_k(\Omega)} x\mathbb{P}(\tau \ge k, X_k = x) \le \sum_{k\ge1} \mathbb{E}(X_k).
\end{aligned}$$

La conclusion suit en prenant la limite $c \to 1$.

6. Soit $Y_n = X_{n+1} - X_n \ge 0$. D'une part,

$$\mathbb{E}(\sum_{n\ge1} Y_n) = \mathbb{E}(\lim_{N\to\infty} \sum_{n=1}^N Y_n) = \mathbb{E}(\lim_{N\to\infty} X_{N+1} - X_1) = \mathbb{E}(X - X_1) = \mathbb{E}(X) - \mathbb{E}(X_1),$$

et, d'autre part,

$$\sum_{n\ge1} \mathbb{E}(Y_n) = \lim_{N\to\infty} \sum_{n=1}^N \mathbb{E}(Y_n) = \lim_{N\to\infty} \mathbb{E}(X_{N+1} - X_1) = \lim_{n\to\infty} \mathbb{E}(X_n) - \mathbb{E}(X_1).$$

Loi	Espérance	Variance
Bernoulli (p)	p	$p(1-p)$
Binomiale (n,p)	np	$np(1-p)$
Poisson (λ)	λ	λ
Géométrique (p)	$1/p$	$(1-p)/p^2$
Hypergéométrique (N,b,n)	bn/N	$nb(N-b)(N-n)/(N^3-N^2)$
Pascal (r,p)	$r(1-p)/p.$	$r(1-p)/p^2$

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

Par conséquent, il suit de *5.* que

$$\lim_{n\to\infty} \mathbb{E}(X_n) - \mathbb{E}(X_1) = \sum_{n\geq 1} \mathbb{E}(Y_n) = \mathbb{E}(\sum_{n\geq 1} Y_n) = \mathbb{E}(X) - \mathbb{E}(X_1).$$

□

Exemple 2.6. On désire trouver le nombre $a \in \mathbb{R}$ qui approxime le mieux une variable aléatoire 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 la valeur de a réalisant le minimum satisfait $-2\mathbb{E}(X) + 2a = 0$, ce qui implique que $a = \mathbb{E}(X)$. ♢

Exemple 2.7. On appelle triangle d'un graphe, un triplet de sommets x,y,z tels que $x \sim y$, $y \sim z$ et $z \sim x$. Quel est l'espérance du nombre de triangles $K_\triangle$ dans le graphe aléatoire $\mathcal{G}(n,m)$? Il suit de la linéarité et du Lemme 2.7 que

$$\mathbb{E}(K_\triangle) = \mathbb{E}(\sum_{\substack{x,y,z \\ \text{distincts}}} \mathbf{1}_{\{x\sim y, y\sim z, z\sim x\}}) = \sum_{\substack{x,y,z \\ \text{distincts}}} \mathbb{P}(x\sim y, y\sim z, z\sim x).$$

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

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

♢

Donnons à présent l'espérance pour les lois introduites plus tôt dans ce chapitre. Observez que 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 2.9. *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 suivant une loi de Bernoulli de paramètre p sur $\{0,1\}$ 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 suivant une loi binomiale de paramètres n et p est d'utiliser le Lemme 2.8, point 1. On peut en effet écrire $X = X_1 + \ldots + X_n$, où les X_i sont des variables de Bernoulli.

En d'autres termes, on exprime X comme le nombre total de succès après n épreuves de Bernoulli. On a alors

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(X_i) = np.$$

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 possible d'interchanger sommation et dérivation. Par conséquent,

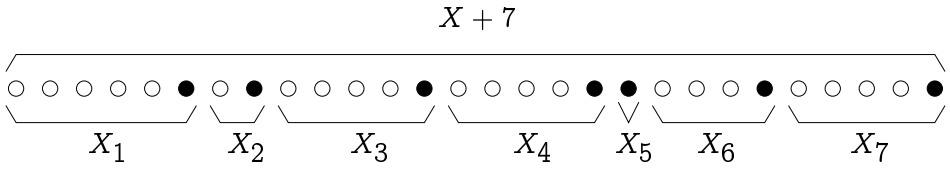
$$G'(x) = \frac{1}{(1-x)^2} = \sum_{k=1}^{\infty} kx^{k-1}.$$

On a donc

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

5. *Loi hypergéométrique.* Nous calculerons l'espérance d'une variable hypergéométrique dans l'Exemple 4.2.

6. *Loi de Pascal.* Si X suit une loi de Pascal de paramètres r et p, on peut la décomposer en $X + r = X_1 + \cdots + X_r$, où les X_i suivent chacun une loi géométrique de paramètre p. Par exemple, pour $r = 7$ (les ronds blancs représentent les échecs, les noirs les succès) :



On a donc

$$\mathbb{E}(X) = \mathbb{E}(X + r) - r = \sum_{i=1}^{r} \mathbb{E}(X_i) - r = \frac{r}{p} - r = \frac{1-p}{p}\, r.$$

□

Exemple 2.8. 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 contenant chacune 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, et ensuite décider si vous préférez garder l'enveloppe choisie, ou prendre plutôt 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 parfois utile.

Lemme 2.10. *Soit X une variable aléatoire à valeurs dans $\mathbb{N}$. Alors,*

$$\mathbb{E}(X) = \sum_{n\geq 0} \mathbb{P}(X > n).$$

Démonstration. Il suffit d'observer que

$$\mathbb{E}(X) = \sum_{m\geq 1} m\mathbb{P}(X = m) = \sum_{m\geq 1} \sum_{n=0}^{m-1} \mathbb{P}(X = m) = \sum_{n\geq 0} \sum_{m=n+1}^{\infty} \mathbb{P}(X = m) = \sum_{n\geq 0} \mathbb{P}(X > n).$$

□

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ un vecteur aléatoire discret et $\varphi : \mathbb{R}^n \to \mathbb{R}$. Dans ce cas, $\varphi(\mathbf{X})$ définit une variable aléatoire discrète. Le résultat suivant permet de déterminer aisément son espérance.

Lemme 2.11. *Soit $\mathbf{X} = (X_1, \ldots, X_n)$ un vecteur aléatoire discret et $\varphi : \mathbb{R}^n \to \mathbb{R}$. Alors,*

$$\mathbb{E}(\varphi(\mathbf{X})) = \sum_{\mathbf{x}\in\mathbf{X}(\Omega)} \varphi(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}),$$

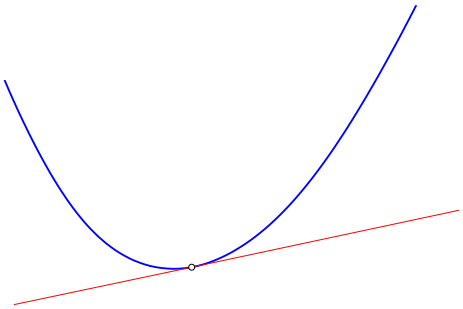
dès que cette somme est absolument convergente.

Démonstration. Notons $E = \mathbf{X}(\Omega)$, $F = \varphi(E)$ et $Y = \varphi(\mathbf{X})$. On a

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_{y\in F} y\,\mathbb{P}(Y = y) = \sum_{y\in F} y\,\mathbb{P}(\varphi(\mathbf{X}) = y) \\
&= \sum_{y\in F} y\,\mathbb{P}(\mathbf{X} \in \varphi^{-1}(y)) = \sum_{y\in F} y \sum_{\mathbf{x}\in\varphi^{-1}(y)} \mathbb{P}(\mathbf{X} = \mathbf{x}) \\
&= \sum_{\substack{y\in F, \mathbf{x}\in E \\ \varphi(\mathbf{x})=y}} y\,\mathbb{P}(\mathbf{X} = \mathbf{x}) = \sum_{\mathbf{x}\in E} \varphi(\mathbf{x})\,\mathbb{P}(\mathbf{X} = \mathbf{x}).
\end{aligned}$$

Observez que la convergence absolue de la série est cruciale pour pouvoir réorganiser les termes comme on l'a fait. □

Définition 2.9. *Une fonction $\varphi : \mathbb{R} \to \mathbb{R}$ est* ***convexe*** *si et seulement si : $\forall x \in \mathbb{R}, \exists a \in \mathbb{R}$: $\forall y \in \mathbb{R}, \varphi(y) \geq \varphi(x) + a(y - x)$. Si l'inégalité est toujours stricte lorsque $y \neq x$, alors on dit que φ est* ***strictement convexe*** *.*



Théorème 2.1 (Inégalité de Jensen[5]). *Soient X une variable aléatoire admettant une espérance et $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction convexe. Alors*

$$\mathbb{E}(\varphi(X)) \geq \varphi(\mathbb{E}(X)).$$

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 $y \in \mathbb{R}$,

$$\varphi(y) \geq \varphi(\mathbb{E}(X)) + a(y - \mathbb{E}(X)).$$

Il suit donc du Lemme 2.11 que

$$\mathbb{E}(\varphi(X)) = \sum_{y \in X(\Omega)} \varphi(y) f_{\mathbf{X}}(y) \geq \varphi(\mathbb{E}(X)) + a(\mathbb{E}(X) - \mathbb{E}(X)) = \varphi(\mathbb{E}(X)).$$

□

2.4.2 Variance, moments d'ordres supérieurs

Définition 2.10. *On appelle $\mathbb{E}(X^n)$ le* ***moment d'ordre*** *n de la variable aléatoire X, pourvu que cette espérance soit bien définie.*

Remarque 2.5. *Si une variable aléatoire possède un moment d'ordre n, alors elle possède également tous les moments d'ordre $1 \leq k < n$. En effet, l'inégalité de Jensen implique que*

$$\infty > \mathbb{E}(|X|^n) = \mathbb{E}((|X|^k)^{n/k}) \geq \mathbb{E}(|X|^k)^{n/k},$$

puisque la fonction $x \mapsto x^{n/k}$ est convexe lorsque $n \geq k$.

Remarque 2.6. *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 : deux variables aléatoires X et Y satisfaisant $\mathbb{E}(e^{\lambda X}) < \infty$ et $\mathbb{E}(e^{\lambda Y}) < \infty$, $\forall \lambda \in \mathbb{R}$, et telles que $\mathbb{E}(X^n) = \mathbb{E}(Y^n)$, pour tout $n \in \mathbb{N}$, ont la même loi.*

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.

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

Définition 2.11. *Soit X une variable aléatoire dont l'espérance existe. On appelle* **variance** *de X la quantité*

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

(la variance de X peut être infinie). On appelle **écart-type** *de X la quantité $\sigma(X) = \sqrt{\mathrm{Var}(X)}$.*

Lemme 2.12. *1. $\mathrm{Var}(X) \geq 0$, 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 $\mathbb{E}(X^2) < \infty$.

3. Si $\mathrm{Var}(X) < \infty$, alors $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.

4. Pour $a,b \in \mathbb{R}$, $\mathrm{Var}(a + bX) = b^2\mathrm{Var}(X)$.

5. Si $\mathrm{Var}(X) < \infty$ et $\mathrm{Var}(Y) < \infty$, alors $\mathrm{Var}(X + Y) < \infty$.

Démonstration. Nous ne démontrerons que deux des affirmations, les autres étant immédiates.
Preuve de 2. Soit Z une variable aléatoire telle que $\mathbb{E}(Z^2) < \infty$. Alors, pour tout $a \in \mathbb{R}$,

$$\mathbb{E}((Z-a)^2) = \mathbb{E}((Z-a)^2 \mathbf{1}_{\{|Z|<2|a|\}}) + \mathbb{E}((Z-a)^2 \mathbf{1}_{\{|Z|\geq 2|a|\}}) \leq 9a^2 + \frac{9}{4}\mathbb{E}(Z^2) < \infty.$$

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 $(a+b)^2 \leq 2(a^2+b^2)$, pour tout $a,b \in \mathbb{R}$, on peut écrire

$$\mathrm{Var}(X+Y) = \mathbb{E}((\check{X}+\check{Y})^2) \leq 2\mathbb{E}(\check{X}^2) + 2\mathbb{E}(\check{Y}^2) = 2\mathrm{Var}(X) + 2\mathrm{Var}(Y) < \infty.$$

□

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 2.13 (Inégalité de Bienaymé[6]-Tchebychev[7])**.**

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}, \qquad \forall a > 0. \tag{2.2}$$

Démonstration. Notons $\check{X} = X - \mathbb{E}(X)$. Il suffit d'observer que

$$\mathrm{Var}(X) = \mathbb{E}(\check{X}^2) \geq \mathbb{E}(\check{X}^2 \mathbf{1}_{\{\check{X}^2 \geq a^2\}}) \geq a^2\,\mathbb{P}(\check{X}^2 \geq a^2) = a^2\,\mathbb{P}(|X - \mathbb{E}(X)| \geq a).$$

□

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

Lemme 2.14. *La table 2.1 donne les variances des principales lois introduites précédemment.*

Démonstration. 1. *Loi de Bernoulli.* La variance d'une variable aléatoire X suivant une loi de Bernoulli de paramètre p sur $\{0,1\}$ est immédiate à calculer :

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 1 \cdot p + 0 \cdot (1-p) - p^2 = p(1-p).$$

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

2. *Loi binomiale.* Voir l'Exemple 2.10.
3. *Loi de Poisson.* Une façon de calculer la variance d'une variable aléatoire X suivant une loi de Poisson est la suivante.

$$\mathbb{E}(X(X-1)) = \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda}\lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = \lambda^2.$$

   Par conséquent, $\mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}(X(X-1)) - \mathbb{E}(X)^2 + \mathbb{E}(X) = \lambda$.
4. *Loi géométrique.* Le second moment d'une variable aléatoire X de loi géométrique est donné par la série

$$\mathbb{E}(X^2) = \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1}.$$

   Pour en 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,

$$\mathrm{Var}(X) = p(1-p)G''(1-p) + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

5. *Loi hypergéométrique.* Voir l'Exemple 4.2.
6. *Loi de Pascal.* Voir l'Exemple 2.10.

□

2.4.3 Covariance et corrélation

En général, $\mathrm{Var}(X+Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$: en effet, un bref calcul montre que

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))).$$

Ceci motive la définition suivante.

Définition 2.12. *On appelle* ***covariance*** *de deux variables aléatoires X et Y la quantité*

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

En particulier,

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

Deux variables aléatoires X et Y sont ***non-corrélées*** *si $\mathrm{Cov}(X,Y) = 0$; dans ce cas, on a $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.*

Attention : la variance n'est pas un opérateur linéaire, même restreint aux variables aléatoires non-corrélées (se souvenir que $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$).

Lemme 2.15. *1.* $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.

2. La covariance est une forme bilinéaire : pour $a,b \in \mathbb{R}$,

$$\mathrm{Cov}(aX,bY) = a\,b\,\mathrm{Cov}(X,Y),$$
$$\mathrm{Cov}(X_1 + X_2,Y) = \mathrm{Cov}(X_1,Y) + \mathrm{Cov}(X_2,Y).$$

3. Pour des variables $X_1,\ldots,X_n$, *on a*

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

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 2.13. *On appelle* ***coefficient de corrélation*** *de deux variables aléatoires* X *et* Y *de variances non-nulles la quantité*

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

Théorème 2.2 (Inégalité de Cauchy-Schwarz)**.**

$$\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2),$$

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 le théorème est trivial). Fixons $b \in \mathbb{R} \setminus \{0\}$. Dans ce cas, on a, pour tout $a \in \mathbb{R}$,

$$a^2\mathbb{E}(X^2) - 2ab\mathbb{E}(XY) + b^2\mathbb{E}(Y^2) = \mathbb{E}((aX - bY)^2) \geq 0.$$

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

$$\mathbb{E}(XY)^2 - \mathbb{E}(X^2)\mathbb{E}(Y^2) \leq 0.$$

Le discriminant est nul si et seulement si il y a un unique zéro, ce qui ne peut avoir lieu que s'il existe $a \in \mathbb{R}$ tels que

$$\mathbb{E}((aX - bY)^2) = 0.$$

□

Il suit de ce théorème que la valeur absolue du coefficient de corrélation est égal à 1 si et seulement si il existe une relation linéaire entre les variables aléatoires.

Corollaire 2.1.

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

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 $X - \mathbb{E}(X)$ et $Y - \mathbb{E}(Y)$. □

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.

Lemme 2.16. *Pour toute paire de variables aléatoires X et Y, on a*

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

et 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, $\check{X} = X - \mathbb{E}(X)$ et $\check{Y} = Y - \mathbb{E}(Y)$, on a

$$\mathbb{E}\left((Y - aX - b)^2\right) = \mathbb{E}\left((\check{Y} - a\check{X} - \check{b})^2\right),$$

où on a posé $\check{b} = b + a\mathbb{E}(X) - \mathbb{E}(Y)$. On vérifie alors aisément que

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

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 2.9. En physiologie, la loi de Kleiber[8] 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 dimensionalité 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 $X = \log M$ et $Y = \log P$, à vérifier qu'il y a une relation linéaire entre X et Y. 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.) ♢

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

2.4.4 Extension aux vecteurs aléatoires

Les notions d'espérance et de covariance s'étendent de façon naturelle aux vecteurs aléatoires.

Définition 2.14. *L'**espérance** du vecteur aléatoire $\mathbf{X} = (X_1, \dots, X_n)$ est le vecteur $\mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), \dots, \mathbb{E}(X_n))$, à condition que chacune de ces espérances existe.*

Définition 2.15. *Soient $\mathbf{X} = (X_1, \dots, X_n)$ et $\mathbf{Y} = (Y_1, \dots, Y_n)$ deux vecteurs aléatoires. Leur **matrice de covariance** est la matrice $n \times n$ $\mathrm{Cov}(\mathbf{X}, \mathbf{Y})$ dont l'élément i,j est donné par*

$$\mathrm{Cov}(X_i, Y_j),$$

pour $1 \leq i, j \leq n$.

2.4.5 Absence de corrélation et indépendance

Voyons à présent quel est le lien entre indépendance et absence de corrélation.

Lemme 2.17. *Deux variables aléatoires indépendantes dont l'espérance existe sont non-corrélées.*
En particulier, si $X_1, \dots, X_n$ sont 2 à 2 indépendantes, $\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i)$.

Démonstration. On applique le Lemme 2.11 avec la fonction $\varphi : \mathbb{R}^2 \to \mathbb{R}$, $\varphi(x, y) = xy$. Cela donne

$$\begin{aligned}
\mathbb{E}(XY) = \mathbb{E}(\varphi(X, Y)) &= \sum_{x \in X(\Omega), y \in Y(\Omega)} \varphi(x, y) f_{(X,Y)}(x, y) \\
&= \sum_{x \in X(\Omega), y \in Y(\Omega)} \varphi(x, y) f_X(x) f_Y(y) \\
&= \sum_{x \in X(\Omega), y \in Y(\Omega)} xy f_X(x) f_Y(y) = \mathbb{E}(X)\mathbb{E}(Y).
\end{aligned}$$

□

Exemple 2.10.

1. *Loi binomiale.* On a vu qu'une variable aléatoire X suivant une loi binomiale de paramètres n et p pouvait s'écrire $X = X_1 + \cdots + X_n$, où les X_i sont des variables de Bernoulli indépendantes de paramètre p. On obtient donc immédiatement que

$$\mathrm{Var}(X) = np(1 - p).$$

2. *Loi de Pascal.* On a également vu qu'une variable aléatoire X suivant une loi de Pascal de paramètres r et p pouvait s'écrire $X + r = X_1 + \cdots + X_r$, où les X_i sont des variables géométriques indépendantes de paramètre p. On obtient donc immédiatement que

$$\mathrm{Var}(X) = \mathrm{Var}(X + r) = r \frac{1 - p}{p^2}.$$

♢

Nous avons vu que deux variables aléatoires indépendantes sont toujours non-corrélées. La réciproque est fausse en général, comme le montre l'exemple suivant.

Exemple 2.11. Considérons $\Omega = \{-1, 0, 1\}$ avec la distribution uniforme. Soient $X(\omega) = \omega$ et $Y(\omega) = |\omega|$ 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. ♢

Dire que X et Y sont indépendants est donc strictement plus fort en général que de demander à ce que $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. Le résultat suivant montre comment il faut renforcer cette dernière propriété pour obtenir l'indépendance.

Lemme 2.18. *Soit $(X_i)_{i\in I}$ une famille de variables aléatoires discrètes. Les propositions suivantes sont équivalentes :*

1. *$(X_i)_{i\in I}$ est indépendante ;*
2. *$\forall \varphi_i : \mathbb{R} \to \mathbb{R}$ telles que $\varphi_i(X_i)$, $i \in I$, admette une espérance,*

$$\mathbb{E}(\prod_{i\in J} \varphi_i(X_i)) = \prod_{i\in J} \mathbb{E}(\varphi_i(X_i)),$$

pour tout $J \subseteq I$ fini.

Démonstration. *1.* $\implies$ *2.* Cela suit immédiatement du Lemme 2.11 et de la factorisation de la fonction de masse conjointe : pour tout $J = \{i_1, \dots, i_n\} \subseteq I$,

$$\begin{aligned}
\mathbb{E}(\prod_{i\in J} \varphi_i(X_i)) &= \sum_{\substack{x_i \in X_i(\Omega) \\ i\in J}} z\varphi_{i_1}(x_{i_1}) \cdots \varphi_{i_n}(x_{i_n})\, f_{(X_{i_1},\dots,X_{i_n})}(x_{i_1}, \dots, x_{i_n}) \\
&= \sum_{\substack{x_i \in X_i(\Omega) \\ i\in J}} \varphi_{i_1}(x_{i_1}) \cdots \varphi_{i_n}(x_{i_n})\, f_{X_{i_1}}(x_{i_1}) \cdots f_{X_{i_n}}(x_{i_n}) \\
&= \prod_{i\in J} \mathbb{E}(\varphi_i(X_i)).
\end{aligned}$$

2. $\implies$ *1.* En appliquant *2.* à $\varphi_i(y) = \mathbf{1}_{\{y\in A_i\}}$, on obtient

$$\mathbb{P}(X_i \in A_i, \forall i \in J) = \mathbb{E}(\prod_{i\in J} \mathbf{1}_{\{X_i\in A_i\}}) = \prod_{i\in J} \mathbb{E}(\mathbf{1}_{\{X_i\in A_i\}}) = \prod_{i\in J} \mathbb{P}(X_i \in A_i).$$

□

2.4.6 Une première version de la loi des grands nombres

Nous avons motivé l'espérance comme étant une approximation de la moyenne des résultats obtenus en mesurant X lors d'une suite d'expériences aléatoires. Nous allons à présent rendre cela un peu plus précis.

Définition 2.16. *Soient $X_1, X_2, \dots, X_n$ une famille de variables aléatoires. Leur* ***moyenne empirique*** *est la variable aléatoire*

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

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

Théorème 2.3 (Loi faible des grands nombres). *Soient $X_1, \dots, X_n$ des variables aléatoires non-corrélées, de même espérance μ et de même variance $\sigma^2 < \infty$. Alors, pour tout $\epsilon > 0$,*

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2 n}.$$

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

Démonstration. Les variables aléatoires X_i étant non-corrélées, il est facile de déterminer la variance de S_n :

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n}.$$

Le résultat suit donc de l'inégalité de Bienaymé-Tchebychev (2.2) :

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2 n}.$$

□

Exemple 2.12. On effectue 10 000 lancers d'une pièce de monnaie équilibrée. Afin de travailler avec des variables centrées, on encode le résultat du $k^{\text{ème}}$ 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 $\bar{X}_n \in [-\epsilon, \epsilon]$ avec grande probabilité lorsque n est suffisamment grand. L'estimée dans la preuve du théorème nous donne

$$\mathbb{P}(|\bar{X}_n| \geq \epsilon) \leq \frac{1}{n\epsilon^2}.$$

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

$$\mathbb{P}(|\bar{X}_{10\ 000}| \geq 0{,}1) \leq \frac{1}{100}.$$

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 $\mathbb{P}(|\bar{X}_{10\ 000}| \geq 0{,}1) \leq 3{,}5 \cdot 10^{-22}$). ♢

Remarque 2.7 (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 que l'on désire 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 $(\Omega,\mathbb{P})$, que l'on répète N fois, de façon indépendante, obtenant une suite de résultats $(\omega_1,\omega_2,\ldots,\omega_N)$. Alors, pour tout événement A, les variables aléatoires $Y_k(\omega_1,\ldots,\omega_N) = \mathbf{1}_A(\omega_k)$ sont i.i.d., avec $\mathbb{E}(Y_k) = \mathbb{P}(A)$. Par conséquent, si l'on note $N(A) = \#\{1 \leq k \leq N : \omega_k \in A\} = \sum_{k=1}^{N} Y_k$ le nombre d'expériences lors desquelles l'événement A est réalisé, on a, pour tout $\epsilon > 0$,*

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

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 7.

2.4.7 Espérance conditionnelle

Soient X et Y deux variables aléatoires discrètes sur $(\Omega,\mathbb{P})$. La notion de probabilité conditionnelle $\mathbb{P}(A \mid B)$, où A et B sont deux événements, peut être étendue à la situation où l'on désire déterminer la loi de Y étant donnée la valeur prise par X.

Définition 2.17. *Soient X,Y deux variables aléatoires discrètes. La* ***fonction de masse conditionnelle*** *de Y sachant que $X = x$ est la fonction $f_{Y|X}(\cdot \mid x) : \mathbb{R} \to [0,1]$ définie par*

$$f_{Y|X}(y \mid x) = \mathbb{P}(Y = y \mid X = x) = \frac{f_{(X,Y)}(x,y)}{f_X(x)},$$

pour tout x tel que $f_X(x) > 0$. La loi correspondante s'appelle la ***loi conditionnelle*** *de Y sachant que $X = x$.*

Étant en possession d'une notion de loi conditionnelle, on peut définir l'espérance conditionnelle, comme étant l'espérance sous la loi conditionnelle.

Définition 2.18. *Soient X,Y deux variables aléatoires discrètes. On appelle* ***espérance conditionnelle*** *de Y étant donné X la variable aléatoire*

$$\mathbb{E}(Y \mid X)(\cdot) \equiv \mathbb{E}(Y \mid X = \cdot\,) = \sum_{y \in Y(\Omega)} y\, f_{Y|X}(y \mid \cdot\,),$$

pourvu que $\sum_{y \in Y(\Omega)} |y|\, f_{Y|X}(y \mid \cdot\,) < \infty$.

Insistons bien sur le fait que l'espérance conditionnelle $\mathbb{E}(Y \mid X)$ *n'est pas un nombre, mais une variable aléatoire* ; il s'agit, en fait, d'une fonction de la variable aléatoire X. Elle possède l'importante propriété suivante.

Lemme 2.19. *L'espérance conditionnelle $\mathbb{E}(Y \mid X)$ satisfait*

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

Plus généralement, pour toute fonction $\varphi : \mathbb{R} \to \mathbb{R}$ telle que les espérances existent,

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

Démonstration. La première affirmation est un cas particulier de la seconde : il suffit de choisir $\varphi \equiv 1$. Démontrons donc la seconde affirmation. Il suit du Lemme 2.11 que

$$\begin{aligned}\mathbb{E}(\mathbb{E}(Y \mid X)\,\varphi(X)) &= \sum_{x,y} y\, f_{Y|X}(y \mid x)\,\varphi(x)\, f_X(x) \\ &= \sum_{x,y} y\,\varphi(x)\, f_{(X,Y)}(x,y) = \mathbb{E}(Y\varphi(X)).\end{aligned}$$

□

Remarque 2.8. *Au-delà de ses applications immédiates, l'intérêt de ce résultat est que cette propriété peut être prise comme* définition *de l'espérance conditionnelle, cette dernière étant la seule fonction de X satisfaisant cette relation pour toutes les fonctions φ admissibles. Ceci permet de définir cette notion dans des situations beaucoup plus générales qu'ici.*

Exemple 2.13. 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. Pour déterminer $\mathbb{E}(S)$, on peut procéder comme suit :

$$\begin{aligned}
\mathbb{E}(S \mid N)(n) &= \sum_s s\, f_{S|N}(s \mid n) = \sum_s s\, \frac{\mathbb{P}(S = s, N = n)}{\mathbb{P}(N = n)} \\
&= \sum_s s\, \frac{\mathbb{P}(X_1 + \cdots + X_n = s, N = n)}{\mathbb{P}(N = n)} \\
&= \sum_s s\, \frac{\mathbb{P}(X_1 + \cdots + X_n = s)\mathbb{P}(N = n)}{\mathbb{P}(N = n)} \\
&= \sum_s s\, \mathbb{P}(X_1 + \cdots + X_n = s) \\
&= \mathbb{E}(X_1 + \cdots + X_n) = \mu\, n.
\end{aligned}$$

Par conséquent, $\mathbb{E}(S \mid N) = \mu N$, et donc, par le Lemme 2.19,

$$\mathbb{E}(S) = \mathbb{E}(\mathbb{E}(S \mid N)) = \mu\, \mathbb{E}(N).$$

♢

Chapitre 3

Marche aléatoire simple sur $\mathbb{Z}$

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 $\mathbb{Z}$.

Notation : Dans ce chapitre, ainsi que dans la suite du cours, nous utiliserons souvent la notation suivante :

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

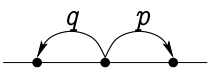
ainsi que ses généralisations naturelles, telles que $\mathbb{P}(A,B,C)$ ou $\mathbb{P}(A \mid B,C)$, etc.

3.1 Description du processus

3.1.1 L'espace de probabilité

On désire modéliser l'évolution (aléatoire) suivante d'une particule sur les sommets du graphe $\mathbb{Z}$:

- ▷ 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 d'un sommet de $\mathbb{Z}$ vers l'un de ses deux voisins ;
- ▷ la probabilité de se déplacer vers le sommet de droite est p ;
- ▷ les sauts sont indépendants.



Afin de n'avoir à considérer que des espaces de probabilité discrets, nous nous restreindrons à la description des N premiers pas de la particule. Soient donc $N \geq 1$ et $a \in \mathbb{Z}$. On considère l'univers

$$\Omega_{N;a} = \left\{ (s_0,\ldots,s_N) \in \mathbb{Z}^{N+1} \,:\, s_0 = a \text{ et } |s_i - s_{i-1}| = 1 \;\forall i \in \{1,\ldots,N\} \right\}.$$

Les éléments de $\Omega_{N;a}$ sont appelés les **trajectoires de longueur** N de la marche aléatoire simple [1] sur $\mathbb{Z}$ partant de a (*cf.* Fig. 3.1).

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.

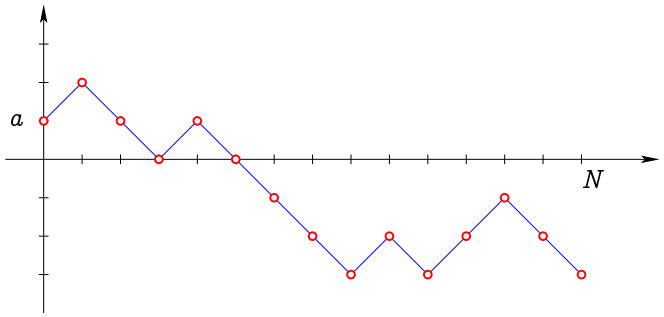


FIGURE 3.1: Une trajectoire de longueur N d'une marche aléatoire simple partant de a.

Soit $p \in [0,1]$ et $q = 1-p$; la mesure de probabilité $\mathbb{P}_{N;p,a}$ sur $\Omega_{N,a}$ désirée est obtenue à partir de la fonction de masse suivante :

$$f_{N;p,a}(s_0,\dots,s_N) = p^{n_+}q^{n_-},$$

où $n_\pm = n_\pm(s_0,\dots,s_N) = \#\{1 \le i \le N \,:\, s_i - s_{i-1} = \pm 1\}$.

L'espace de probabilité $(\Omega_{N;a},\mathbb{P}_{N;p,a})$ décrit (les N premiers pas de) la **marche aléatoire simple sur** $\mathbb{Z}$ partant de a, de paramètre p. La marche aléatoire simple est dite **symétrique** lorsque $p = q = \frac{1}{2}$.

3.1.2 Quelques variables aléatoires utiles

Pour $k \in \mathbb{N}$, $k \le N$, on note $S_k(s_0,\dots,s_N) = s_k$ la variable aléatoire donnant la position de la particule après k pas (on dira également « au temps k »). Les variables aléatoires $X_k = S_k - S_{k-1}$ correspondent alors au déplacement effectué lors du $k^{\text{ème}}$ saut de la particule, et on peut écrire

$$S_n = S_0 + \sum_{i=1}^{n} X_i.$$

Par définition, $\mathbb{P}_{N;p,a}(S_0 = a) = 1$ et les variables aléatoires X_i sont i.i.d. et satisfont

$$\mathbb{P}_{N;p,a}(X_i = 1) = p, \quad \mathbb{P}_{N;p,a}(X_i = -1) = q.$$

En particulier, on peut écrire

$$\mathbb{P}_{N;p,a}(S_0 = s_0, S_1 = s_1, \dots, S_N = s_N) = \delta_{s_0,a} \prod_{i=1}^{N} \mathbb{P}_{N;p,a}(X_i = s_i - s_{i-1}).$$

Exemple 3.1. On considère une suite de lancers d'une pièce de monnaie équilibrée. Après chaque lancer, le joueur gagne 1 CHF si « pile » sort, et perd 1 CHF sinon. S'il commence avec une fortune égale à a CHF, alors l'évolution de sa fortune au cours des N premiers lancers est décrite par une marche aléatoire simple symétrique sur $\mathbb{Z}$ partant de a, sa fortune au temps k étant égale à S_k. ♢

On appelle **cylindres** les événements de la forme

$$[s_0,\dots,s_n] = \{S_0 = s_0, \dots, S_n = s_n\},$$

c'est-à-dire l'ensemble des trajectoires dont les $n \leq N$ premiers pas sont donnés par $s_0, \ldots, s_n$. On note $\mathcal{F}_n$ l'ensemble des événements ne dépendant que des n premiers pas de la trajectoires. Clairement, pour tout événement $A \in \mathcal{F}_n$, on peut trouver $\tilde{A} \subset \Omega_{n;a}$ tel que $A = \bigcup_{(s_0,\ldots,s_n)\in\tilde{A}}[s_0, \ldots, s_n]$. En particulier, on vérifie aisément que

$$\mathbb{P}_{N;a,p}(A) = \mathbb{P}_{n;a,p}(\tilde{A}).$$

Par conséquent, la probabilité d'un événement appartenant à $\mathcal{F}_n$ ne dépend pas de la valeur de N (pourvu que $N \geq n$). Dans la suite, nous nous intéresserons toujours à des événements ne dépendant que d'un nombre fini n de pas de la marche. On peut donc sans ambiguïté omettre le paramètre N de la notation : pour le calcul, on peut prendre n'importe quelle mesure $\mathbb{P}_{N;p,a}$ avec $N \geq n$ (et l'univers $\Omega_{N,a}$ associé). Afin d'alléger la notation, nous omettrons également le paramètre p et écrirons donc simplement $\mathbb{P}_a$.

3.2 Quelques propriétés importantes

La marche aléatoire simple possède les importantes propriétés suivantes.

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) Soit $B \in \mathcal{F}_n$ un événement ne dépendant que des n premiers pas de la marche. Alors, pour tout $s \in \mathbb{Z}$ tel que $\mathbb{P}_a(S_n = s, B) > 0$, on a

$$\mathbb{P}_a((S_n, \ldots, S_N) \in A \mid S_n = s, B) = \mathbb{P}_s((S_0, \ldots, S_{N-n}) \in A),$$

pour tout ensemble de trajectoires $A \subset \Omega_{N-n;s}$.

Il est important que 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 à partir du temps n.

Démonstration. La première affirmation est immédiate, puisque, par définition,

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

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

Pour la seconde propriété, on observe que

$$\begin{aligned}
&\mathbb{P}_a((S_n, S_{n+1}, \ldots, S_N) \in A \mid S_n = s, B) \\
&\qquad = \mathbb{P}_a((s, s + X_{n+1}, \ldots, s + X_{n+1} + \cdots + X_N) \in A \mid S_n = s, B) \\
&\qquad = \mathbb{P}((s, s + X_{n+1}, \ldots, s + X_{n+1} + \cdots + X_N) \in A) \\
&\qquad = \mathbb{P}_s((S_0, S_1, \ldots, S_{N-n}) \in A),
\end{aligned}$$

où l'on a utilisé le fait que $\{S_n = s\} \cap B \in \mathcal{F}_n$ est indépendant de $(X_{n+k})_{k\geq 1}$ pour la deuxième égalité, et le fait que $(s, s + X_{n+1}, \ldots, s + X_{n+1} + \cdots + X_N)$ et $(s, s + X_1, \ldots, s + X_1 + \cdots + X_{N-n})$ ont même loi, puisque les $(X_i)_{i\geq 1}$ sont i.i.d.. □

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

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

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

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

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

$$\mathbb{P}_0(S_{2n} = 0) = \frac{1 + o(1)}{\sqrt{\pi n}}. \tag{3.1}$$

Démonstration. Par homogénéité spatiale, il suffit de considérer le cas $a = 0$. Soit $n_\pm = \#\{1 \leq i \leq n \,:\, X_i = \pm 1\}$. On a manifestement $n_+ + n_- = n$ et $n_+ - n_- = s$. Par conséquent, pour que la marche atteigne s au temps n, il faut qu'elle ait fait $n_+ = \frac{n+s}{2}$ pas vers le haut, et $n_- = \frac{n-s}{2}$ pas vers le bas, ce qui n'est possible que si $n + s$ est pair, et si $|s| \leq n$. Chaque portion de trajectoire contribuant à cet événement a donc probabilité $p^{n_+} q^{n_-}$, et le nombre de telles portions de trajectoires est donné par $\binom{n}{n_+}$. □

3.3 Le premier retour au point de départ

Nous allons à présent étudier le temps mis par la marche pour retourner à son point de départ. Plus précisément, nous allons nous intéresser aux événements

$$\{\tau_0 > n\} \equiv \{S_1 \neq 0, S_2 \neq 0, \ldots, S_n \neq 0\},$$
$$\{\tau_0 = n\} \equiv \{\tau_0 > n-1\} \cap \{S_n = 0\}.$$

Manifestement, ces deux événements appartiennent à $\mathcal{F}_n$.

Remarque 3.1. *Considérons la variable aléatoire $\tau_0 = \min\{n \geq 1 \,:\, S_n = 0\}$ (avec la convention habituelle que $\min \varnothing = +\infty$), définie sur les trajectoires de longueur infinie. On vérifie immédiatement que les deux événements ci-dessus correspondent effectivement à $\tau_0 > n$ et $\tau_0 = n$, respectivement. Cette approche alternative sera justifiée dans la seconde partie du cours, où l'on traitera d'univers généraux.*

Théorème 3.1. *Pour tout $n \geq 1$,*

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

et donc

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

Démonstration. Toutes les portions de trajectoire joignant $(0,0)$ à (n,b) étant 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.

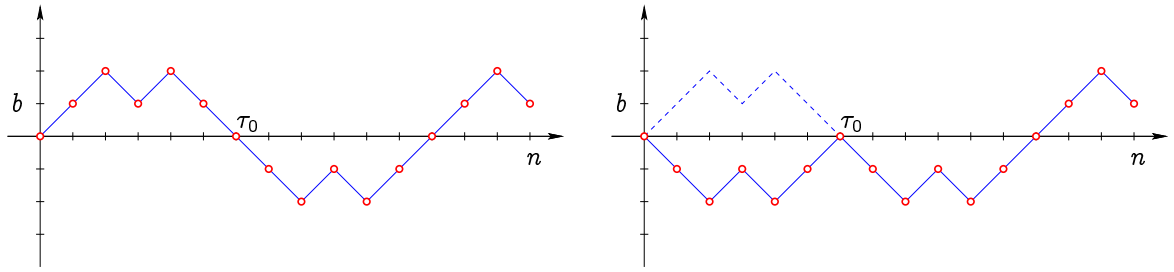
▷ $\mathcal{T}_{\pm}[(1,1),(n,b)]$: ensemble de toutes les portions de trajectoires joignant $(1,1)$ à (n,b) intersectant l'axe des abscisses.

▷ $\mathcal{T}[(1,1),(n,b)]$: ensemble de toutes les portions de trajectoires joignant $(1,1)$ à (n,b).

Manifestement,

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

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 $\#\mathcal{T}_{\pm}[(1,1),(n,b)]$. L'observation essentielle, appelée **principe de réflexion**, est la suivante : l'ensemble $\mathcal{T}_{\pm}[(1,1),(n,b)]$ 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

$$\#\mathcal{T}_{\pm}[(1,1),(n,b)] = \binom{n-1}{\frac{n+b}{2}-1} - \binom{n-1}{\frac{n+b}{2}} = \frac{b}{n}\binom{n}{\frac{n+b}{2}}. \tag{3.2}$$

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. □

On peut facilement déduire du résultat précédent une relation très simple dans le cas symétrique.

Corollaire 3.1. *Dans le cas symétrique,*

$$\mathbb{P}_0(\tau_0 > 2n) = \mathbb{P}_0(S_{2n} = 0).$$

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

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

la quatrième ligne suivant de (3.2). □

Remarque 3.2. *Combiné à (3.1), le résultat précédent montre que*

$$\mathbb{P}_0(\tau_0 > 2n) = \frac{1 + o(1)}{\sqrt{\pi n}}. \tag{3.3}$$

En particulier,

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

Il serait pratique de pouvoir formuler ce dernier résultat sous la forme $\mathbb{P}(\tau_0 = +\infty) = 0$, mais le problème est que l'événement $\{\tau_0 = +\infty\}$ dépend de la trajectoire infinie, et l'univers correspondant n'est pas dénombrable et sort donc du cadre que nous considérons pour le moment. Nous verrons toutefois plus tard qu'il est effectivement possible de construire une mesure de probabilité $\mathbb{P}$ sur les trajectoires infinies satisfaisant, d'une part, $\mathbb{P}(A) = \mathbb{P}_n(A)$ pour tout événement $A \in \mathcal{F}_n$ et, d'autre part, telle que si $A_1 \supset A_2 \supset \cdots$ est une suite décroissante d'événements, alors[2] *$\mathbb{P}(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n)$. Dans notre cas, $\{\tau_0 = +\infty\} = \lim_{n\to\infty}\{\tau > n\}$ et la suite $(\{\tau_0 > n\})_{n\geq 1}$ est manifestement décroissante ; l'écriture ci-dessus est donc justifiée.*

Bien que la marche symétrique retourne presque-sûrement à son point de départ ($\tau_0 < \infty$ avec probabilité 1), cela peut prendre beaucoup de temps : l'espérance de τ_0 est infinie ! En effet, il suit de (3.3) et du Lemme 2.10 que

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

Ceci conduit à des propriétés plutôt 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,

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

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 3.2. *Pour tout $n > 0$ pair,*

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

(La probabilité de cet événement est nulle si n est impair.)

2. Si $(A_k)_{k\geq 1}$ est suite décroissante d'événements, alors $\lim_{k\to\infty} A_k = \bigcap_{k\geq 1} A_k$. Similairement, pour une suite croissante, $\lim_{k\to\infty} A_k = \bigcup_{k\geq 1} A_k$.

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 Théorème 3.1. □

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 Corollaire 3.1.)

3.4 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, $\nu_0(2n) = \max\{0 \le k \le 2n \,:\, S_k = 0\}$.

Théorème 3.2 (Loi de l'arcsinus pour la dernière visite en 0). *On suppose que $p = 1/2$. Pour tout $0 \le k \le n$,*

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

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 :

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

la dernière identité résultant du Corollaire 3.1.

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

$$\mathbb{P}_0(S_{2k} = 0)\mathbb{P}_0(S_{2n-2k} = 0) = \binom{2k}{k}\binom{2n-2k}{n-k}2^{-2n} = \frac{1+o(1)}{\pi\sqrt{k(n-k)}}, \tag{3.4}$$

lorsque k et $n-k$ tendent 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,

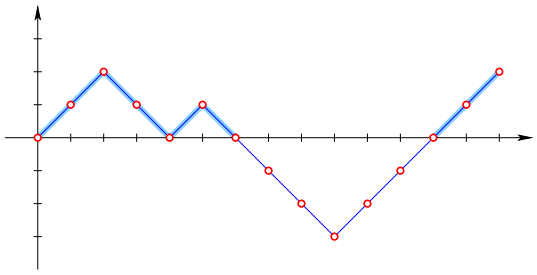
$$\mathbb{P}_0(\nu(10000) \leq 100) \cong \frac{2}{\pi} \arcsin \sqrt{0{,}01} \cong 6{,}4\%\,,$$
$$\mathbb{P}_0(\nu(10000) \geq 9900) \cong \frac{2}{\pi} \arcsin \sqrt{0{,}01} \cong 6{,}4\%\,,$$
$$\mathbb{P}_0(\nu(10000) \leq 1000) \cong \frac{2}{\pi} \arcsin \sqrt{0{,}1} \cong 20{,}5\%\,.$$

3.5 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

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

donnant le temps pendant lequel la marche est positive. Observez que t_{2n}^{+} est nécessairement pair.



Théorème 3.3 (loi de l'arcsinus pour les temps de séjour). *On suppose que $p = 1/2$. Alors,*

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

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 $g_{2k} = \mathbb{P}_0(S_{2k} = 0)$. Nous voulons montrer que

$$f_{2n}(2k) = g_{2k}\, g_{2n-2k}. \tag{3.5}$$

La première observation est que, par le Corollaire 3.1,

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

L'avant-dernière identité suit du fait que, S_{2n-1} étant impair, $S_{2n-1} \geq 0$ implique que $S_{2n} \geq 0$. Ceci établit (3.5) lorsque $k = n$. L'identité pour $k = 0$ suit alors par symétrie.

Soit $k \in \{1, \ldots, n-1\}$. Dans ce cas, lorsque l'événement $t_{2n}^{+} = 2k$ est réalisé, le temps τ_0 du premier retour à l'origine satisfait $\tau_0 = 2r$, avec $1 \leq r < n$. Pour $1 \leq k < \tau_0$, la marche reste toujours strictement positive ou strictement négative, chacune de ces deux possibilités ayant probabilité $1/2$. 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'en τ_0, et la seconde celle des trajectoires négatives jusqu'en τ_0.

Pour conclure la preuve, on fait une récurrence. On a déjà vérifié la validité de (3.5) pour tous les $0 \leq k \leq n$ lorsque $n = 1$. Supposons donc (3.5) vérifiée pour tous les $0 \leq k \leq n$ 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 (3.5). La dernière identité suit de l'observation que, pour tout $\ell \geq 1$,

$$\begin{aligned}\mathbb{P}_0(S_{2\ell} = 0) &= \sum_{r=1}^{\ell} \mathbb{P}_0(S_{2\ell} = 0 \mid \tau_0 = 2r)\mathbb{P}_0(\tau_0 = 2r) \\ &= \sum_{r=1}^{\ell} \mathbb{P}_0(S_{2\ell-2r} = 0)\mathbb{P}_0(\tau_0 = 2r),\end{aligned}$$

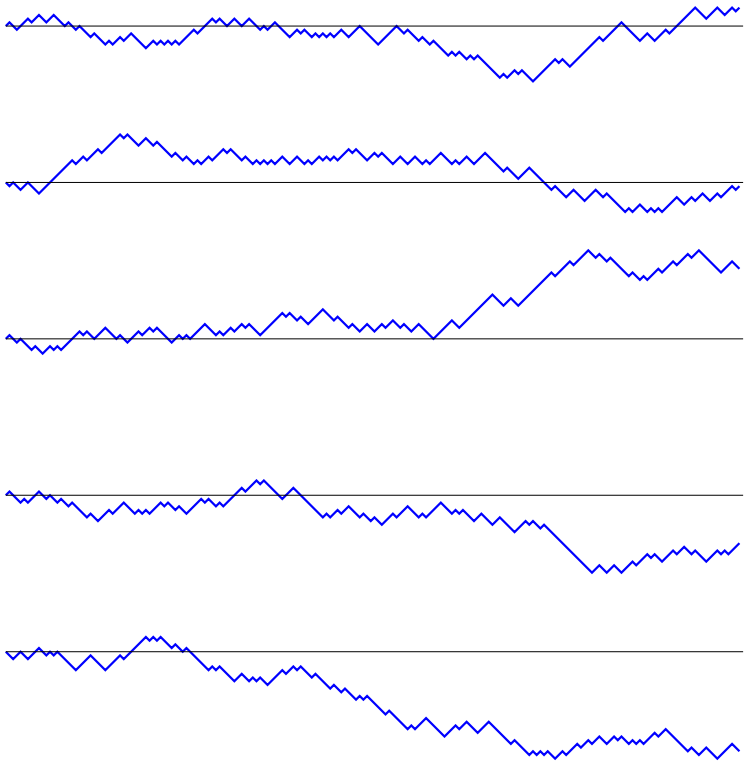
c'est-à-dire $g_{2\ell} = \sum_{r=1}^{\ell} g_{2\ell-2r} h_{2r}$.

La seconde affirmation a déjà été démontrée dans la preuve du Théorème 3.2. □

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 grand temps n, 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 : 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 le plus chanceux pendant la partie domine l'autre joueur *pendant plus de* 364 *jours et* 10 *heures* !

La figure suivante 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) !



Chapitre 4

Fonctions génératrices

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

$$G_a(s) = \sum_{i=0}^{\infty} a_i s^i \qquad \text{pour les } s \in \mathbb{C} \text{ tels que la série converge.}$$

Rappelons quelques propriétés de base de ce type de fonctions.

Convergence. Il existe un **rayon de convergence** $0 \leq R \leq \infty$ 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 $\{s : |s| \leq R'\}$, 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 $0 < R' \leq R$ tel que $G_a(s) = G_b(s)$ pour tout $|s| < R'$, alors $a_n = b_n$ pour tout n. De plus,

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

Continuité. (Théorème d'Abel) Si $a_i \geq 0$ pour tout i, et $G_a(s)$ est finie pour $|s| < 1$, alors $\lim_{s \uparrow 1} G_a(s) = \sum_{i=0}^{\infty} a_i$, que cette somme soit finie ou égale à $+\infty$. (Ce résultat est particulièrement 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 $(f_X(k))_{k=0}^{\infty}$; 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*

$$G_X(s) = \mathbb{E}(s^X) = \sum_{k=0}^{\infty} s^k f_X(k).$$

Remarque 4.1. *Puisque $G_X(1) = \mathbb{E}(1) = 1$, il suit que le rayon de convergence R de G_X est toujours supérieur ou égal à 1.*

Exemple 4.1. 1. *Variable aléatoire constante.* Si $\mathbb{P}(X = c) = 1$, alors $G_X(s) = s^c$.

2. *Loi de Bernoulli.* Si $\mathbb{P}(X = 1) = p$ et $\mathbb{P}(X = 0) = 1 - p$, on a

$$G_X(s) = (1 - p) + ps.$$

3. *Loi binomiale.* Pour une loi binomiale de paramètres n et p, la formule du binôme implique que
$$G_X(s) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} s^k = ((1-p)+ps)^n.$$
4. *Loi de Poisson.* Pour X suivant une loi de Poisson de paramètre λ, on obtient
$$G_X(s) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} s^k = e^{\lambda(s-1)}.$$
5. *Loi géométrique.* Pour X suivant une loi géométrique de paramètre p, on a
$$\sum_{k=1}^{\infty} p(1-p)^{k-1} s^k = \frac{ps}{1-(1-p)s}.$$

♢

Le théorème d'Abel fournit une technique efficace pour calculer les moments de X ; par exemple ($G_X^{(k)}(1)$ devant s'interpréter comme $\lim_{s\uparrow 1} G_X^{(k)}(s)$ lorsque $R = 1$)

$$\begin{aligned}
G_X'(s) &= \sum_{k=0}^{\infty} k s^{k-1} f_X(k) && \Longrightarrow G_X'(1) = \mathbb{E}(X),\\
G_X''(s) &= \sum_{k=0}^{\infty} k(k-1) s^{k-2} f_X(k) && \Longrightarrow G_X''(1) = \mathbb{E}(X(X-1)),\\
G_X^{(\ell)}(s) &= \sum_{k=0}^{\infty} k\cdots(k-\ell+1) s^{k-\ell} f_X(k) && \Longrightarrow G_X^{(\ell)}(1) = \mathbb{E}(X\cdots(X-\ell+1)).
\end{aligned}$$

On a donc en particulier le résultat suivant.

Proposition 4.1. *Si $G_X(s)$ est la fonction génératrice de X, alors*
$$\mathbb{E}(X) = G_X'(1), \qquad \mathrm{Var}(X) = G_X''(1) + G_X'(1) - G_X'(1)^2,$$
les expressions dans les membres de droite devant être compris comme des limites $s \uparrow 1$ lorsque le rayon de convergence de G_X est égal à 1.

Exemple 4.2. *Espérance et variance de la loi hypergéométrique.* La formule du binôme montre que la fonction génératrice d'une variable hypergéométrique X de paramètres N, n et b,
$$G_X(s) = \sum_{k=(n-r)\vee 0}^{b\wedge n} s^k \binom{b}{k}\binom{N-b}{n-k} \Big/ \binom{N}{n},$$
est précisément le coefficient de x^n du polynôme
$$Q(x,s) = (1+sx)^b (1+x)^{N-b} \Big/ \binom{N}{n}.$$
Il suit que la moyenne de X coïncide avec le coefficient de x^n de
$$\frac{\partial Q}{\partial s}(x,1) = xb(1+x)^{N-1} \Big/ \binom{N}{n},$$
et est donc donnée par $G_X'(1) = bn/N$. Similairement, on trouve que la variance de X est égale à $nb(N-b)(N-n)/(N^3-N^2)$. ♢

Remarque 4.2. *En général, si l'on désire calculer les moments d'une variable aléatoire X, il se révèle avantageux de travailler avec la* **fonction génératrice des moments** *de X, qui est définie par*

$$M_X(t) = G_X(e^t),$$

pourvu que $e^t < R$, le rayon de convergence de G_X. En effet, on a alors

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

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

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

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

Proposition 4.2. *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes à valeurs dans $\mathbb{N}$. Alors la fonction génératrice de $S_n = X_1 + \cdots + X_n$ est donnée par*

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

Démonstration. En utilisant le Lemme 2.18, on a

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

□

Exemple 4.3. *Loi de Pascal.* On peut à présent calculer aisément la fonction génératrice d'une variable de Pascal X de paramètres r et p. En effet, celle-ci peut se décomposer en $X + r = X_1 + \cdots + X_r$, où les X_i sont des variables géométriques de paramètre p indépendantes, et on a donc

$$G_X(s) = s^{-r}G_{X+r}(s) = s^{-r}\left(G_{X_1}(s)\right)^r = \Big(\frac{p}{1-(1-p)s}\Big)^r.$$

♢

Exemple 4.4. Soient X et Y deux variables aléatoires indépendantes, suivant des lois binomiales de paramètres m et p, et n et p, respectivement. Alors

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

et donc $X + Y$ suit une loi binomiale de paramètres $m + n$ et p.

Similairement, si X et Y sont deux variables aléatoires indépendantes suivant des lois de Poisson de paramètre λ et μ, respectivement, alors $X + Y$ suit une loi de Poisson de paramètre $\lambda + \mu$:

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

De même, on vérifie facilement que si X et Y sont des variables aléatoires indépendantes suivant des lois de Pascal de paramètres r_1 et p, et r_2 et p, alors $X + Y$ suit une loi de Pascal de paramètres $r_1 + r_2$ et p. ♢

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

Proposition 4.3. *Soient $X_1, X_2, \dots$ 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 des X_i et dont la fonction génératrice est G_N. Alors la fonction génératrice de $S = X_1 + \cdots + X_N$ est donnée par*

$$G_S = G_N \circ G_X .$$

Démonstration. En utilisant le Lemme 2.19,

$$\begin{aligned} G_S(s) = \mathbb{E}(s^S) = \mathbb{E}(\mathbb{E}(s^S \mid N)) &= \sum_n \mathbb{E}(s^S \mid N)(n)\,\mathbb{P}(N=n) \\ &= \sum_n \mathbb{E}(s^{X_1+\cdots+X_n})\,\mathbb{P}(N=n) = \sum_n \mathbb{E}(s^{X_1})\cdots\mathbb{E}(s^{X_n})\,\mathbb{P}(N=n) \\ &= \sum_n (G_X(s))^n\,\mathbb{P}(N=n) = G_N(G_X(s)). \end{aligned}$$

□

Exemple 4.5. En prenant la dérivée de G_S en 1, on retrouve immédiatement le résultat de l'Exemple 2.13. ♢

Exemple 4.6. 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 distribution de K ? Comme on l'a vu,

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

Par conséquent,

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

ce qui est la fonction génératrice d'une variable de Poisson de paramètre λp. ♢

Le théorème de continuité suivant montre que les fonctions génératrices permettent l'étude de la convergence de certaines suites de variables aléatoires.

Théorème 4.1. *Soient $(X_n)_{n\geq 1}$ une suite de variables aléatoires à valeurs dans $\mathbb{N}$. Les deux propositions suivantes sont équivalentes :*

1. *$p_k = \lim_{n\to\infty} \mathbb{P}(X_n = k)$ existe pour tout $k \in \mathbb{N}$;*
2. *$G(s) = \lim_{n\to\infty} G_{X_n}(s)$ existe pour tout $0 < s < 1$.*

De plus, on a alors $G(s) = \sum_{k\geq 0} s^k p_k$. En particulier, lorsque $\sum_{k\geq 0} p_k = 1$, il existe une variable aléatoire X à valeurs dans $\mathbb{N}$ telle que

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

Remarque 4.3. *Observez que l'on peut très bien avoir l'existence de toutes les limites $\lim_{n\to\infty} \mathbb{P}(X_n = k)$ sans qu'il existe une variable aléatoire limite : il suffit de considérer, par exemple, des variables aléatoires $(X_n)_{n\geq 1}$ telles que $\mathbb{P}(X_n = n) = 1$. Dans des situations de ce type, une partie de la probabilité est « perdue à l'infini ».*

Démonstration. *1. $\Longrightarrow$ 2.* Supposons tout d'abord que $p_k = \lim_{n\to\infty} \mathbb{P}(X_n = k)$ existe pour tout $k \in \mathbb{N}$, et posons $G(s) = \sum_{k\geq 0} s^k p_k$. Soit $0 < s < 1$. Comme $|\mathbb{P}(X_n = k) - p_k| \leq 1$, on a

$$|G_{X_n}(s) - G(s)| \leq \sum_{k=0}^{r} |\mathbb{P}(X_n = k) - p_k| + \sum_{k=r+1}^{\infty} s^k \leq \sum_{k=0}^{r} |\mathbb{P}(X_n = k) - p_k| + \frac{s^r}{1-s}.$$

Les deux termes du membre de droite pouvant être rendus arbitrairement petits en prenant r suffisamment grand, puis n suffisamment grand, la conclusion suit.

2. $\Longrightarrow$ 1. Supposons à présent que $G(s) = \lim_{n\to\infty} G_{X_n}(s)$ existe pour tout $0 < s < 1$. D'une part, $G(s)$ étant nécessairement croissante en s, la limite $G(0) = \lim_{s\downarrow 0} G(s)$ existe. D'autre part, on a

$$\mathbb{P}(X_n = 0) = G_{X_n}(0) \leq G_{X_n}(s) \leq \mathbb{P}(X_n = 0) + \sum_{k\geq 1} s^k = \mathbb{P}(X_n = 0) + \frac{s}{1-s}.$$

Ceci implique que

$$G(s) - \frac{s}{1-s} \leq \liminf_{n\to\infty} \mathbb{P}(X_n = 0) \leq \limsup_{n\to\infty} \mathbb{P}(X_n = 0) \leq G(s),$$

et donc, en laissant $s \downarrow 0$,

$$\lim_{n\to\infty} \mathbb{P}(X_n = 0) = G(0).$$

On procède à présent par récurrence. Notons $p_k^n = \mathbb{P}(X_n = k)$, et supposons que $p_k = \lim_{n\to\infty} p_k^n$ existe pour tout $k < r$. On peut alors écrire, pour tout $0 < s < 1$,

$$\lim_{n\to\infty} \frac{G_{X_n}(s) - p_0^n - p_1^n s - \cdots - p_{r-1}^n s^{r-1}}{s^r} = \frac{G(s) - p_0 - p_1 s - \cdots - p_{r-1} s^{r-1}}{s^r} \equiv H_r(s).$$

La fraction dans le membre de gauche peut s'écrire $\sum_{k\geq 0} p_{k+r}^n s^k$, qui est à nouveau une série à termes positifs. On peut donc répéter le raisonnement précédent afin de conclure à l'existence de la limite $H_r(0) = \lim_{s\downarrow 0} H_r(s)$. En procédant comme ci-dessus, on en déduit alors que $\lim_{n\to\infty} p_r^n = H_r(0)$.

Ceci montre l'existence de $p_k = \lim_{n\to\infty} p_k^n$, pour tout $k \geq 0$. L'identification $G(s) = \sum_{k\geq 0} p_k s^k$ suit alors de l'implication *1. $\Longrightarrow$ 2.*. □

Exemple 4.7. Soit $(X_n)_{n\geq 0}$ une suite de variables aléatoires de loi binom(n, p_n), avec $\lim_{n\to\infty} np_n = \lambda > 0$. On a

$$\lim_{n\to\infty} G_{X_n}(s) = \lim_{n\to\infty} (1 + (s-1)p_n)^n = e^{(s-1)\lambda}.$$

Cette dernière expression étant la fonction génératrice associée à la loi poisson(λ), on retrouve la loi des petits nombres. ♢

Remarque 4.4. *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 défectives prenant valeurs dans $\mathbb{N} \cup \{+\infty\}$. Pour une telle variable aléatoire X, on voit que $G_X(s) = \mathbb{E}(s^X)$ converge tant que $|s| < 1$, et que*

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

Il n'est bien sûr plus possible d'obtenir les moments de X à partir de G_X : ceux-ci sont tous infinis !

4.2 Application aux processus de branchement

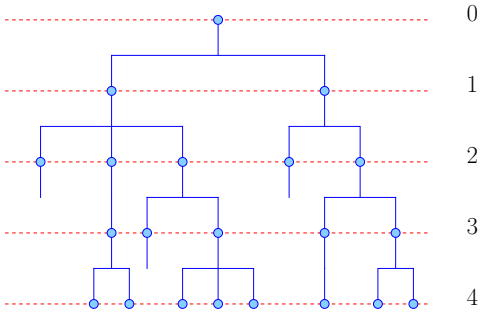
Dans cette section, nous allons illustrer la puissance des fonctions génératrices dans l'étude d'une classe intéressante de processus stochastiques : les **processus de branchement**.

À l'époque victorienne, certaines personnes ont craint la disparition des noms des familles aristocratiques. Sir Francis Galton [1] posa originellement la question de déterminer la probabilité d'un tel événement dans le *Educational Times* de 1873, et le Révérend Henry William Watson [2] 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 $n^{\text{ème}}$ génération. Chaque membre de la $n^{\text{ème}}$ génération 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.



On s'intéresse à la suite aléatoire $Z_0, Z_1, Z_2, \ldots$ des tailles des générations successives. Ce processus peut être défini de la façon suivante : on considère une collection $(X_k^n)_{n\geq 0, k\geq 1}$ de variables aléatoires i.i.d. à valeurs dans $\mathbb{N}$; X_k^n représente le nombre de fils du $k^{\text{ème}}$ individu de la $n^{\text{ème}}$ génération (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 $Z_0 = 1$ et, pour $n \geq 1$,

$$Z_n = X_1^{n-1} + X_2^{n-1} + \cdots + X_{Z_{n-1}}^{n-1}, \tag{4.1}$$

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

Notons que, comme dans le chapitre 3, l'étude du processus au temps n ne requiert qu'un espace de probabilité discret (les trajectoires du processus entre les temps 0 et n).

Théorème 4.2. *Pour tout $n \geq 1$,*

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

1. 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.
2. Henry William Watson (1827 – 1903), mathématicien britanique.

Démonstration. Soit $n \geq 1$. Il suit immédiatement de (4.1) et de la Proposition 4.3 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 :

$$G_n = G_{n-1} \circ G = (G_{n-2} \circ G) \circ G = \cdots = \underbrace{G \circ G \circ \cdots \circ G}_{n \text{ fois}}.$$

□

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 4.1. *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 Théorème 4.2, $G_n = G^{\circ n} = G \circ G^{\circ(n-1)} = G \circ G_{n-1}$. Par conséquent, il suit de la Proposition 4.1 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, $\mathbb{E}(Z_n) = \mu^n$. Similairement,

$$G_n''(1) = G''(1)\,(G_{n-1}'(1))^2 + G'(1)\, G_{n-1}''(1).$$

Par conséquent, la Proposition 4.1 implique que

$$\mathrm{Var}(Z_n) = \sigma^2\, \mu^{2n-2} + \mu\, \mathrm{Var}(Z_{n-1}),$$

et la conclusion suit. □

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 ? Cette question peut être reformulée sous la forme 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.)

Remarque 4.5. *Comme pour le cas du retour à l'origine de la marche aléatoire au chapitre 3, il serait plus commode de pouvoir considérer directement des trajectoires « infinies » du processus. Dans ce cas, l'événement « la population s'éteint après un temps fini » pourrait s'écrire (puisque $Z_n \in \mathbb{N}$ pour tout n)*

$$\{\textit{extinction}\} = \{\lim_{n\to\infty} Z_n = 0\} = \{\exists n_0 \,:\, Z_n = 0,\ \forall n \geq n_0\} = \bigcup_{n\geq 1} \{Z_n = 0\}.$$

Comme on l'a déjà mentionné au chapitre 3, on discutera plus tard comment construire une mesure de probabilité $\mathbb{P}$ sur les trajectoires infinies. De plus, cette mesure limite satisfera $\mathbb{P}(\lim_{n\to\infty} Z_n = 0) = \lim_{n\to\infty} \mathbb{P}(Z_n = 0)$, car la suite d'événements $\{Z_n = 0\}_{n\geq 1}$ est croissante.

Le théorème suivant montre que le destin de la population est étroitement lié à la taille moyenne des familles.

Théorème 4.3. *Soit $\mu = \mathbb{E}(Z_1)$, la taille moyenne d'une famille. La probabilité d'extinction*

$$\eta = \lim_{n\to\infty} \mathbb{P}(Z_n = 0)$$

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 $\eta_n = \mathbb{P}(Z_n = 0) = G_n(0)$. L'existence de la limite $\eta = \lim_{n\to\infty} \eta_n$ est immédiate puisque $(\eta_n)_{n\geq 1}$ est une suite croissante et bornée (par 1). On a

$$\eta_n = G_n(0) = G(G_{n-1}(0)) = G(\eta_{n-1}).$$

Par continuité de G, on peut passer à la limite ($n \to \infty$), ce qui montre que la probabilité d'extinction satisfait

$$\eta = \lim_{n\to\infty} \eta_n = \lim_{n\to\infty} G(\eta_{n-1}) = G(\lim_{n\to\infty} \eta_{n-1}) = G(\eta).$$

Vérifions à présent que si a est une racine positive de cette équation, alors $\eta \leq a$. G étant croissante sur $[0, 1]$, on a

$$\eta_1 = G(0) \leq G(a) = a.$$

Similairement,

$$\eta_2 = G(\eta_1) \leq G(a) = a.$$

Il suit, par induction, que $\eta_n \leq a$, pour tout n, et donc que $\eta \leq a$. Par conséquent, η est bien la plus petite racine positive de l'équation $s = G(s)$.

Pour démontrer la seconde affirmation, on utilise le fait que G est convexe sur $[0, 1]$; ceci est vrai, car

$$G''(s) = \mathbb{E}(Z_1(Z_1 - 1)s^{Z_1-2}) = \sum_{k\geq 2} k(k-1)s^{k-2}\mathbb{P}(Z_1 = k) \geq 0, \qquad \text{pour } s \in [0, 1].$$

G est donc convexe (en fait, strictement convexe si $\mathbb{P}(Z_1 \geq 2) > 0$) 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 $G(0) \geq 0$). Un coup d'œil à la Figure 4.1 (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 $G(0) \geq 0$, et donc $\eta < 1$. Finalement, dans le cas $\mu = 1$, le deux courbes sont tangentes en 1, et donc $\eta = 1$ 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 la la marche aléatoire simple sur $\mathbb{Z}$. On a déjà vu, dans le 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.

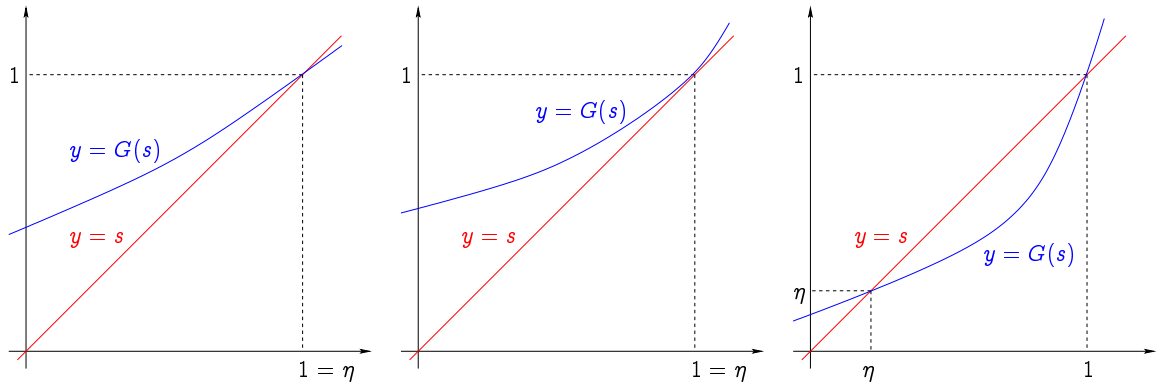


FIGURE 4.1: Solutions de l'équation $G(s) = s$. Gauche : $\mu < 1$. Milieu : $\mu = 1$ et $\mathrm{Var}(Z_1) > 0$. Droite : $\mu > 1$.

On note $g_n = \mathbb{P}_0(S_n = 0)$ et $h_n = \mathbb{P}_0(\tau_0 = n)$. Les fonctions génératrices correspondantes sont

$$\mathbb{G}(s) = \sum_{n=0}^{\infty} g_n s^n, \qquad \mathbb{H}(s) = \sum_{n=1}^{\infty} h_n s^n.$$

Il convient de remarquer que τ_0 peut être défective, auquel cas $\mathbb{H}(1) = \mathbb{P}_0(\tau_0 < \infty) < 1$.

Lemme 4.2. *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, on a, pour $n \geq 1$,

$$g_{2n} = \mathbb{P}_0(S_{2n} = 0) = \sum_{k=1}^{n} \mathbb{P}_0(\tau_0 = 2k)\mathbb{P}_0(S_{2n} = 0 \,|\, \tau_0 = 2k)$$
$$= \sum_{k=1}^{n} \mathbb{P}_0(\tau_0 = 2k)\mathbb{P}_0(S_{2n-2k} = 0) = \sum_{k=1}^{n} h_{2k} g_{2n-2k}.$$

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 $\mathbb{G}(s) = \sum_{n\geq 0} \binom{2n}{n}(pqs^2)^n$. 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 $n \in \mathbb{N}$,

$$\binom{a}{n} = \frac{a(a-1)(a-2)\cdots(a-n+1)}{n!}.$$

On a vu (Lemme 1.4) que, pour tout $a \in \mathbb{R}$ et tout x tel que $|x| < 1$,

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

Par conséquent, on a que, pour $|4pqs^2| < 1$ (c'est-à-dire $|s| < 1$, puisque $pq \leq \frac{1}{4}$),

$$\mathbb{G}(s) = \sum_{n\geq 0} \binom{-\frac{1}{2}}{n} (-4pqs^2)^n = (1-4pqs^2)^{-1/2}.$$

3. suit immédiatement de *1* et *2*. □

Corollaire 4.1. *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,

$$\mathbb{E}_0(\tau_0) = \sum_{n=1}^{\infty} nh(n) = \mathbb{H}'(1) = \infty.$$

Démonstration. La première affirmation suit après avoir pris la limite $s \uparrow 1$ dans l'expression pour $\mathbb{H}(s)$ donnée dans le Lemme 4.2 (observez que $1 - 4pq = (p-q)^2$).

Lorsque $p = \frac{1}{2}$, 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 4.2. *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 de temps de retour est infinie, et* ***récurrente-positive*** *si cette espérance est finie.*

Le corollaire précédent montre que la marche aléatoire simple unidimensionnelle est récurrente-nulle si $p = \frac{1}{2}$ et transiente dans les autres cas.

4.4 Fonction génératrice conjointe

Tout comme la loi d'une variable aléatoire à valeurs dans $\mathbb{N}$ peut être encodée par sa fonction génératrice, la loi conjointe d'une famille de variables aléatoires à valeurs dans $\mathbb{N}$ peut être encodée par leur fonction génératrice conjointe.

Définition 4.3. *La* ***fonction génératrice conjointe*** *du vecteur aléatoire* $\mathbf{X} = (X_1, \dots, X_n)$ *prenant valeurs dans* $\mathbb{N}^n$ *est définie par*

$$G_{(X_1,\dots,X_n)}(s_1, \dots, s_n) = \mathbb{E}(s_1^{X_1} \cdots s_n^{X_n}).$$

La fonction génératrice conjointe peut être utilisée pour caractériser l'indépendance de variables aléatoires.

Proposition 4.4. *$X_1, \dots, X_n$, à valeurs dans $\mathbb{N}$, sont indépendantes si et seulement si*

$$G_{(X_1,\dots,X_n)}(s_1, \dots, s_n) = G_{X_1}(s_1) \cdots G_{X_n}(s_n),$$

pour tout $s_1, \dots, s_n$.

Démonstration. Les X_i étant indépendantes, c'est aussi le cas des $s_i^{X_i}$. Par conséquent,

$$\begin{aligned} G_{(X_1,\dots,X_n)}(s_1, \dots, s_n) &= \mathbb{E}(s_1^{X_1} \cdots s_n^{X_n}) = \mathbb{E}(s_1^{X_1}) \cdots \mathbb{E}(s_n^{X_n}) \\ &= G_{X_1}(s_1) \cdots G_{X_n}(s_n). \end{aligned}$$

Pour démontrer l'autre direction, on procède comme suit :

$$\begin{aligned} G_{(X_1,\dots,X_n)}(s_1, \dots, s_n) - G_{X_1}(s_1) \cdots G_{X_n}(s_n) = \\ \sum_{x_1,\dots,x_n} s_1^{x_1} \cdots s_n^{x_n} \left(\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) - \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n)\right). \end{aligned}$$

Comme, par hypothèse, cette fonction est identiquement nulle sur son domaine de définition, on en conclut que

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) - \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n) = 0,$$

pour tout $x_1, \dots, x_n$ (observez qu'on peut obtenir les coefficients d'une telle série entière en la dérivant par rapport à ses variables en $s_1 = \cdots = s_n = 0$, et ici toutes les dérivées sont nulles). Les X_i sont donc indépendants. □

Deuxième partie

Espaces de probabilité généraux

Résumé

Dans cette partie du cours, nous verrons comment traiter les univers infinis non nécessairement dénombrables. Pour ce faire, le formalisme général de la théorie des probabilités, basé sur les axiomes de Kolmogorov, sera introduit.

Chapitre 5

Approche axiomatique

5.1 Construction d'espaces de probabilité

Nous allons à présent discuter des espaces de probabilité associés à des univers Ω généraux (c'est-à-dire, potentiellement non dénombrables). Cette situation est substantiellement plus subtile que celle considérée dans la première partie.

Manifestement, on ne peut en général plus considérer l'approche utilisée dans la première partie, consistant à construire la mesure de probabilité $\mathbb{P}$ à partir de la probabilité des événements élémentaires. On va donc chercher à définir $\mathbb{P}$ directement au niveau des événements généraux. Quelles sont les propriétés qu'il est naturel d'exiger d'une telle mesure ? Les propriétés suivantes semblent être le minimum :

- $\mathbb{P}(A) \in [0,1]$ pour tout événement A ;
- normalisation : $\mathbb{P}(\Omega) = 1$;
- σ-additivité : $\mathbb{P}(\bigcup_{k\geq 1} A_k) = \sum_{k\geq 1} \mathbb{P}(A_k)$, pour toute collection $(A_k)_{k\geq 1}$ d'événements 2 à 2 disjoints.

En effet, si ces propriétés sont satisfaites, alors on retrouve les autres propriétés utilisées abondamment dans la première partie, le Corollaire 1.1 restant valide. Une justification supplémentaire de l'importance de l'hypothèse de σ-additivité est donnée par le lemme suivant, qui montre qu'elle implique une forme de continuité de $\mathbb{P}$, dont on a vu à plusieurs reprises dans la première partie de ce cours à quel point elle est désirable.

Lemme 5.1. *Supposons l'hypothèse de σ-additivité satisfaite. Alors, pour toute suite croissante d'événements $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$, on a*

$$\mathbb{P}(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n),$$

où $\lim_{n\to\infty} A_n = \bigcup_{n\geq 1} A_n$. Similairement, on a, pour toute suite décroissante $B_1 \supseteq B_2 \supseteq B_3 \supseteq \cdots$,

$$\mathbb{P}(\lim_{n\to\infty} B_n) = \lim_{n\to\infty} \mathbb{P}(B_n),$$

où $\lim_{n\to\infty} B_n = \bigcap_{n\geq 1} B_n$.

Démonstration. On peut écrire $\lim_{n\to\infty} A_n = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \cdots$ comme union

d'une famille d'événements deux-à-deux disjoints. Par conséquent,

$$\begin{aligned}\mathbb{P}(\lim_{n\to\infty} A_n) &= \mathbb{P}(A_1) + \sum_{i=1}^{\infty} \mathbb{P}(A_{i+1} \setminus A_i) \\ &= \mathbb{P}(A_1) + \lim_{n\to\infty} \sum_{i=1}^{n} (\mathbb{P}(A_{i+1}) - \mathbb{P}(A_i)) \\ &= \mathbb{P}(A_1) + \lim_{n\to\infty} (\mathbb{P}(A_{n+1}) - \mathbb{P}(A_1)) \\ &= \lim_{n\to\infty} \mathbb{P}(A_n).\end{aligned}$$

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

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

□

Remarque 5.1. *Soit $(A_k)_{k\geq 1}$ une famille d'événements 2 à 2 disjoints. La suite d'événements $B_k = \bigcup_{i=1}^{k} A_i$ est croissante et $\lim_{k\to\infty} B_k = \bigcup_{i\geq 1} A_i$. Par conséquent, la propriété de continuité ci-dessus et l'additivité finie de $\mathbb{P}$ implique sa σ-additivité :*

$$\mathbb{P}(\bigcup_{i\geq 1} A_i) = \mathbb{P}(\lim_{k\to\infty} B_k) = \lim_{k\to\infty} \mathbb{P}(B_k) = \lim_{k\to\infty} \mathbb{P}(\bigcup_{i=1}^{k} A_i) = \lim_{k\to\infty} \sum_{i=1}^{k} \mathbb{P}(A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

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

La question qui se pose à présent est de déterminer s'il est toujours possible de construire une mesure de probabilité $\mathbb{P} : \mathscr{P}(\Omega) \to \mathbb{R}$ possédant ces trois propriétés. Le lemme suivant montre que cet espoir est vain.

Lemme 5.2. *Soit $\Omega = \{0,1\}^{\mathbb{N}}$ l'univers correspondant à une suite infinie de lancers d'une pièce de monnaie équilibrée. Il n'existe pas d'application $\mathbb{P} : \mathscr{P}(\Omega) \to [0,1]$ possédant les propriétés suivantes :*

- *$\mathbb{P}(\Omega) = 1$;*
- *$\mathbb{P}(\bigcup_{k\geq 1} A_k) = \sum_{k\geq 1} \mathbb{P}(A_k)$, pour toute collection $(A_k)_{k\geq 1}$ d'événements 2 à 2 disjoints ;*
- *Pour tout $A \subset \Omega$ et $n \geq 1$, $\mathbb{P}(T_n A) = \mathbb{P}(A)$, où*

$$T_n : \omega = (\omega_1, \omega_2, \ldots) \mapsto (\omega_1, \ldots, \omega_{n-1}, 1 - \omega_n, \omega_{n+1}, \ldots)$$

est l'application inversant le résultat du $n^{ème}$ lancer.

Remarque 5.2. *La troisième condition exprime à la fois l'indépendance des lancers successifs et le fait que la pièce est équilibrée.*

Démonstration. Notons $\sim$ la relation d'équivalence sur Ω donnée par $\omega \sim \omega'$ si et seulement si il existe $N(\omega, \omega')$ tel que $\omega_k = \omega'_k$ pour tout $k \geq N$. L'axiome du choix permet alors de construire un sous-ensemble A de Ω contenant exactement un représentant de chacune des classes d'équivalence.

Soit $\mathcal{S} = \{S \subset \mathbb{N}^* : |S| < \infty\}$ l'ensemble de toutes les collections finies d'entiers naturels. Comme $\mathcal{S} = \bigcup_{m \geq 1} \{S \subset \mathbb{N}^* : \max S = m\}$ est une union dénombrable d'ensembles finis, $\mathcal{S}$ est dénombrable.

Étant donné $S = \{n_1, \ldots, n_k\} \in \mathcal{S}$, on note $T_S = \prod_{n \in S} T_n = T_{n_1} \circ \cdots \circ T_{n_k}$ l'application inversant le résultat des lancers spécifiés par S.

On a alors :

- ▷ $\Omega = \bigcup_{S \in \mathcal{S}} T_S(A)$. En effet, pour chaque $\omega \in \Omega$, il existe un $\omega' \in A$ (le représentant de sa classe d'équivalence) tel que $\omega \sim \omega'$, et donc un ensemble $S \in \mathcal{S}$ tel que $\omega = T_S(\omega') \in T_S(A)$.
- ▷ Les ensembles $(T_S(A))_{S \in \mathcal{S}}$ sont 2 à 2 disjoints. En effet, supposons que $T_{S_1}(A) \cap T_{S_2}(A) \neq \varnothing$. Il existe alors $\omega_1, \omega_2 \in A$ tels que $T_{S_1}(\omega_1) = T_{S_2}(\omega_2)$. On a alors $\omega_1 \sim T_{S_1}(\omega_1) = T_{S_2}(\omega_2) \sim \omega_2$, et donc $\omega_1 \sim \omega_2$, ce qui implique que $\omega_1 = \omega_2$ et donc $S_1 = S_2$.

On en déduit que

$$1 = \mathbb{P}(\Omega) = \mathbb{P}\Big(\bigcup_{S \in \mathcal{S}} T_S(A)\Big) = \sum_{S \in \mathcal{S}} \mathbb{P}(T_S(A)) = \sum_{S \in \mathcal{S}} \mathbb{P}(A),$$

ce qui est impossible. □

Remarque 5.3. *La preuve précédente repose sur l'axiome du choix (non-dénombrable). On peut montrer que cela est nécessaire.*

Au vu du résultat précédent, il nous faut faire des concessions. Il n'est pas souhaitable de renoncer aux propriétés énoncées ci-dessus, car cela appauvrirait substantiellement la théorie. Une autre solution est de renoncer à chercher à définir $\mathbb{P}$ sur *tous* les sous-ensembles de Ω. En effet, l'applicabilité de la théorie ne sera pas diminuée si les sous-ensembles auxquels on n'associe pas de probabilité sont suffisamment pathologiques. Le fait mentionné précédemment que la construction de sous-ensembles problématiques, comme celui donné dans la preuve ci-dessus, requiert l'axiome du choix montre qu'aucun de ceux-ci ne peut être décrit explicitement. En particulier, leur exclusion n'a aucun impact dans la pratique.

5.1.1 La tribu des événements

La discussion précédente nous conduit donc à restreindre la notion d'événements à une classe $\mathcal{F} \subseteq \mathscr{P}(\Omega)$. Afin de pouvoir travailler, $\mathcal{F}$ doit être stable sous les manipulations habituelles. Ceci conduit à la définition suivante.

Définition 5.1. *Un ensemble $\mathcal{F} \subset \mathscr{P}(\Omega)$ est une* ***tribu*** *sur Ω si elle possède les propriétés suivantes :*

- ▷ $\Omega \in \mathcal{F}$;
- ▷ $\forall A \in \mathcal{F},\ A^c \equiv \Omega \setminus A \in \mathcal{F}$;
- ▷ *pour toute collection $A_1, A_2, \ldots \in \mathcal{F}$, $\bigcup_{n \geq 1} A_n \in \mathcal{F}$.*

La paire $(\Omega, \mathcal{F})$ est appelée un ***espace probabilisable****.*

On en déduit facilement les propriétés suivantes.

Lemme 5.3. *Une tribu $\mathcal{F}$ sur Ω possède les propriétés suivantes :*

- ▷ $\varnothing \in \mathcal{F}$;
- ▷ $\forall A, B \in \mathcal{F},\ A \cup B, A \cap B, A \setminus B \in \mathcal{F}$;
- ▷ *pour toute collection $A_1, A_2, \ldots \in \mathcal{F}$, $\bigcap_{n \geq 1} A_n \in \mathcal{F}$.*

Démonstration. Laissée en exercice. □

La façon habituelle de construire une tribu est de partir d'un ensemble $\mathcal{G}$ de « bons » sous-ensembles de Ω, faciles à décrire, et de définir $\mathcal{F}$ comme étant la plus petite tribu contenant $\mathcal{G}$. Cette approche repose sur l'observation suivante.

Lemme 5.4. *Soit $(\mathcal{F}_i, i \in I)$ une famille quelconque de tribus sur Ω. Alors $\bigcap_{i\in I} \mathcal{F}_i$ est également une tribu sur Ω.*

Démonstration. Exercice. □

Définition 5.2. *Soit $\mathcal{G} \subseteq \mathscr{P}(\Omega)$. On appelle* ***tribu engendrée par*** *$\mathcal{G}$, notée $\sigma(\mathcal{G})$, la plus petite tribu contenant $\mathcal{G}$,*

$$\sigma(\mathcal{G}) = \bigcap_{i\in I} \mathcal{F}_i,$$

où $(\mathcal{F}_i, i \in I)$ est la famille de toutes les tribus sur Ω contenant $\mathcal{G}$ (cette famille étant non-vide puisqu'elle contient toujours $\mathscr{P}(\Omega)$).

Exemple 5.1. *1.* Soit Ω dénombrable et $\mathcal{G} = \{\{\omega\} : \omega \in \Omega\}$ l'ensemble des singletons. Alors, $\sigma(\mathcal{G}) = \mathscr{P}(\Omega)$, puisque $A = \bigcup_{\omega\in A}\{\omega\}$ pour tout $A \in \mathscr{P}(\Omega)$.

2. Soit $\Omega = \mathbb{R}^n$ et soit

$$\mathcal{G} = \left\{ \prod_{i=1}^{n} [a_i, b_i] \,:\, a_i < b_i,\, a_i, b_i \in \mathbb{Q} \right\}.$$

La tribu $\mathcal{B}^n = \sigma(\mathcal{G})$ est appelée **tribu borélienne** sur $\mathbb{R}^n$ et les éléments de $\mathcal{B}^n$ sont appelés les **boréliens** de $\mathbb{R}^n$. Observons que $\mathcal{B}^n$ est un ensemble très riche :

- $\mathcal{B}^n$ contient tous les ouverts de $\mathbb{R}^n$. Il suffit en effet d'observer que, si A est un ouvert, alors on peut trouver, pour tout élément $\omega \in A$, un ensemble $B \in \mathcal{G}$ tel que $\omega \in B \subset A$. Par conséquent, $A = \bigcup_{\substack{B\in\mathcal{G}\\ B\subset A}} B$ est une union dénombrable d'éléments de $\mathcal{G}$ et appartient donc à $\mathcal{B}^n$.
- $\mathcal{B}^n$ contient tous les fermés de $\mathbb{R}^n$ (par stabilité de $\mathcal{F}$ sous passage au complémentaire).

3. La tribu $\mathcal{B} \equiv \mathcal{B}^1$ peut également être engendrée par

$$\mathcal{G}' = \{(-\infty, c] \,:\, c \in \mathbb{R}\}.$$

En effet, d'une part, les éléments de $\mathcal{G}'$ étant fermés, il suit que $\mathcal{G}' \subseteq \mathcal{B}$ et donc $\sigma(\mathcal{G}') \subseteq \mathcal{B}$ (par minimalité de la tribu engendrée). D'autre part, $\sigma(\mathcal{G}')$ contient tous les intervalles $(a, b] = (-\infty, b] \setminus (-\infty, a]$, et donc tous les intervalles $[a, b] = \bigcap_{n\geq 1}(a - \frac{1}{n}, b]$. $\sigma(\mathcal{G}')$ contient donc la tribu $\mathcal{B}$ engendrée par ces derniers (à nouveau, par minimalité).

4. Soit $\varnothing \neq A \subseteq \mathbb{R}^n$. L'ensemble $\mathcal{B}(A) = \{B \cap A \,:\, B \in \mathcal{B}^n\}$ est une tribu sur A, appelée la tribu des boréliens de A. ♢

Un exemple important est donné par la construction de la tribu-produit.

Définition 5.3. *Soit $(E_i, \mathcal{E}_i)$, $i \in I \neq \varnothing$, une collection d'espaces probabilisables, et $\Omega = \prod_{i\in I} E_i$ le produit cartésien des ensembles E_i. Soit $X_i : \Omega \to E_i$ l'application associant à $\omega = (\omega_i)_{i\in I} \in \Omega$ la composante ω_i. Soit encore $\mathcal{G} = \left\{ X_i^{-1}(A_i) \,:\, i \in I,\, A_i \in \mathcal{E}_i \right\}$ la collection des sous-ensembles de Ω spécifiés par un événement concernant une unique composante. Alors, la tribu $\bigotimes_{i\in I} \mathcal{E}_i = \sigma(\mathcal{G})$ est appelée la* ***tribu-produit*** *des $\mathcal{E}_i$ sur Ω.*

Dans le cas où $E_i = E$ et $\mathcal{E}_i = \mathcal{E}$, pour tout $i \in I$, on utilise la notation $\mathcal{E}^{\otimes I} \equiv \bigotimes_{i\in I} \mathcal{E}_i$.

Exemple 5.2. *1.* La tribu $\mathcal{B}^n$ est la tribu-produit de n copies de $\mathcal{B}$, $\mathcal{B}^n = \mathcal{B}^{\otimes n}$.

2. L'ensemble des trajectoires d'un processus stochastique en temps discret, à valeurs dans un ensemble au plus dénombrable, est de la forme $E^{\mathbb{N}}$, où E est l'ensemble au plus dénombrable

des états dans lequel peut se trouver le système en un temps donné. La tribu-produit $E^{\otimes\mathbb{N}}$ est alors précisément celle qui est engendrée par les cylindres

$$[s_1, \dots, s_n] = \{\omega \in \Omega : \omega_1 = s_1, \dots, \omega_n = s_n\},$$

où $n \in \mathbb{N}^*$ et $s_1, \dots, s_n \in E$. ♢

5.1.2 La mesure de probabilité

On veut à présent associer à chaque événement $A \in \mathcal{F}$ sa probabilité. La définition suivante, élaborée durant les premières décennies du XXème siècle, est généralement attribuée à Andreï Kolmogorov.

Définition 5.4. *Une* **mesure de probabilité** *sur un espace probabilisable $(\Omega, \mathcal{F})$ est une application $\mathbb{P} : \mathcal{F} \to [0,1]$ telle que*

- ▷ $\mathbb{P}(\Omega) = 1$,
- ▷ $\mathbb{P}(\bigcup_{k\geq 1} A_k) = \sum_{k\geq 1} \mathbb{P}(A_k)$, *pour toute collection $(A_k)_{k\geq 1}$ d'événements 2 à 2 disjoints.*

Le triplet $(\Omega, \mathcal{F}, \mathbb{P})$ est alors appelé un **espace de probabilité**.

Exemple 5.3. On vérifie facilement que, pour tout $\omega \in \Omega$, l'application $\delta_\omega : \mathscr{P}(\Omega) \to [0,1]$ définie par

$$\delta_\omega(A) = \begin{cases} 1 & \text{si } \omega \in A, \\ 0 & \text{sinon,} \end{cases}$$

est une mesure de probabilité sur $(\Omega, \mathscr{P}(\Omega))$. On l'appelle la **masse de Dirac** en ω.

Similairement, si $p \in [0,1]$, l'application $\mathbb{P} : \mathscr{P}(\mathbb{R}) \to [0,1]$ définie par

$$\mathbb{P} = p\,\delta_1 + (1-p)\,\delta_0$$

est une mesure de probabilité sur $(\mathbb{R}, \mathscr{P}(\mathbb{R}))$. Il s'agit évidemment de la loi de Bernoulli de paramètre p. ♢

Remarque 5.4. *Comme ses hypothèses sont satisfaites par définition, les propriétés listées dans le Corollaire 1.1 sont encore vérifiées pour une mesure de probabilité $\mathbb{P}$ générale (il faut évidement restreindre les événements concernés à la tribu $\mathcal{F}$).*

Nous aurons besoin des définitions suivantes.

Définition 5.5. *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 $A \subseteq B$;*
- ▷ $\bigcup_{i\geq 1} A_i \in \mathcal{M}$, *pour toute suite $A_1, A_2, \dots \in \mathcal{M}$ d'ensembles 2 à 2 disjoints.*

Comme on l'a fait pour les tribus, étant donné $\mathcal{C} \subseteq \mathscr{P}(\Omega)$, 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

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

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

Lemme 5.5 (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 $\sigma(\mathcal{C}) \subseteq \mathcal{M}(\mathcal{C})$.

La collection $\mathcal{D}_1 = \{A \subset \Omega : A \cap B \in \mathcal{M}(\mathcal{C}), \forall B \in \mathcal{C}\}$ est manifestement une classe monotone. De plus, $\mathcal{C}$ étant un π-système, $\mathcal{D}_1 \supseteq \mathcal{C}$. Par conséquent, $\mathcal{D}_1 \supseteq \mathcal{M}(\mathcal{C})$, 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 $B \in \mathcal{C}$.

Similairement, la collection $\mathcal{D}_2 = \{A \subset \Omega : A \cap B \in \mathcal{M}(\mathcal{C}), \forall B \in \mathcal{M}(\mathcal{C})\}$ est également une classe monotone, et par l'observation précédente, $\mathcal{D}_2 \supseteq \mathcal{C}$. Par conséquent, $\mathcal{D}_2 \supseteq \mathcal{M}(\mathcal{C})$, 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. Soient $A_1, A_2 \ldots \in \mathcal{M}(\mathcal{C})$. $\mathcal{M}(\mathcal{C})$ étant un π-système, il suit que les ensembles

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

appartiennent également à $\mathcal{M}(\mathcal{C})$ et sont disjoints 2 à 2. On en conclut que $\bigcup_{i\geq 1} A_i = \bigcup_{i\geq 1} B_i \in \mathcal{M}(\mathcal{C})$ ce qui termine la preuve. □

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

Théorème 5.1 (Théorème d'unicité). *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité, et supposons que $\mathcal{F} = \sigma(\mathcal{G})$ pour un certain $\mathcal{G} \subseteq \mathscr{P}(\Omega)$. Lorsque $\mathcal{G}$ est un π-système, $\mathbb{P}$ 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 $\mathcal{D} = \{A \in \mathcal{F} : \mathbb{P}(A) = \mathbb{Q}(A)\}$ est une classe monotone. De plus, $\mathcal{D} \supseteq \mathcal{G}$ par hypothèse. Il suit donc de la minimalité de $\mathcal{M}(\mathcal{G})$ que $\mathcal{D} \supseteq \mathcal{M}(\mathcal{G})$.

$\mathcal{G}$ étant un π-système, le Lemme 5.5 implique que $\mathcal{F} = \sigma(\mathcal{G}) = \mathcal{M}(\mathcal{G}) \subseteq \mathcal{D}$, et donc que $\mathcal{F} = \mathcal{D}$. □

Le problème qui se pose à présent est de construire une mesure de probabilité sur une tribu donnée. Grâce au Théorème 5.1, ceci revient à déterminer les conditions sous lesquelles il est possible d'étendre une fonction $\mathbb{P}$ définie sur un π-système à la tribu que ce dernier engendre. Une réponse à ce problème est donnée par le théorème de Carathéodory, qui sera démontré en Analyse III. Il repose sur la définition suivante.

Définition 5.6. *Une collection $\mathcal{A}$ de sous-ensembles de Ω est une* ***algèbre*** *si*

- $\Omega \in \mathcal{A}$;
- $\forall A \in \mathcal{A}$, $A^c \equiv \Omega \setminus A \in \mathcal{A}$;
- *pour tout* $A, B \in \mathcal{A}$, $A \cup B \in \mathcal{A}$.

Théorème 5.2 (Théorème d'extension de Carathéodory). *Une mesure de probabilité définie sur une algèbre $\mathcal{A}$ possède une unique extension à la tribu engendrée par $\mathcal{A}$.*

Esquissons à présent quelques applications importantes de ce résultat.

La mesure de Lebesgue sur $[0,1]$.

Soit $\Omega = [0,1]$. On aimerait construire une mesure de probabilité λ modélisant le tirage au hasard, uniformément (c'est-à-dire sans favoriser aucun nombre), d'un élément de $[0,1]$. En particulier, ceci implique que la mesure d'un intervalle $I = [a,b) \subseteq [0,1]$ arbitraire devrait être donné par sa longueur $b-a$. On sait donc comment définir la mesure λ sur la collection $\mathcal{I} = \{[a,b) : 0 \leq a < b \leq 1\}$. Cela ne suffit pas tout à fait, puisque $\mathcal{I}$ n'est pas une algèbre. On considère donc la collection

$$\mathcal{B}_0 = \{I_1 \cup \cdots \cup I_n : I_k \in \mathcal{I} \cup \{1\}, 1 \leq k \leq n, n \geq 1\}$$

de toutes les unions finies d'intervalles de $\mathcal{I}$ et du singleton $\{1\}$. $\mathbb{P}$ peut alors être facilement étendue aux éléments de $\mathcal{B}_0$ par additivité finie : si A est l'union de k intervalles disjoints $I_1, \ldots, I_k$, alors $\lambda(A) = \lambda(I_1) + \cdots + \lambda(I_k)$. On vérifie facilement que $\mathcal{B}_0$ est une algèbre et que $\sigma(\mathcal{B}_0) = \mathcal{B}([0,1])$.

Afin de pouvoir appliquer le Théorème 5.2, il reste à vérifier que l'application λ ainsi définie sur $\mathcal{B}_0$ est bien une mesure de probabilité, c'est-à-dire que si $A_1, A_2, \ldots \in \mathcal{B}_0$ sont 2 à 2 disjoints et si $A = \bigcup_{i=1}^{\infty} A_i \in \mathcal{B}_0$, alors $\lambda(A) = \sum_{i=1}^{\infty} \lambda(A_i)$. Nous ne le ferons pas ici (cela sera fait en Analyse III).

Une fois ceci établi, le Théorème 5.2 implique que λ possède une unique extension à une mesure de probabilité sur $\mathcal{B}([0,1])$.

Définition 5.7. *La mesure de probabilité λ sur $([0,1], \mathcal{B}([0,1]))$ construite ci-dessus est appelée* ***mesure de Lebesgue*** *sur $[0,1]$.*

Remarque 5.5. *On peut évidemment construire de la même façon la mesure sur $[0,1)$, $(0,1)$ ou $(0,1]$.*

Mesure-produit.

Soient $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ et $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ deux espaces de probabilité. On désire construire une mesure de probabilité $\mathbb{P}_1 \otimes \mathbb{P}_2$ sur $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ satisfaisant

$$\mathbb{P}_1 \otimes \mathbb{P}_2(A \times B) = \mathbb{P}_1(A)\mathbb{P}_2(B), \quad \forall A \in \mathcal{F}_1, B \in \mathcal{F}_2.$$

Ceci détermine donc les valeurs prises par $\mathbb{P}_1 \otimes \mathbb{P}_2$ sur la collection $\mathcal{I} = \{A \times B : A \in \mathcal{F}_1, B \in \mathcal{F}_2\}$. En utilisant l'additivité finie, on peut étendre la définition de $\mathbb{P}_1 \otimes \mathbb{P}_2$ à l'algèbre $\mathcal{F}_0$ obtenue en considérant toutes les unions finies d'éléments de $\mathcal{I}$. On montre ensuite que l'application $\mathbb{P}_1 \otimes \mathbb{P}_2$ est en fait une mesure de probabilité sur $\mathcal{F}_0$. La conclusion suit alors du Théorème 5.2, puisque $\sigma(\mathcal{F}_0) = \mathcal{F}_1 \otimes \mathcal{F}_2$.

Ceci permet, par exemple, la construction de la mesure de Lebesgue sur $([0,1]^n, \mathcal{B}([0,1]^n))$ à partir de la mesure de Lebesgue sur $[0,1]$.

5.2 Variables aléatoires

Nous allons à présent étendre la notion de variables aléatoires au contexte plus général considéré dans cette partie du cours.

Définition 5.8. *Une* ***variable aléatoire*** *sur un espace probabilisable $(\Omega, \mathcal{F})$ est une application $X : \Omega \to \mathbb{R}$ telle que*

$$X^{-1}(B) \in \mathcal{F}, \qquad \forall B \in \mathcal{B}.$$

La condition donnée dans cette définition est naturelle. En effet, les événements d'intérêt seront de la forme $\{X \in B\} = X^{-1}(B)$, $B \in \mathcal{B}$ (comme on l'a vu, $\mathcal{B}$ contient tous les sous-ensembles pertinents dans la pratique). Si l'on veut pouvoir associer une probabilité à un événement de ce type, il est nécessaire qu'il appartienne à $\mathcal{F}$.

Remarque 5.6. *Plus généralement, on peut considérer une paire d'espaces probabilisables $(\Omega, \mathcal{F})$ et $(\Omega', \mathcal{F}')$. Une* ***variable aléatoire*** *de $(\Omega, \mathcal{F})$ dans $(\Omega', \mathcal{F}')$ est une application $X : \Omega \to \Omega'$ telle que*

$$X^{-1}(A) \in \mathcal{F}, \qquad \forall A \in \mathcal{F}'.$$

*Dans ce cours, nous réserverons le terme de variable aléatoire aux applications à valeurs dans $\mathbb{R}$, et on parlera d'***applications mesurables** *de $(\Omega, \mathcal{F})$ dans $(\Omega', \mathcal{F}')$ pour ces applications plus générales.*

Exemple 5.4. Si Ω est au plus dénombrable, toute application $X : \Omega \to \mathbb{R}$ est une variable aléatoire sur $(\Omega, \mathscr{P}(\Omega))$. ♢

Lemme 5.6. *Si $\mathcal{F}' = \sigma(\mathcal{G}')$, alors*

$$X^{-1}(A) \in \mathcal{F}, \quad \forall A \in \mathcal{F}' \quad \Leftrightarrow \quad X^{-1}(A) \in \mathcal{F}, \quad \forall A \in \mathcal{G}'.$$

Démonstration. Supposons que $X^{-1}(A) \in \mathcal{F}$ pour tout $A \in \mathcal{G}'$. La collection

$$\mathcal{A}' = \left\{ A \subseteq \Omega' \,:\, X^{-1}(A) \in \mathcal{F} \right\}$$

est alors une tribu contenant $\mathcal{G}'$. Comme $\mathcal{F}' = \sigma(\mathcal{G}')$, on a donc $\mathcal{A}' \supseteq \mathcal{F}'$, et donc $X^{-1}(A) \in \mathcal{F}$ pour tout $A \in \mathcal{F}'$. □

Exemple 5.5. Il suit du Lemme 5.6 et de l'Exemple 5.1 (point 3) que X est une variable aléatoire si et seulement si

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

(On peut bien sûr remplacer $\leq$ par $<$, $>$ ou $\geq$.) ♢

Exemple 5.6. Soit $\Omega \subseteq \mathbb{R}^n$ et $\mathcal{F} = \mathcal{B}^n(\Omega)$. Alors, toute fonction continue $X : \Omega \to \mathbb{R}$ est une variable aléatoire. En effet, pour tout $c \in \mathbb{R}$, $\{X \leq c\}$ est un fermé de Ω et appartient donc à $\mathcal{B}^n(\Omega)$, par le point 2 de l'Exemple 5.1. ♢

5.2.1 Loi et fonction de répartition

Comme dans le cas discret, la mesure de probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{F})$ et la variable aléatoire X induisent une mesure de probabilité $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$ sur $\mathbb{R}$. Cela suit du résultat général suivant.

Théorème 5.3. *Soit X une application mesurable d'un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ dans un espace probabilisable $(\Omega', \mathcal{F}')$. Alors l'application $\mathbb{P}' : \mathcal{F}' \to [0, 1]$ définie par*

$$\mathbb{P}'(A') = \mathbb{P}(X^{-1}(A')), \quad \forall A' \in \mathcal{F}',$$

est une mesure de probabilité sur $(\Omega', \mathcal{F}')$.

Démonstration. On vérifie immédiatement que $\mathbb{P}'(\Omega') = 1$. D'autre part, si $A'_1, A'_2, \ldots \in \mathcal{F}'$ sont 2 à 2 disjoints, alors c'est également le cas de leur préimages $X^{-1}(A'_1), X^{-1}(A'_2), \ldots$, ce qui implique que

$$\mathbb{P}'\Big(\bigcup_{n\geq 1} A'_n\Big) = \mathbb{P}\Big(X^{-1}\Big(\bigcup_{n\geq 1} A'_n\Big)\Big) = \mathbb{P}\Big(\bigcup_{n\geq 1} X^{-1}(A'_n)\Big) = \sum_{n\geq 1} \mathbb{P}(X^{-1}(A'_n)) = \sum_{n\geq 1} \mathbb{P}'(A'_n).$$

□

Définition 5.9. *La mesure de probabilité $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$ sur $(\mathbb{R}, \mathcal{B})$ induite par une variable aléatoire X est appelée la* ***loi*** *de X.*

On dit de deux variables aléatoires X et Y qu'elles ont la même loi, ou qu'elles sont ***identiquement distribuées****, lorsque $\mathbb{P}_X = \mathbb{P}_Y$. Dans ce cas, on écrira $X \overset{\text{loi}}{=} Y$.*

Exemple 5.7. On peut clairement caractériser une variable aléatoire discrète comme étant une variable aléatoire dont la loi est de la forme

$$\mathbb{P}_X = \sum_{k \in I} p_k \delta_{x_k},$$

où $I \neq \varnothing$ est un ensemble au plus dénombrable, les x_k, $k \in I$, sont des réels distincts, et les p_k, $k \in I$, des réels strictement positifs tels que $\sum_{k \in I} p_k = 1$. ♢

Il suit du Théorème d'unicité 5.1 et du point 3 de l'Exemple 5.1 qu'une mesure de probabilité $\mathbb{P}$ sur $(\mathbb{R}, \mathcal{B})$ est entièrement déterminée par la fonction $F_{\mathbb{P}}(x) = \mathbb{P}((-\infty, x])$. En particulier, la loi d'une variable aléatoire X est entièrement déterminée par la fonction $F_X(x) = \mathbb{P}(X \leq x)$.

Définition 5.10. *Soit $\mathbb{P}$ une mesure de probabilité sur $(\mathbb{R}, \mathcal{B})$. La fonction $F_{\mathbb{P}} : \mathbb{R} \to [0, 1]$, $F_{\mathbb{P}}(x) = \mathbb{P}((-\infty, x])$, est appelée* ***fonction de répartition*** *de $\mathbb{P}$.*

Soit X une variable aléatoire sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$. La fonction $F_X : \mathbb{R} \to [0, 1]$, $F_X(x) = \mathbb{P}(X \leq x)$, est appelée ***fonction de répartition*** *de X.*

Lemme 5.7. *La fonction de répartition $F_{\mathbb{P}} : \mathbb{R} \to [0, 1]$ associée à une mesure de probabilité sur $(\mathbb{R}, \mathcal{B})$ possède les propriétés suivantes :*

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

Démonstration. Les trois premières affirmations sont immédiates. La quatrième est une conséquence du Lemme 5.1 : pour toute suite $x_n \downarrow x$,

$$\lim_{n \to \infty} F_{\mathbb{P}}(x_n) = \lim_{n \to \infty} \mathbb{P}((-\infty, x_n]) = \mathbb{P}\Big(\bigcap_{n \geq 1} (-\infty, x_n]\Big) = \mathbb{P}((-\infty, x]) = F_{\mathbb{P}}(x).$$

□

En fait, on peut montrer que les quatre propriétés ci-dessus caractérisent les fonctions de répartition : toute fonction $F : \mathbb{R} \to \mathbb{R}$ les possédant est la fonction de répartition associée à une mesure de probabilité sur $(\Omega, \mathcal{F})$.

Lemme 5.8. *Soit $F : \mathbb{R} \to \mathbb{R}$ une fonction possédant les quatre propriétés du Lemme 5.7. Alors, il existe une variable aléatoire X sur $((0, 1), \mathcal{B}((0, 1)), \lambda)$ telle que $F_X = F$. Cette variable aléatoire est explicitement donnée par*

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

Démonstration. Les propriétés 2 et 3 impliquent que $X(u) \in (-\infty, \infty)$ pour tout $u \in (0, 1)$. Observons à présent que l'infimum est en fait un minimum, puisque F est continue à droite. Ceci implique que $X(u) \leq c$ si et seulement si $u \leq F(c)$. En particulier, $\{X \leq c\} = (0, F(c)] \cap (0, 1) \in \mathcal{B}((0, 1))$, et donc X est une variable aléatoire, par la discussion de l'Exemple 5.5. De plus, $F_X(c) = \lambda((0, F(c)]) = F(c)$ pour tout $c \in \mathbb{R}$. □

Lemme 5.9. *Soit X une variable aléatoire de fonction de répartition F_X. Alors,*

1. $\mathbb{P}(X > x) = 1 - F_X(x)$,
2. $\mathbb{P}(x < X \leq y) = F_X(y) - F_X(x)$,
3. $\mathbb{P}(X = x) = F_X(x) - \lim_{y \uparrow x} F_X(y)$.

Démonstration. Les deux premières affirmations sont immédiates. Pour la troisième, on considère les événements $A_n = \{x - \frac{1}{n} < X \leq x\}$. Puisque $\lim_{n\to\infty} A_n = \{X = x\}$, il suit du Lemme 5.1 que

$$\mathbb{P}(X = x) = \lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} (F_X(x) - F_X(x - \tfrac{1}{n})),$$

par le point *2.* (la limite existe par monotonicité de F_X). □

5.3 Indépendance

Les notions de probabilité conditionnelle et d'indépendance d'événements et de variables aléatoires introduites dans la première partie du cours sont inchangées, modulo la restriction des événements à la tribu $\mathcal{F}$.

Définition 5.11. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité.*

Soient $B \in \mathcal{F}$ tel que $\mathbb{P}(B) > 0$ et $A \in \mathcal{F}$. La ***probabilité conditionnelle de A sachant B*** *est définie par $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$.*

Une famille d'événements $(A_i)_{i\in I}$ est ***indépendante*** *sous $\mathbb{P}$ si*

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

pour tous les sous-ensembles finis J de I.

Une famille de variables aléatoires $(X_i)_{i\in I}$ est ***indépendante*** *si les événements*

$$\{X_i \in A_i\}, i \in J,$$

sont indépendants pour tout $A_i \in \mathcal{B}$, $i \in J$, et tout $J \subset I$ fini.

À nouveau, on vérifie aisément que $\mathbb{P}(\cdot \mid B)$ est un mesure de probabilité sur les espaces probabilisables $(\Omega, \mathcal{F})$ et $(B, \mathcal{F}(B))$, où $\mathcal{F}(B) = \{A \cap B : A \in \mathcal{F}\}$.

On a vu dans la première partie du cours qu'afin de vérifier l'indépendance de variables aléatoires discrètes, il suffisait de considérer des singletons $A_i = \{x_i\}$ avec $x_i \in X_i(\Omega)$. Dans le cas général considéré ici, on peut montrer qu'il est également possible de se restreindre à une classe particulière d'événements.

Lemme 5.10. *La famille de variables aléatoire $(X_i)_{1\leq i\leq n}$ est indépendante si et seulement si*

$$\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x_i),$$

pour tout $x_i \in \mathbb{R}$, $1 \leq i \leq n$.

On a également vu, dans le cas de variables aléatoires discrètes, que si les variables aléatoires $(X_i)_{i\in I}$ sont indépendantes, alors il en est de même des variables aléatoires $(\varphi_i(X_i))_{i\in I}$. On peut également étendre ce résultat au contexte actuel.

Lemme 5.11. *Soit $(X_i)_{i\in I}$ des variables aléatoires indépendantes sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, et soient $(\varphi_i)_{i\in I}$ une famille de variables aléatoires sur $(\mathbb{R}, \mathcal{B})$. Alors, les variables aléatoires $(\varphi_i(X_i))$ sont indépendantes.*

Démonstration. Identique à celle du Lemme 2.4. □

5.4 Espérance

Étant à présent en possession d'une notion appropriée de variable aléatoire, il est temps de nous tourner vers la généralisation du concept d'espérance.

L'idée de départ est simple : nous sommes déjà en possession d'un concept d'espérance pour les variables aléatoires discrètes. Il est donc naturel d'essayer d'approximer nos variables aléatoires générales par des variables discrètes.

Soit donc $X : \Omega \to \mathbb{R}$ une variable aléatoire, et $n \in \mathbb{N}^*$. On considère la discrétisation de X à l'échelle $1/n$:

$$X^{(n)} = \frac{1}{n}\lfloor nX \rfloor.$$

Ainsi, $X^{(n)}(\omega) = k/n$ lorsque $k/n \leq X(\omega) < (k+1)/n$, $k \in \mathbb{Z}$. Manifestement, $X^{(n)}$ est une variable aléatoire discrète. De plus, pour tout $n \in \mathbb{N}^*$,

$$X^{(n)} \leq X < X^{(n)} + \frac{1}{n}. \tag{5.1}$$

On aimerait définir l'espérance de X comme limite des espérance des variables aléatoires discrètes $X^{(n)}$ lorsque $n \to \infty$. Cela est possible, car la suite $(\mathbb{E}(X^{(n)}))_{n\geq 1}$ est une suite de Cauchy. En effet, il suit de (5.1) que, pour tout $m, n \in \mathbb{N}^*$, $X^{(n)} \leq X < X^{(m)} + 1/m$; en particulier,

$$|X^{(n)} - X^{(m)}| \leq \max(\frac{1}{m}, \frac{1}{n}).$$

On en déduit que si $\mathbb{E}(|X^{(m)}|) < \infty$ pour un certain $m \in \mathbb{N}^*$, alors c'est le cas pour tout $n \in \mathbb{N}^*$. On a donc bien, pour tout $m, n \in \mathbb{N}^*$,

$$|\mathbb{E}(X^{(n)}) - \mathbb{E}(X^{(m)})| \leq \mathbb{E}(|X^{(n)} - X^{(m)}|) \leq \max(\frac{1}{m}, \frac{1}{n}).$$

Définition 5.12. *Soit $X : \Omega \to \mathbb{R}$ une variable aléatoire sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. On dit que X possède une espérance si $\mathbb{E}(|X^{(n)}|) < \infty$ pour un (et donc tous les) $n \in \mathbb{N}^*$. Dans ce cas, l'**espérance** de X est*

$$\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(X^{(n)}).$$

On écrira alors $X \in \mathcal{L}^1 \equiv \mathcal{L}^1(\mathbb{P})$.

Remarque 5.7. *Évidemment, si X est une variable aléatoire discrète, la définition précédente de l'espérance coïncide avec celle de la première partie du cours, puisque*

$$\lim_{n\to\infty} |\mathbb{E}(X^{(n)}) - \mathbb{E}(X)| \leq \lim_{n\to\infty} \mathbb{E}(|X^{(n)} - X|) \leq \lim_{n\to\infty} \tfrac{1}{n} = 0.$$

Remarque 5.8. *Il est important d'observer que l'espérance de X ne dépend en fait que de la loi $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$ de X. Plus précisément, on a*

$$\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(X^{(n)}) = \lim_{n\to\infty} \sum_{k\in\mathbb{Z}} \frac{k}{n} \mathbb{P}_X([k/n, (k+1)/n)) = \mathbb{E}_{\mathbb{P}_X}(\mathrm{Id}_{\mathbb{R}}),$$

où $\mathrm{Id}_{\mathbb{R}}$ est la variable aléatoire identité sur $\mathbb{R}$: $\mathrm{Id}_{\mathbb{R}} : \mathbb{R} \to \mathbb{R}$, $\mathrm{Id}_{\mathbb{R}}(x) = x$, et l'espérance du membre de droite est prise par rapport à la mesure de probabilité $\mathbb{P}_X$ sur $(\mathbb{R}, \mathcal{B})$.

Les principales propriétés de l'espérance établie dans le cas de variables aléatoires discrètes s'étendent aisément au cas général.

Théorème 5.4. *Soient $Y, X, X_n \in \mathcal{L}^1$, pour tout $n \geq 1$. Alors,*

1. $X \geq 0 \implies \mathbb{E}(X) \geq 0$.
2. $\mathbb{E}(|X|) \geq |\mathbb{E}(X)|$.
3. *(Linéarité) Pour tout $\alpha, \beta \in \mathbb{R}$, $\alpha X + \beta Y \in \mathcal{L}^1$ et*
$$\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}(X) + \beta \mathbb{E}(Y).$$
4. *(σ-additivité) Supposons que les variables aléatoires X_n, $n \geq 1$, soient positives et que $X = \sum_{n\geq 1} X_n$. Alors,*
$$\mathbb{E}(X) = \sum_{n\geq 1} \mathbb{E}(X_n).$$
5. *(Convergence monotone) Lorsque les variables aléatoires X_n, $n \geq 1$, satisfont $X_n \nearrow X$ ponctuellement, on a*
$$\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(X_n).$$
6. *Si X et Y sont indépendantes, alors $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.*
7. *Les inégalités de Cauchy-Schwarz et Jensen sont vérifiées.*

Démonstration. On se ramène aux propriétés établies dans le cas discret, *cf.* Lemme 2.8.

1. Évident, puisque $X \geq 0$ implique $X_n \geq 0$ pour tout n.

2. Comme $\big||X^{(n)}| - |X|^{(n)}\big| \leq 2/n$,
$$\mathbb{E}(|X|) = \lim_{n\to\infty} \mathbb{E}(|X|^{(n)}) = \lim_{n\to\infty} \mathbb{E}(|X^{(n)}|) \geq \lim_{n\to\infty} |\mathbb{E}(X^{(n)})| = |\mathbb{E}(X)|.$$

3. Comme $|(\alpha X + \beta Y)^{(n)} - \alpha X^{(n)} - \beta Y^{(n)}| \leq (1 + |\alpha + \beta|)/n$,
$$\mathbb{E}(\alpha X + \beta Y) = \lim_{n\to\infty} \mathbb{E}(\alpha X^{(n)} + \beta Y^{(n)}) = \lim_{n\to\infty} \alpha \mathbb{E}(X^{(n)}) + \beta \mathbb{E}(Y^{(n)}) = \alpha \mathbb{E}(X) + \beta \mathbb{E}(Y).$$

4. Le même argument que dans le cas discret montre que $\mathbb{E}(X) \geq \sum_{n\geq 1} \mathbb{E}(X_n)$. Pour l'inégalité dans l'autre direction, on introduit les variables aléatoires discrètes $Y_{n,k} = X_n^{(2^{n+k})}$. Il suit alors de (5.1) que $Y_{n,k} \leq X_n < Y_{n,k} + 2^{-n-k}$. Par conséquent, en utilisant les propriétés déjà démontrées, on obtient, pour tout $k \geq 1$,
$$\mathbb{E}(X) = \mathbb{E}(\sum_{n\geq 1} X_n) \leq \mathbb{E}(\sum_{n\geq 1} Y_{n,k} + 2^{-n-k}) = \sum_{n\geq 1} \mathbb{E}(Y_{n,k}) + 2^{-k} \leq \sum_{n\geq 1} \mathbb{E}(X_n) + 2^{-k}.$$

5. suit de *4* comme dans la preuve du Lemme 2.8.

6. Comme $|(XY)^{(n)} - X^{(n)}Y^{(n)}| \leq (|X| + |Y| + 1)/n$, on a
$$\mathbb{E}(XY) = \lim_{n\to\infty} \mathbb{E}((XY)^{(n)}) = \lim_{n\to\infty} \mathbb{E}(X^{(n)}Y^{(n)}) = \lim_{n\to\infty} \mathbb{E}(X^{(n)})\mathbb{E}(Y^{(n)}) = \mathbb{E}(X)\mathbb{E}(Y).$$

7. Au vu des propriétés déjà établies, les preuves données dans le cas discret s'appliquent ici. □

Les deux résultats suivants se révèlent également très souvent utiles.

Théorème 5.5. *Soient $X_n \in \mathcal{L}^1$, $n \geq 1$.*

1. *(Lemme de Fatou) Supposons que $X_n \geq 0$. Alors,*

$$\mathbb{E}(\liminf_{n\to\infty} X_n) \leq \liminf_{n\to\infty} \mathbb{E}(X_n).$$

2. *(Convergence dominée) Supposons que $X_n \to X$ ponctuellement, et qu'il existe une variable aléatoire $Y \in \mathcal{L}^1$ telle que $|X_n| \leq Y$ pour tout n. Alors, $X \in \mathcal{L}^1$ et*

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

Démonstration. 1. Soit $Y_n = \inf_{k\geq n} X_k$ et $Y = \lim_{n\to\infty} Y_n = \liminf_{n\to\infty} X_n$. On a donc d'une part que $Y_n \leq X_n$, et d'autre part que $Y_n \nearrow Y$. Il suit donc, par convergence monotone, que

$$\liminf_{n\to\infty} \mathbb{E}(X_n) \geq \liminf_{n\to\infty} \mathbb{E}(Y_n) = \mathbb{E}(Y) = \mathbb{E}(\liminf_{n\to\infty} X_n).$$

2. Comme $X_n + Y \geq 0$, il suit du lemme de Fatou que

$$\mathbb{E}(X) + \mathbb{E}(Y) = \mathbb{E}(X+Y) \leq \liminf_{n\to\infty} \mathbb{E}(X_n + Y) = \liminf_{n\to\infty} \mathbb{E}(X_n) + \mathbb{E}(Y),$$

et donc, puisque $\mathbb{E}(Y) < \infty$, que $\mathbb{E}(X) \leq \liminf_{n\to\infty} \mathbb{E}(X_n)$.

De la même façon, puisque $-X_n + Y \geq 0$, on obtient que $\mathbb{E}(X) \geq \limsup_{n\to\infty} \mathbb{E}(X_n)$. La conclusion suit. □

Remarque 5.9. *On vérifie aisément que les théorèmes de convergences monotone et dominée restent vrais si l'on remplace la convergence ponctuelle par la convergence* ***presque-sûre****, c'est-à-dire la convergence ponctuelle sur un sous-ensemble $\Omega_0 \subseteq \Omega$ tel que $\mathbb{P}(\Omega_0) = 1$.*

Ayant à disposition une notion d'espérance, on peut définir comme précédemment les moments d'ordres supérieurs et la variance. Les propriétés démontrées dans le cas discret restent vraies, car leurs preuves ne reposent que sur les propriétés de l'espérance que l'on vient de généraliser. On ne les répètera pas ici.

5.5 Variables aléatoires à densité

5.5.1 Espérance et intégration

La notion d'espérance introduite précédemment partage de nombreuses propriétés avec l'intégrale de Riemann : elle est linéaire, elle préserve l'ordre, etc. En fait, cela va beaucoup plus loin.

Théorème 5.6. *On considère l'espace de probabilité $([0,1], \mathcal{B}([0,1]), \lambda)$, où λ est la mesure de Lebesgue. Soit $X : [0,1] \to \mathbb{R}$ une fonction Riemann-intégrable bornée. Alors, X est une variable aléatoire et $\mathbb{E}(X) = \int_0^1 X(s)\,\mathrm{d}s$.*

Nous ne démontrerons pas ce théorème et nous contenterons de renvoyer à la Figure ?? ; une preuve sera donnée dans le cours de théorie de la mesure.

Au vu du Théorème 5.6, on voit que $\mathbb{E}(X)$ fournit une généralisation de la notion d'intégrale au sens de Riemann : c'est ce que l'on appelle l'intégrale de Lebesgue de X. On la notera $\mathbb{E}(X) = \int_0^1 X(s)\,\lambda(\mathrm{d}s) \equiv \int_0^1 X\,\mathrm{d}\lambda$. Cette dernière permet d'intégrer beaucoup plus de fonctions que l'intégrale de Riemann, et a l'avantage considérable de se comporter beaucoup mieux sous passage à la limite comme le montrent les théorèmes de convergence monotone et dominée.

Remarque 5.10. *Si X est une variable aléatoire sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$, on peut également considérer $\mathbb{E}(X)$ comme définissant l'intégrale de la fonction X sur Ω, relativement à la mesure de probabilité $\mathbb{P}$. Pour cette raison, on utilise souvent la notation*

$$\mathbb{E}(X) = \int_\Omega X(\omega)\mathbb{P}(\mathrm{d}\omega) \equiv \int X \,\mathrm{d}\mathbb{P}.$$

Finalement, observons qu'il n'est pas nécessaire de restreindre le domaine d'intégration à l'intervalle $[0, 1]$. Si $X : \mathbb{R} \to \mathbb{R}$ est mesurable, on peut en effet définir

$$\int_{-\infty}^{\infty} X(s)\,\lambda(\mathrm{d}s) = \sum_{n\in\mathbb{Z}} \int_0^1 X(s+n)\,\lambda(\mathrm{d}s).$$

Bien entendu, λ dans le membre de gauche n'est pas une mesure de probabilité : $\lambda(\mathbb{R}) = +\infty$. Une application σ-additive $\mu : \mathcal{F} \to \mathbb{R}^+$ est appelée une mesure. On va voir à présent que cette mesure permet de définir de nombreuses mesures de probabilité d'intérêt sur $\mathcal{B}(\mathbb{R})$.

Terminologie : on dira qu'une propriété des nombres réels est vraie **presque partout** si l'ensemble des points où elle n'est pas vérifiée est inclus dans un ensemble de mesure de Lebesgue nulle.

Notation : Afin d'alléger les notations, nous noterons simplement $\int f(s)\,\mathrm{d}s$ au lieu de $\int f(s)\,\lambda(\mathrm{d}s)$. Ceci ne devrait pas prêter à confusion, toutes les intégrales considérées étant prises au sens de Lebesgue. De plus, on utilisera la notation habituelle $\int_A f(s)\,\mathrm{d}s \equiv \int f(s)\mathbf{1}_A(s)\,\mathrm{d}s$.

5.5.2 Densité de probabilité, absolue continuité

Définition 5.13. *Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction mesurable positive, telle que $\int_{-\infty}^{\infty} f(s)\,\mathrm{d}s = 1$. Alors l'application $\mathbb{P} : \mathcal{B} \to [0, 1]$ définie par*

$$\mathbb{P}(A) = \int_A f(s)\,\mathrm{d}s, \qquad \forall A \in \mathcal{B},$$

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

Remarque 5.11.

1. *Insistons sur le fait que la valeur $f(s)$ n'est pas une probabilité (en particulier, $f(s)$ peut être plus grande que 1). Par contre, il peut être utile de penser à $f(s)\,\mathrm{d}s$ comme à la probabilité de l'intervalle $[s, s + \mathrm{d}s]$.*
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 0, alors g est également une densité de probabilité pour $\mathbb{P}$. En effet, si $B = \{x : f(x) \neq g(x)\}$, alors*

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

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

En particulier, la fonction de répartition associée à une mesure de probabilité absolument continue $\mathbb{P}$ de densité de probabilité f satisfait

$$F_{\mathbb{P}}(x) = \mathbb{P}((-\infty, x]) = \int_{-\infty}^{x} f(s)\,\mathrm{d}s.$$

En particulier, $F_{\mathbb{P}}$ est continue [1].

Remarque 5.12. *On peut fabriquer des fonctions de répartition F (assez pathologiques) qui sont continues, mais qui ne sont pas associées à des mesures de probabilité absolument continues. Les mesures de probabilité correspondantes sont dites* ***singulières****.*

Remarque 5.13. *On peut démontrer (c'est le Théorème de différentiation de Lebesgue) que $F_{\mathbb{P}}$ est différentiable presque partout. Ceci permet d'associer à $\mathbb{P}$ une densité de probabilité de façon canonique : on prendra $f(x) = F'_{\mathbb{P}}(x)$ en tout point où $F_{\mathbb{P}}$ est différentiable, et $f(x) = 0$ ailleurs.*

Tout ceci s'étend naturellement aux variables aléatoires.

Définition 5.14. *Une variable aléatoire X est* ***à densité*** *si sa loi est absolument continue. La densité de probabilité associée à $\mathbb{P}_X$ est alors notée f_X et appelée* ***densité de probabilité*** *de X.*

Ainsi, pour toute variable aléatoire X à densité,

$$\mathbb{P}(X \in A) = \int_A f_X(s)\,\mathrm{d}s, \qquad \forall A \in \mathcal{B}.$$

Théorème 5.7. *Soit $\Omega \in \mathcal{B}$ et $\mathbb{P}$ une mesure de probabilité sur $(\Omega, \mathcal{B}(\Omega))$ avec densité de probabilité f, et soit X une variable aléatoire sur Ω. Alors, $X \in \mathcal{L}^1(\mathbb{P})$ si et seulement si $\int_\Omega |X(s)|f(s)\,\mathrm{d}s < \infty$ et, dans ce cas,*

$$\mathbb{E}(X) = \int_\Omega X(s)f(s)\,\mathrm{d}s.$$

Démonstration. Pour chaque $n \geq 1$, on voit que $\mathbb{E}(|X^{(n)}|) < \infty$ si et seulement si l'expression

$$\begin{aligned}\sum_{k\in\mathbb{Z}} |\tfrac{k}{n}|\,\mathbb{P}(X^{(n)} = \tfrac{k}{n}) &= \sum_{k\in\mathbb{Z}} |\tfrac{k}{n}| \int_{\frac{k}{n}\leq X < \frac{k+1}{n}} f(s)\,\mathrm{d}s \\ &= \sum_{k\in\mathbb{Z}} \int_{\frac{k}{n}\leq X < \frac{k+1}{n}} |X^{(n)}(s)|f(s)\,\mathrm{d}s = \int_\Omega |X^{(n)}(s)|f(s)\,\mathrm{d}s\end{aligned}$$

est finie. Comme $|X - X^{(n)}| < \frac{1}{n}$, ceci a lieu si et seulement si $\int_\Omega |X(s)|f(s)\,\mathrm{d}s < \infty$.

On a alors

$$\mathbb{E}(X^{(n)}) = \int_\Omega X^{(n)}(s)f(s)\,\mathrm{d}s \to \int_\Omega X(s)f(s)\,\mathrm{d}s,$$

puisque $\int_\Omega X^{(n)}(s)f(s)\,\mathrm{d}s \leq \int_\Omega X(s)f(s)\,\mathrm{d}s \leq \int_\Omega X^{(n)}(s)f(s)\,\mathrm{d}s + \frac{1}{n}$. □

Corollaire 5.1. *Soit X une variable aléatoire à densité, et $\varphi : \mathbb{R} \to \mathbb{R}$ une application mesurable telle que $\varphi(X) \in \mathcal{L}^1$. Alors, l'espérance de la variable aléatoire $\varphi(X)$ satisfait*

$$\mathbb{E}(\varphi(X)) = \int_{\mathbb{R}} \varphi(s)\,f_X(s)\,\mathrm{d}s.$$

1. Cela n'est pas complètement évident si f n'est pas bornée. Une façon de procéder est la suivante. On fixe $\epsilon > 0$. Pour $n \geq 1$, on introduit $f_n = \min(f, n)$. On a alors $f_n \uparrow f$ lorsque $n \to \infty$. Par le Théorème de convergence monotone, on a que $\int_{\mathbb{R}} f_n(x)\mathrm{d}x \to \int_{\mathbb{R}} f(x)\mathrm{d}x$. On peut donc trouver n assez grand pour que $\int_{\mathbb{R}}(f(x) - f_n(x)) < \epsilon$. On a alors

$$\int_A f(x)\mathrm{d}x = \int_A (f(x) - f_n(x))\mathrm{d}x + \int_A f_n(x)\mathrm{d}x \leq \int_{\mathbb{R}} (f(x) - f_n(x))\mathrm{d}x + n\lambda(A) \leq \epsilon + n\lambda(A) \leq 2\epsilon,$$

pour tout $A \in \mathcal{B}$ tel que $\lambda(A) \leq \delta = \epsilon/n$. La continuité suit, puisque $\lambda([x, x+\delta]) = \delta$.

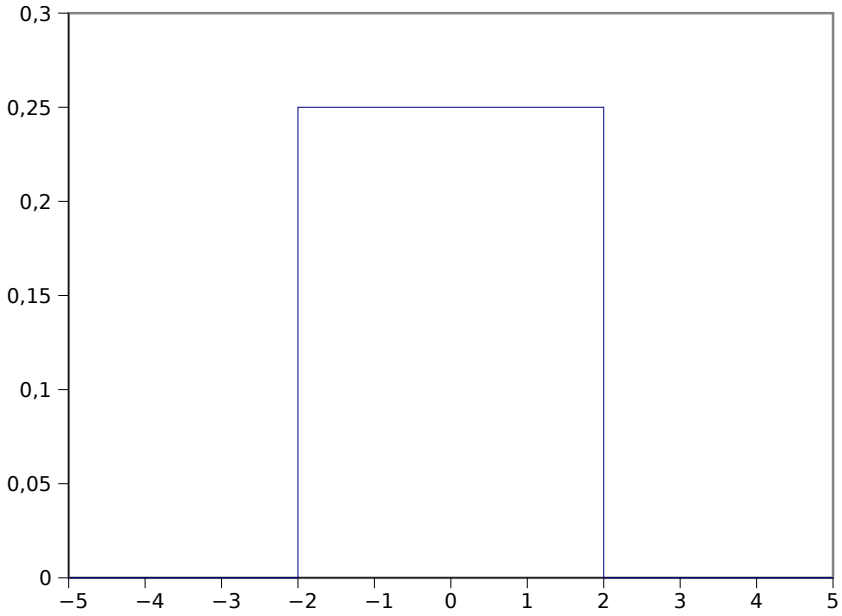


FIGURE 5.1: Densité de probabilité de la loi uniforme sur $[-2, 2]$.

Remarque 5.14. *Observez la similarité formelle avec le résultat correspondant pour les variables discrètes :* $\mathbb{E}(\varphi(X)) = \sum_{x \in X(\Omega)} \varphi(x) f_X(x)$.

Démonstration. Il suit de la Remarque 5.8 que

$$\mathbb{E}_{\mathbb{P}}(\varphi \circ X) = \mathbb{E}_{\mathbb{P} \circ (\varphi \circ X)^{-1}}(\mathrm{Id}_{\mathbb{R}}),$$

où l'on a explicité les mesures par rapport auxquelles les espérances sont prises. Une seconde application de la Remarque 5.8 implique donc que

$$\mathbb{E}_{\mathbb{P} \circ X^{-1}}(\varphi) = \mathbb{E}_{\mathbb{P} \circ X^{-1} \circ \varphi^{-1}}(\mathrm{Id}_{\mathbb{R}}) = \mathbb{E}_{\mathbb{P} \circ (\varphi \circ X)^{-1}}(\mathrm{Id}_{\mathbb{R}}) = \mathbb{E}_{\mathbb{P}}(\varphi \circ X).$$

La conclusion suit du théorème précédent, puisque

$$\mathbb{E}_{\mathbb{P} \circ X^{-1}}(\varphi) = \int_{\mathbb{R}} \varphi(s)\, f_X(s)\, \mathrm{d}s.$$

□

5.5.3 Exemples importants de variables aléatoires à densité

On présente ici quelques-unes des lois à densité les plus importantes. Elles sont introduites à partir de leur densité de probabilité, et il est laissé en exercice de vérifier que ses densités de probabilité sont proprement normalisées (c'est-à-dire d'intégrale 1).

Loi uniforme

Soient $a < b$. X est **uniforme** sur $[a, b]$, noté $X \sim \mathrm{U}(a, b)$, si elle a densité de probabilité

$$f_X(x) = \frac{1}{b - a}\, \mathbf{1}_{[a,b]}(x).$$

Cette distribution modélise le tirage d'un élément de l'intervalle $[a, b]$ de façon uniforme, c'est-à-dire sans en privilégier aucun.

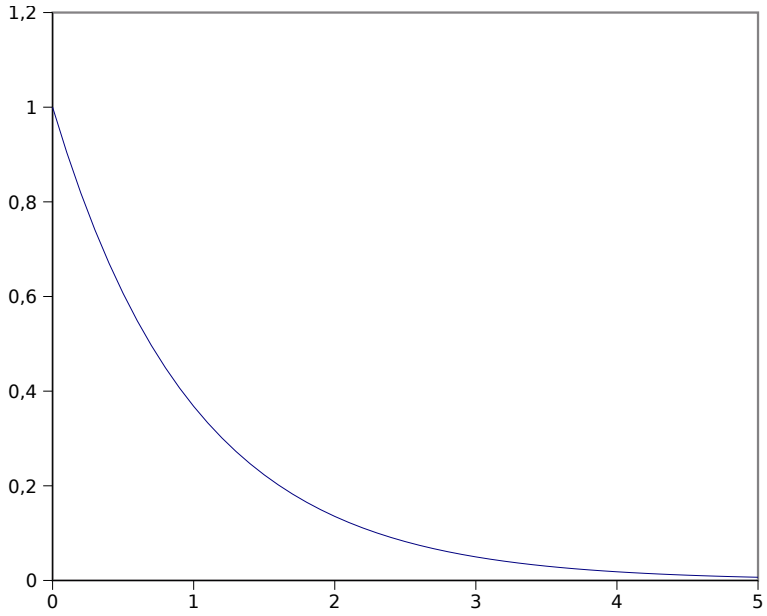


FIGURE 5.2: Densité de probabilité de la loi exponentielle pour $\lambda = 1$.

L'espérance et la variance de X se calcule aisément :

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

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \tfrac{1}{b-a}\int_a^b s^2\,\mathrm{d}s - \tfrac{(a+b)^2}{4} = \tfrac{a^2+ab+b^2}{3} - \tfrac{(a+b)^2}{4} = \tfrac{(b-a)^2}{12}.$$

Plus généralement, pour tout $A \in \mathcal{B}$, on X est uniforme sur A, $X \sim \mathrm{U}(A)$ si

$$f_X(x) = \frac{1}{\lambda(A)}\,\mathbf{1}_A(x).$$

Loi exponentielle

X est exponentielle de paramètre $\lambda > 0$, $X \sim \exp(\lambda)$ si elle admet pour densité de probabilité

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

Cette loi joue un rôle central dans la théorie des processus markoviens à temps continu ; nous en aurons un aperçu dans le chapitre 11.

Elle 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 $\delta, 2\delta, 3\delta, \ldots$, 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 approximativement $k = t/\delta$ épreuves. On veut laisser δ tendre vers 0. Pour que le résultat ne soit pas trivial, il faut également que p tende vers 0 de façon à ce que p/δ tende vers une constante $\lambda > 0$. Dans ce cas,

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

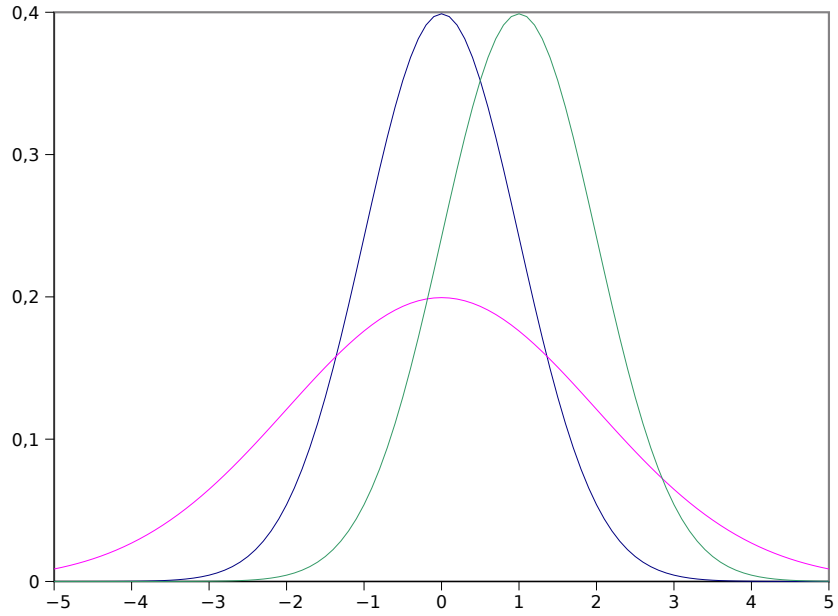


FIGURE 5.3: Densité de probabilité de la loi normale : $\mu = 0, \sigma^2 = 1$ (bleu), $\mu = 0, \sigma^2 = 2$ (magenta) et $\mu = 1, \sigma^2 = 1$ (vert).

Il est aussi aisé de voir (exercice) que la loi exponentielle possède la même propriété de perte de mémoire que la loi géométrique, *cf.* Lemme 2.1.

À nouveau, l'espérance et la variance de $X \sim \exp(\lambda)$ se calculent aisément :

$$\mathbb{E}(X) = \lambda \int_0^\infty s e^{-\lambda s} \, \mathrm{d}s = \int_0^\infty e^{-\lambda s} \, \mathrm{d}s = \lambda^{-1},$$
$$V(X) = \lambda \int_0^\infty s^2 e^{-\lambda s} \, \mathrm{d}s - \lambda^{-2} = \lambda^{-2}.$$

Loi normale

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). X suit une loi **normale** (ou **gaussienne**) de paramètres μ et σ^2, $X \sim \mathcal{N}(\mu, \sigma^2)$, si elle a densité de probabilité

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

pour tout $x \in \mathbb{R}$. Lorsque $\mu = 0$ et $\sigma^2 = 1$, on parle de **loi normale standard**. La fonction de répartition de la loi normale standard est habituellement notée Φ.

Les paramètres μ et σ^2 ont des interprétations immédiates : lorsque $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\mathbb{E}(X) = \tfrac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^\infty s e^{-(s-\mu)^2/2\sigma^2} \, \mathrm{d}s = \mu + \tfrac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^\infty s e^{-s^2/2\sigma^2} \, \mathrm{d}s = \mu,$$
$$\mathrm{Var}(X) = \tfrac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^\infty (s-\mu)^2 e^{-(s-\mu)^2/2} \, \mathrm{d}s = \sigma^2 \tfrac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^\infty e^{-s^2/2\sigma^2} \, \mathrm{d}s = \sigma^2.$$

Loi gamma

X suit la loi **gamma** de paramètres $\lambda, t > 0$, $X \sim \mathrm{gamma}(\lambda, t)$, si elle a densité de probabilité

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

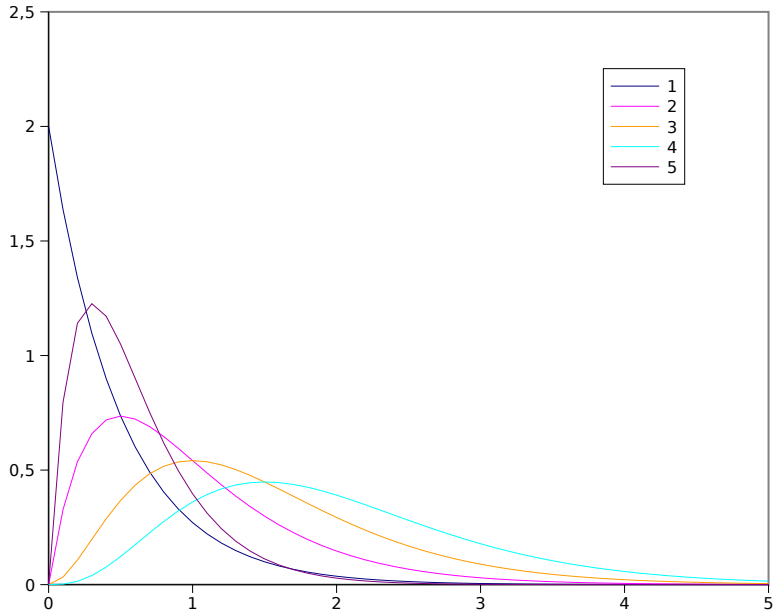


FIGURE 5.4: Densité de probabilité de la loi Gamma pour $\lambda = 0.5$ et diverses valeurs de t.

où Γ est la **fonction gamma**,

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

Lorsque $\lambda = \frac{1}{2}$, et $t = \frac{1}{2}d$, d entier, on dit que X suit la **loi du χ^2 à d degrés de liberté**. Cette distribution joue un rôle important en statistiques.

L'espérance et la variance de $X \sim \text{gamma}(\lambda, t)$ sont données par

$$\mathbb{E}(X) = \frac{\lambda^t}{\Gamma(t)} \int_0^\infty s s^{t-1} e^{-\lambda s} \, \mathrm{d}s = \frac{1}{\Gamma(t)\lambda} \int_0^\infty (\lambda s)^{(t+1)-1} e^{-\lambda s} \, \mathrm{d}s = \frac{\Gamma(t+1)}{\Gamma(t)\lambda} = \frac{t}{\lambda},$$
$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{\Gamma(t+2)}{\Gamma(t)\lambda^2} - \frac{t^2}{\lambda^2} = \frac{t(t+1)}{\lambda^2} - \frac{t^2}{\lambda^2} = \frac{t}{\lambda^2}.$$

Loi de Cauchy

X suit la loi **de Cauchy**[2], $X \sim$ cauchy, si elle a densité de probabilité

$$f_X(x) = \frac{1}{\pi(1+x^2)},$$

pour tout $x \in \mathbb{R}$.

Cette loi a un certain nombre de propriétés « pathologiques », et apparaît souvent dans des contre-exemples. En particulier, elle ne possède pas d'espérance, puisque $\mathbb{E}(|X|) = \infty$ (et donc pas de variance non plus).

Loi bêta

X suit une loi **beta** de paramètres $a, b > 0$, $X \sim \text{beta}(a, b)$, si elle a densité de probabilité

$$f_X(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \mathbf{1}_{[0,1]}(x),$$

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

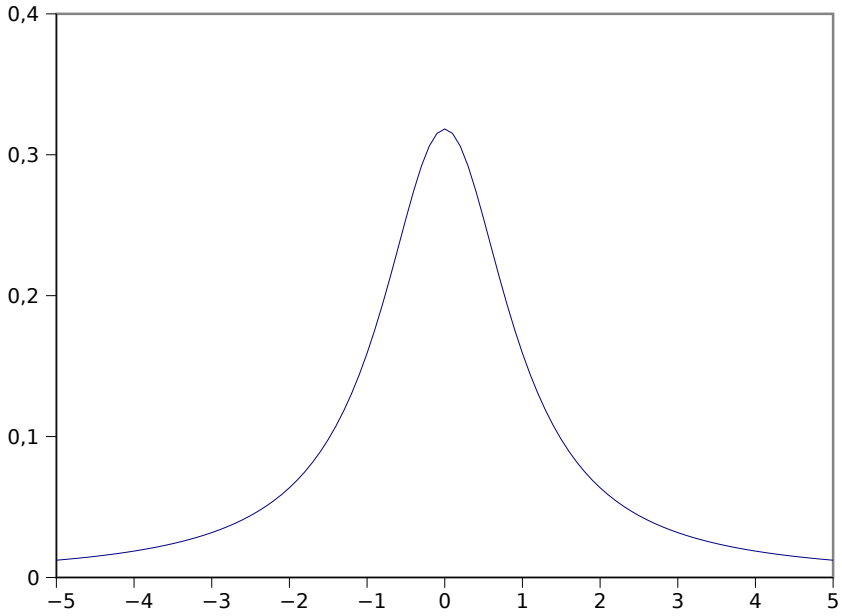


FIGURE 5.5: Densité de probabilité de la loi de Cauchy.

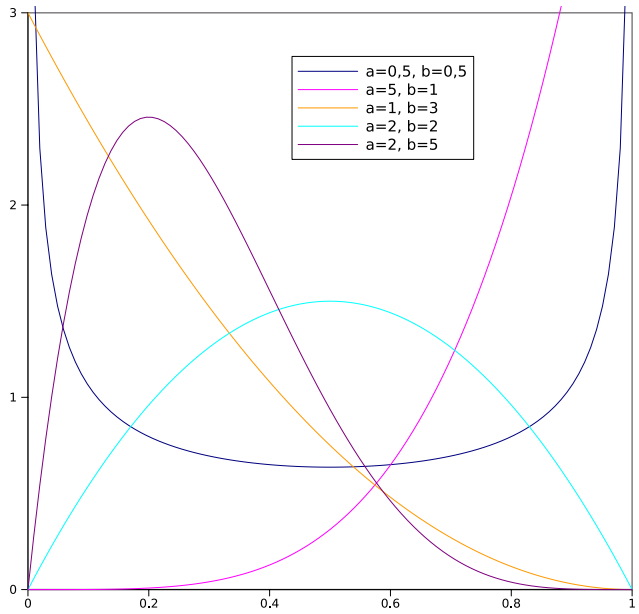


FIGURE 5.6: Densité de probabilité de la loi bêta pour diverses valeurs de a et b.

où $B(a,b)$ est la constante de normalisation. On peut montrer que

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Si $a = b = 1$, X est uniforme sur $[0,1]$.

La distribution bêta est très utilisée en statistiques bayesiennes.

On calcule facilement son espérance et sa variance :

$$\mathbb{E}(X) = \tfrac{1}{B(a,b)} \int_0^1 s s^{a-1}(1-s)^{b-1}\,\mathrm{d}s = \tfrac{1}{B(a,b)} \int_0^1 s^{(a+1)-1}(1-s)^{b-1}\,\mathrm{d}s = \tfrac{B(a+1,b)}{B(a,b)} = \tfrac{a}{a+b},$$
$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \tfrac{B(a+2,b)}{B(a,b)} - \tfrac{a^2}{(a+b)^2} = \tfrac{a(a+1)}{(a+b)(a+b+1)} - \tfrac{a^2}{(a+b)^2} = \tfrac{ab}{(a+b)^2(a+b+1)}.$$

Loi de Student

X suit une loi **de Student**[3] ou **loi** t à $\nu > 0$ degrés de liberté, $X \sim \mathrm{student}(\nu)$, si elle a densité de probabilité

$$f_X(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}(1 + \frac{x^2}{\nu})^{-(\nu+1)/2},$$

pour $x \in \mathbb{R}$.

Cette distribution apparaît dans le problème de l'estimation de la moyenne d'une population normalement distribuée lorsque l'échantillon est petit. C'est la base des célèbres tests de Student en statistiques.

Son espérance est nulle lorsque $\nu > 1$, et n'existe pas pour $\nu \leq 1$. Sa variance n'est pas définie lorsque $\nu \leq 1$, elle est infinie lorsque $1 < \nu \leq 2$, et elle est égale à $\nu/(\nu-2)$ lorsque $\nu > 2$.

Loi de Weibull

X suit une loi **de Weibull**[4] de paramètre de forme $k > 0$ et de paramètre d'échelle $\lambda > 0$ si elle a densité de probabilité

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

Lorsque $k = 1$, on retrouve la distribution exponentielle.

La loi de Weibull est très populaire dans les modèles statistiques en fiabilité. Elle est également utilisée, par exemple, pour analyser les signaux reçus par les radars, ou dans les réseaux de communication sans fil. D'un point de vue plus théorique, elle joue un rôle important dans l'analyse des valeurs extrêmes lors d'expériences aléatoires.

On trouve que son espérance et sa variance sont données par

$$\mathbb{E}(X) = \lambda\Gamma(1+\tfrac{1}{k}),$$
$$\mathrm{Var}(X) = \lambda^2\Gamma(1+\tfrac{2}{k}) - \lambda^2\Gamma(1+\tfrac{1}{k})^2.$$

3. William Sealy Gosset (1876, Canterbury – 1937, Beaconsfield), connu sous le pseudonyme Student, chimiste et statisticien irlandais. Employé de la brasserie Guinness pour stabiliser le goût de la bière, il a ainsi inventé le célèbre test de Student.

4. Ernst Hjalmar Waloddi Weibull (1887, ? ? ? – 1979, Annecy), ingénieur et mathématicien suédois.

5.5.4 Vecteurs aléatoires à densité

La notion de vecteur aléatoire s'étend sans difficulté au cas d'univers généraux.

Définition 5.15. *Un vecteur aléatoire de dimension n est une application mesurable d'un espace probabilisable $(\Omega, \mathcal{F})$ vers l'espace probabilisable $(\mathbb{R}^n, \mathcal{B}^n)$.*

Nous nous intéresserons plus particulièrement au cas des vecteurs aléatoires à densité.

Définition 5.16. *Un vecteur aléatoire $\mathbf{X} = (X_1, \dots, X_n)$ est* ***à densité*** *s'il existe une fonction positive $f_{\mathbf{X}} : \mathbb{R}^n \to \mathbb{R}$ telle que*

$$\mathbb{P}(\mathbf{X} \in A) = \int_A f_{\mathbf{X}}(x_1, \dots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n, \qquad \forall A \in \mathcal{B}(\mathbb{R}^n).$$

$f_{\mathbf{X}}$ est la ***densité de probabilité conjointe*** *du vecteur aléatoire $\mathbf{X}$.*

Remarque 5.15. *On peut montrer qu'il suffit de vérifier la condition pour des ensembles A de la forme $(-\infty, x_1] \times \cdots \times (-\infty, x_n]$, $x_1, \dots, x_n \in \mathbb{R}$, c'est-à-dire que*

$$F_{\mathbf{X}}(x_1, \dots, x_n) = \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \int_{-\infty}^{x_1} \mathrm{d}s_1 \cdots \int_{-\infty}^{x_n} \mathrm{d}s_n\, f_{\mathbf{X}}(s_1, \dots, s_n).$$

La fonction $F_{\mathbf{X}}$ est appelée ***fonction de répartition conjointe*** *de $\mathbf{X}$.*

À nouveau, il n'y a pas unicité de la densité conjointe, et on choisira toujours une version de $f_{\mathbf{X}}$ satisfaisant $f_{\mathbf{X}}(x_1, \dots, x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F_{\mathbf{X}}(x_1, \dots, x_n)$, en chaque point où la fonction de répartition conjointe est suffisamment différentiable.

Les densités de probabilité des composantes d'un vecteur aléatoire $\mathbf{X}$ peuvent être aisément extraites de la densité de probabilité conjointe.

Lemme 5.12. *Soit $\mathbf{X} = (X_1, \dots, X_n)$ un vecteur aléatoire à densité. Alors, pour tout $1 \leq k \leq n$,*

$$f_{X_k}(x_k) = \int_{-\infty}^{\infty} \mathrm{d}x_1 \cdots \int_{-\infty}^{\infty} \mathrm{d}x_{k-1} \int_{-\infty}^{\infty} \mathrm{d}x_{k+1} \cdots \int_{-\infty}^{\infty} \mathrm{d}x_n\, f_{\mathbf{X}}(x_1, \dots, x_n).$$

Démonstration.

$$\begin{aligned}
\mathbb{P}(X_k \in A) &= \mathbb{P}(\mathbf{X} \in \mathbb{R}^{k-1} \times A \times \mathbb{R}^{n-k}) \\
&= \int_{\mathbb{R}^{k-1}} \mathrm{d}x_1 \cdots \mathrm{d}x_{k-1} \int_A \mathrm{d}x_k \int_{\mathbb{R}^{n-k}} \mathrm{d}x_{k+1} \cdots \mathrm{d}x_n f_{\mathbf{X}}(x_1, \dots, x_n) \\
&= \int_A \mathrm{d}x_k \Big\{ \int_{\mathbb{R}^{k-1}} \mathrm{d}x_1 \cdots \mathrm{d}x_{k-1} \int_{\mathbb{R}^{n-k}} \mathrm{d}x_{k+1} \cdots \mathrm{d}x_n f_{\mathbf{X}}(x_1, \dots, x_n) \Big\}
\end{aligned}$$

et une version de f_{X_k} est donc donnée par l'expression entre accolades. □

Définition 5.17. *Étant donné un vecteur aléatoire $\mathbf{X} = (X_1, \dots, X_n)$, les densités de probabilité f_{X_k}, $1 \leq k \leq n$, sont appelées ses* ***densités de probabilité marginales****.*

L'indépendance de variables aléatoires peut se caractériser simplement en termes de leur densité de probabilité conjointe.

Lemme 5.13. *Soit $\mathbf{X} = (X_1, \dots, X_n)$ un vecteur aléatoire à densité. Les variables aléatoires $X_1, \dots, X_n$ sont indépendantes si et seulement si*

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

pour presque tout $(x_1, \dots, x_n)$.

Démonstration. Supposons $X_1, \dots, X_n$ indépendantes. Pour tout $x_1, \dots, x_n \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) &= \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n) \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1}(y_1) \cdots f_{X_n}(y_n)\, \mathrm{d}y_1 \cdots \mathrm{d}y_n,\end{aligned}$$

et par conséquent $f_{X_1}(x_1) \cdots f_{X_n}(x_n)$ est une densité de probabilité conjointe de $\mathbb{P}_{\mathbf{X}}$. □

Exemple 5.8. Soit $\Omega = D_1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$ muni de la densité de probabilité uniforme. On considère les quatre variables aléatoires suivantes : $X(\omega) = x$, $Y(\omega) = y$, $R(\omega) = \sqrt{x^2 + y^2}$ et $\Theta(\omega) \in [0, 2\pi)$ telle que $x = r\cos(\Theta(\omega))$ et $y = r\sin(\Theta(\omega))$. Ainsi les vecteurs aléatoires (X, Y) et (R, Θ) correspondent à la position d'un point du disque tiré uniformément au hasard, exprimée, respectivement, en coordonnées cartésiennes et polaires. Déterminons leurs lois conjointes, ainsi que les lois de ces quatre variables aléatoires.

Pour le couple (X, Y), on a

$$\mathbb{P}((X, Y) \in A) = |A \cap D_1|/\pi = \iint_A \frac{1}{\pi} \mathbf{1}_{\{x^2+y^2<1\}} \mathrm{d}x\mathrm{d}y,$$

et donc $f_{X,Y}(x, y) = \frac{1}{\pi}\mathbf{1}_{\{x^2+y^2<1\}}$. La loi de X est obtenue en prenant la marginale correspondante,

$$f_X(x) = \int_{-1}^{1} \frac{1}{\pi} \mathbf{1}_{\{x^2+y^2<1\}} \mathrm{d}y = \frac{1}{\pi} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \mathrm{d}y = \frac{2}{\pi}\sqrt{1 - x^2},$$

pour $-1 < x < 1$ et 0 sinon. De la même façon, $f_Y(y) = \frac{2}{\pi}\sqrt{1 - y^2}\mathbf{1}_{\{y^2<1\}}$. En particulier, on voit que $f_{(X,Y)}(x, y) \neq f_X(x) f_Y(y)$, et donc X et Y ne sont pas indépendantes.

Passons au couple (R, Θ). Étant donné $A \subset \mathbb{R}^2$, notons $\tilde{A} = \{(x, y) : (R(x, y), \Theta(x, y)) \in A\}$. Alors,

$$\mathbb{P}((R, \Theta) \in A) = |\tilde{A} \cap D_1|/\pi = \iint_A \frac{1}{\pi} \mathbf{1}_{\{0\le r<1, 0\le\theta<2\pi\}} r\mathrm{d}r\mathrm{d}\theta,$$

d'où l'on tire la densité de probabilité conjointe $f_{R,\Theta}(r, \theta) = \frac{r}{\pi}\mathbf{1}_{\{0\le r<1, 0\le\theta<2\pi\}}$. La densité de R est donc donnée par

$$f_R(r) = \frac{r}{\pi} \int_0^{2\pi} \mathrm{d}\theta = 2r,$$

si $0 \le r < 1$ et 0 sinon. Pour Θ,

$$f_\Theta(\theta) = \frac{1}{\pi} \int_0^1 r\mathrm{d}r = \frac{1}{2\pi},$$

si $0 \le \theta < 2\pi$ et 0 sinon. On a donc $f_{(R,\Theta)}(r, \theta) = f_R(r) f_\Theta(\theta)$, et R et Θ sont indépendantes. ♢

Finalement, si $\mathbf{X} = (X_1, \dots, X_n)$ est un vecteur aléatoire à densité, et $\Psi : \mathbb{R}^n \to \mathbb{R}^n$ possède de bonnes propriétés, le théorème suivant permet de déterminer la loi conjointe du vecteur aléatoire $\Psi(\mathbf{X})$ en termes de $f_{\mathbf{X}}$.

Soient $U \subseteq \mathbb{R}^n$ un ouvert, et $\Psi : U \to \mathbb{R}^n$, $\Psi(\mathbf{x}) = (\psi_1(\mathbf{x}), \dots, \psi_n(\mathbf{x}))$. On dit que Ψ est continuement différentiable si les dérivées partielles $\partial\psi_i/\partial x_j$ existent et sont continues sur U. On note $D_\Psi(\mathbf{x}) = (\partial\psi_i(\mathbf{x})/\partial x_j)_{1\le i,j\le n}$ la matrice Jacobienne, $J_\Psi(\mathbf{x}) = \det D_\Psi(\mathbf{x})$ le Jacobien, et $V = \Psi(U)$.

Théorème 5.8. *Soient $U \subseteq \mathbb{R}^n$ un ouvert, et $\Psi : U \to V$ une application continuement différentiable et bijective, telle que $J_\Psi(\mathbf{x}) \neq 0$, pour tout $\mathbf{x} \in U$. Alors, pour toute fonction $f : V \to \mathbb{R}$, $f \in \mathcal{L}^1$, on a*

$$\int_U f(\Psi(\mathbf{x}))\, |J_\Psi(\mathbf{x})|\, \mathrm{d}x_1 \cdots \mathrm{d}x_n = \int_V f(\mathbf{y})\, \mathrm{d}y_1 \cdots \mathrm{d}y_n.$$

Démonstration. Dans le cas où f est suffisamment régulière, il s'agit simplement du résultat classique sur les changements de variables. La preuve lorsque $f \in \mathcal{L}^1$ sera faite en Analyse III. □

Corollaire 5.2. *On considère un vecteur aléatoire $\mathbf{X} = (X_1, \dots, X_n)$ à valeurs dans un ouvert $U \subseteq \mathbb{R}^n$, et une application $\Psi : \mathbb{R}^n \to \mathbb{R}^n$ comme dans le théorème précédent. Alors la densité de probabilité conjointe du vecteur aléatoire $\mathbf{Y} = \Psi(\mathbf{X})$ est donnée par*

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\Psi^{-1}(\mathbf{y}))\,|J_{\Psi^{-1}}(\mathbf{y})|.$$

Démonstration. Soit $A \subseteq V$. On a

$$\mathbb{P}(\mathbf{Y} \in A) = \mathbb{P}(\Psi(\mathbf{X}) \in A) = \mathbb{P}(\mathbf{X} \in \Psi^{-1}(A)) = \int_{\Psi^{-1}(A)} f_{\mathbf{X}}(\mathbf{x})\,\mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

Une application du théorème à l'intégrale du membre de droite (attention, on l'applique à la transformation inverse Ψ^{-1}) donne donc

$$\mathbb{P}(\mathbf{Y} \in A) = \int_A f_{\mathbf{X}}(\Psi^{-1}(\mathbf{y}))\,|J_{\Psi^{-1}}(\mathbf{y})|\,\mathrm{d}y_1 \cdots \mathrm{d}y_n,$$

d'où le résultat suit. □

On en déduit immédiatement le résultat suivant, très important, sur la loi d'une somme de variables aléatoires.

Lemme 5.14. *Soient X, Y deux variables aléatoires à densité. Alors la loi de leur somme est donnée par*

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f_{(X,Y)}(x, u-x)\,\mathrm{d}x.$$

En particulier, si X et Y sont indépendantes, la densité de probabilité de $X+Y$ est donnée par la convolution des densités de probabilité de X et Y,

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f_X(x) f_Y(u-x)\,\mathrm{d}x.$$

Démonstration. On considère l'application $\Psi : \mathbb{R}^2 \to \mathbb{R}^2$ donnée par $\Psi(x, y) = (x, x+y)$. Elle satisfait à toutes les hypothèses du Corollaire précédent. On a donc

$$f_{(X,X+Y)}(u, v) = f_{(X,Y)}(u, v-u),$$

puisque le Jacobien vaut 1. Par conséquent la première affirmation suit en prenant la seconde marginale,

$$f_{X+Y}(v) = \int_{-\infty}^{\infty} f_{(X,Y)}(u, v-u)\,\mathrm{d}u.$$

Si X et Y sont indépendantes, leur densité de probabilité conjointe se factorise et la seconde affirmation suit. □

Une autre conséquence utile (et immédiate) du Corollaire précédent est le résultat suivant.

Lemme 5.15. *Soit X une variable aléatoire à densité et $a, b \in \mathbb{R}$, $a \neq 0$. La densité de probabilité de la variable aléatoire $aX + b$ est donnée par*

$$f_{aX+b}(y) = \frac{1}{|a|} f_X\left((y-b)/a\right).$$

Démonstration. Laissée en exercice. □

On déduit immédiatement des deux lemmes précédents l'important résultat suivant.

Lemme 5.16. *Soient X_1 et X_2 deux variables aléatoires indépendantes de loi $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$ respectivement. La variable aléatoire $X_1 + X_2$ suit une loi $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

Démonstration. Soient $Y_1 = X_1 - \mu_1$ et $Y_2 = X_2 - \mu_2$; par le lemme 5.15, ces variables suivent respectivement les lois $\mathcal{N}(0, \sigma_1^2)$ et $\mathcal{N}(0, \sigma_2^2)$. Une application du Lemme 5.14 montre que la densité de probabilité de la variable aléatoire $Y_1 + Y_2$ est donnée par

$$\frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \int_{\mathbb{R}} \exp\{-\frac{x^2}{2\sigma_1^2} - \frac{(z-x)^2}{2\sigma_2^2}\}\, \mathrm{d}x.$$

Puisque

$$\sigma_2^2 x^2 + \sigma_1^2 (z-x)^2 = (\sqrt{\sigma_1^2 + \sigma_2^2}\, x - \frac{\sigma_1^2 z}{\sqrt{\sigma_1^2 + \sigma_2^2}})^2 + \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} z^2,$$

l'intégration sur x montre que cette densité de probabilité est bien celle d'une variable aléatoire de loi $\mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$, et donc $X_1 + X_2$ suit bien une loi $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. □

Vecteurs aléatoires gaussiens

Nous allons voir à présent un exemple particulièrement important de vecteur aléatoire. Si $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, on note leur produit scalaire $\langle \mathbf{x}, \mathbf{y} \rangle$.

Définition 5.18. *Un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$ est un* ***vecteur aléatoire gaussien*** *si les variables aléatoires $\langle \mathbf{a}, \mathbf{X} \rangle$ suivent des lois normales, pour tout $\mathbf{a} \in \mathbb{R}^n$.*

Lemme 5.17. *Les propriétés suivantes sont vérifiées pour tout vecteur gaussien $\mathbf{X} = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$.*

1. *X_i est une variable aléatoire gaussienne pour chaque $i = 1, \ldots, n$.*
2. *Si $A : \mathbb{R}^n \to \mathbb{R}^n$ est une application linéaire, le vecteur $A\mathbf{X} : \Omega \to \mathbb{R}^n$ est un vecteur gaussien.*

Démonstration. La première affirmation suit en prenant $\mathbf{a} = \mathbf{e}_i$ dans la Définition 5.18. Pour la seconde affirmation, il suffit d'observer que, pour tout $\mathbf{a} \in \mathbb{R}^n$,

$$\langle \mathbf{a}, A\mathbf{X} \rangle = \langle {}^{\mathrm{t}}\mathbf{A}\mathbf{a}, \mathbf{X} \rangle$$

est bien gaussien. □

Remarque 5.16. *La réciproque de la première affirmation est fausse : un vecteur aléatoire dont chaque composante est gaussienne n'est pas nécessairement gaussien. Nous le verrons sur un exemple plus tard (Exemple 6.1).*

Exemple 5.9. Un exemple de vecteur aléatoire gaussien est le vecteur $(X_1, \ldots, X_n)$ composé de n variables aléatoires *indépendantes* suivant des lois normales. En effet, $a_1 X_1 + \cdots + a_n X_n$ est une somme de variables aléatoires normales, et donc, par le Lemme 5.16, suit également une loi normale. ♢

Il suit de l'exemple précédent et du Lemme 5.17 que l'image d'un vecteur $(X_1, \ldots, X_n)$ composé de n variables aléatoires *indépendantes* suivant des lois normales sous l'action d'une transformation linéaire A est également un vecteur gaussien. En particulier, on obtient la classe suivante de vecteur aléatoires gaussiens.

Lemme 5.18. *Soient $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n) \in \mathbb{R}^n$ et $\mathbf{C} = (C_{ij})$ une matrice $n \times n$ symétrique définie positive. Le vecteur aléatoire $\mathbf{X} = (X_1, \dots, X_n)$ de densité de probabilité conjointe*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{C}}} \exp\left(-\tfrac{1}{2}\langle \mathbf{x} - \boldsymbol{\mu}, \mathbf{C}^{-1}(\mathbf{x} - \boldsymbol{\mu})\rangle\right), \tag{5.2}$$

est un vecteur gaussien. On dira qu'un tel vecteur suit une loi $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$.

Démonstration. Puisque $\mathbf{C}^{-1}$ est symétrique, on peut trouver une matrice orthogonale $\mathbf{O}$ et une matrice diagonale $\mathbf{D}$ telles que $\mathbf{C}^{-1} = {}^{\mathrm{t}}\mathbf{O}\mathbf{D}\mathbf{O}$. Par conséquent, en posant $\mathbf{Y} = \mathbf{O}(\mathbf{X} - \boldsymbol{\mu})$, on voit que les variables aléatoires $Y_1, \dots, Y_n$ sont indépendantes, et suivent des lois normales. Par l'exemple précédent, le vecteur $\mathbf{Y}$ est donc gaussien. Par conséquent, il suit du point *2.* du Lemme 5.17 que ${}^{\mathrm{t}}\mathbf{O}\mathbf{Y}$, et donc également le vecteur est gaussien, et donc également le vecteur $\mathbf{X} = {}^{\mathrm{t}}\mathbf{O}\mathbf{Y} + \boldsymbol{\mu}$. □

5.5.5 Loi conditionnelle et espérance conditionnelle

Soient X et Y deux variables aléatoires possédant la densité de probabilité conjointe $f_{(X,Y)}$. On aimerait donner un sens à la loi conditionnelle de Y sachant que X prend la valeur x. Le problème est que la probabilité $\mathbb{P}(Y \leq y \,|\, X = x)$ n'est pas définie puisque l'événement $\{X = x\}$ a probabilité nulle. Afin de déterminer la généralisation appropriée, nous pouvons procéder comme suit. Soit x tel que $f_X(x) > 0$; alors, pour $\epsilon > 0$ petit,

$$\begin{aligned}
\mathbb{P}(Y \leq y \,|\, x \leq X \leq x + \epsilon) &= \frac{\mathbb{P}(Y \leq y, x \leq X \leq x + \epsilon)}{\mathbb{P}(x \leq X \leq x + \epsilon)} \\
&\simeq \frac{\epsilon \int_{-\infty}^{y} f_{(X,Y)}(x, v)\mathrm{d}v}{f_X(x)\,\epsilon} \\
&= \int_{-\infty}^{y} \frac{f_{(X,Y)}(x, v)}{f_X(x)}\,\mathrm{d}v.
\end{aligned}$$

En laissant $\epsilon \downarrow 0$, le membre de gauche converge vers ce que l'on aimerait définir comme $\mathbb{P}(Y \leq y \,|\, X = x)$, et le membre de droite conduit donc à la définition suivante.

Définition 5.19. *Soient X, Y deux variables aléatoires avec densité de probabilité conjointe $f_{(X,Y)}$. La* ***densité de probabilité conditionnelle*** *de Y sachant que $X = x$ est définie par*

$$f_{Y|X}(y \,|\, x) = \frac{f_{(X,Y)}(x, y)}{f_X(x)},$$

pour tout x tel que $f_X(x) > 0$. La loi correspondante s'appelle la ***loi conditionnelle*** *de Y sachant que $X = x$.*

Remarque 5.17. *Soient X_1 et X_2 deux variables aléatoires indépendantes de loi $\exp(\lambda)$. Quelle est la densité de probabilité conditionnelle de $X_1 + X_2$ étant donné que $X_1 = X_2$?*

Première solution : *Soit $Y_1 = X_1 + X_2$ et $Y_2 = X_1/X_2$. Manifestement, $X_1 = X_2$ si et seulement si $Y_2 = 1$. On vérifie facilement (exercice) que la densité de probabilité conditionnelle de Y_1 étant donné que $Y_2 = 1$ est donnée par*

$$f_{Y_1|Y_2}(y_1 \,|\, 1) = \lambda^2 y_1 e^{-\lambda y_1}, \qquad y_1 \geq 0.$$

Deuxième solution : *Soit $Y_1 = X_1 + X_2$ et $Y_3 = X_1 - X_2$. Manifestement, $X_1 = X_2$ si et seulement si $Y_3 = 0$. On vérifie facilement (exercice) que la densité de probabilité conditionnelle de Y_1 étant donné que $Y_3 = 0$ est donnée par*

$$f_{Y_1|Y_3}(y_1 \,|\, 0) = \lambda e^{-\lambda y_1}, \qquad y_1 \geq 0.$$

Il y a clairement un problème : les deux réponses obtenues sont différentes ! L'erreur trouve sa source dans la question elle-même : qu'entend-on par la condition $X_1 = X_2$? Ce dernier est un événement de probabilité nulle, et il est crucial de décrire précisément de quelle suite d'événements de probabilité positive il est la limite. Dans la première solution, on interprète essentiellement cet événement comme $\{X_1 \leq X_2 \leq (1+\epsilon)X_1\}$ (ϵ petit), alors que dans la seconde, on l'interprète comme $\{X_1 \leq X_2 \leq X_1 + \epsilon\}$. Il convient donc de déterminer au préalable quelle est l'interprétation désirée, et cela dépend du problème considéré.

Étant en possession d'une notion de loi conditionnelle, on peut, comme dans le cas discret, définir l'espérance conditionnelle, comme étant l'espérance sous la loi conditionnelle.

Définition 5.20. *Soient X et Y deux variables aléatoires de densité de probabilité conjointe $f_{(X,Y)}$. L'**espérance conditionnelle** de Y sachant X est la variable aléatoire*

$$\mathbb{E}(Y \mid X)(\cdot) \equiv \mathbb{E}(Y \mid X = \cdot) = \int_{\mathbb{R}} y\, f_{Y|X}(y \mid \cdot)\, \mathrm{d}y,$$

pourvu que $\int_{\mathbb{R}} |y|\, f_{Y|X}(y \mid \cdot)\, \mathrm{d}y < \infty$.

On a le même résultat que dans le cas discret.

Lemme 5.19. *Soient X et Y deux variables aléatoires de densité de probabilité conjointe $f_{(X,Y)}$. Pour toute fonction mesurable φ telle que les espérances existent,*

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

Démonstration. La preuve est formellement identique à celle du cas discret :

$$\begin{aligned}\mathbb{E}(\mathbb{E}(Y \mid X)\varphi(X)) &= \int_{\mathbb{R}}\int_{\mathbb{R}} y f_{Y\mid X}(y \mid x)\mathrm{d}y\, \varphi(x) f_X(x)\mathrm{d}x \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} y\varphi(x) f_{(X,Y)}(x,y)\mathrm{d}x\mathrm{d}y = \mathbb{E}(Y\varphi(X)).\end{aligned}$$

□

5.6 Processus en temps discret.

On considère une succession d'expériences, pas forcément identiques, ni forcément indépendantes. On suppose que l'on connaisse l'espace de probabilité décrivant la première expérience, ainsi que l'espace de probabilité décrivant la $k^{\text{ème}}$ expérience lorsque les résultats des $k-1$ expériences précédentes sont connus. Le théorème suivant montre comment construire un espace de probabilité décrivant une telle situation, dans le cas où les expériences sont décrites par des espaces de probabilités discrets.

Théorème 5.9. *Soit $n \geq 2$. Pour chaque $i \in \{1, \dots, n\}$, soit $\Omega_i \neq \varnothing$ un univers au plus dénombrable. Soit ρ_1 une mesure de probabilité sur Ω_1 et, pour $k = 2, \dots, n$, et pour tout $\omega_i \in \Omega_i$, $i < k$, soit $\rho_k|\omega_1, \dots, \omega_{k-1}$ une mesure de probabilité sur Ω_k. Notons $\Omega = \Omega_1 \times \cdots \times \Omega_n$ et $X_i : \Omega \to \Omega_i$ la projection sur la $i$ème composante. Alors, il existe une unique mesure de probabilité $\mathbb{P}$ sur $(\Omega, \mathscr{P}(\Omega))$ satisfaisant*

- $\mathbb{P}(X_1 = \omega_1) = \rho_1(\omega_1)$, *pour tout* $\omega_1 \in \Omega_1$;
- *pour tout* $k = 2, \dots, n$,

$$\mathbb{P}(X_k = \omega_k \mid X_1 = \omega_1, \dots, X_{k-1} = \omega_{k-1}) = \rho_k|\omega_1, \dots, \omega_{k-1}(\omega_k),$$

pour tout $\omega_i \in \Omega_i$ tels que $\mathbb{P}(X_1 = \omega_1, \dots, X_{k-1} = \omega_{k-1}) > 0$.

La mesure de probabilité $\mathbb{P}$ est explicitement donnée par

$$\mathbb{P}(\{\omega\}) = \rho_1(\omega_1)\rho_{2|\omega_1}(\omega_2)\rho_{3|\omega_1,\omega_2}(\omega_3)\cdots\rho_n|\omega_1, \dots, \omega_{n-1}(\omega_n), \tag{5.3}$$

pour tout $\omega = (\omega_1, \dots, \omega_n) \in \Omega$.

Démonstration. Le fait que $\mathbb{P}$ prenne nécessairement la forme (5.3) est une conséquence immédiate de la définition des probabilités conditionnelles. Cela prouve évidemment l'unicité de $\mathbb{P}$.

Vérifions donc la validité des conditions pour cette définition de $\mathbb{P}$. On a

$$\begin{aligned}\mathbb{P}(X_1 = \omega_1, \dots, X_k = \omega_k) &= \sum_{\omega_{k+1} \in \Omega_{k+1}, \dots, \omega_n \in \Omega_n} \mathbb{P}(\{\omega_1, \dots, \omega_n\}) \\ &= \rho_1(\omega_1)\cdots\rho_{k|\omega_1,\dots,\omega_{k-1}}(\omega_k) \sum_{\omega_{k+1}\in\Omega_{k+1}} \rho_{k+1|\omega_1,\dots,\omega_k}(\omega_{k+1})\cdots \sum_{\omega_n\in\Omega_n} \rho_n|\omega_1, \dots, \omega_{n-1}(\omega_n).\end{aligned}$$

À présent, pour chaque j,

$$\sum_{\omega_j \in \Omega_j} \rho_{j|\omega_1,\dots,\omega_{j-1}}(\omega_j) = 1,$$

puisque $\rho_{j|\omega_1,\dots,\omega_{j-1}}$ est une mesure de probabilité. On obtient donc

$$\mathbb{P}(X_1 = \omega_1, \dots, X_k = \omega_k) = \rho_1(\omega_1)\cdots\rho_{k|\omega_1,\dots,\omega_{k-1}}(\omega_k).$$

Pour $k = 1$, ceci se réduit donc bien à $\mathbb{P}(X_1 = \omega_1) = \rho_1(\omega_1)$. Pour $k > 1$, on a donc bien

$$\mathbb{P}(X_k = \omega_k \mid X_1 = \omega_1, \dots, X_{k-1} = \omega_{k-1}) = \frac{\mathbb{P}(X_1 = \omega_1, \dots, X_k = \omega_k)}{\mathbb{P}(X_1 = \omega_1, \dots, X_{k-1} = \omega_{k-1})} = \rho_{k|\omega_1,\dots,\omega_{k-1}}(\omega_k).$$

□

On aimerait à présent étendre cette construction au cas d'une infinité d'expériences.

Théorème 5.10. *Pour chaque $i \in \mathbb{N}$, on se donne un ensemble au plus dénombrable $\Omega_i \neq \varnothing$. Soit ρ_1 une mesure de probabilité sur Ω_1 et, pour chaque $k \in \{2, \dots, n\}$ et tout $\omega_i \in \Omega_i$, $i < k$, soit $\rho_k|\omega_1, \dots, \omega_{k-1}$ une mesure de probabilité sur Ω_k. Notons $\Omega = \times_{n\geq 1}\Omega_n$, $X_i : \Omega \to \Omega_i$ la projection sur la $i$ème composante, et $\mathcal{F} = \bigotimes_{n\geq 1}\mathscr{P}(\Omega_n)$ la tribu-produit sur Ω. Alors, il existe une unique mesure de probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{F})$ satisfaisant*

- $\mathbb{P}(X_1 = \omega_1) = \rho_1(\omega_1)$, *pour tout* $\omega_1 \in \Omega_1$;
- *pour tout* $k \geq 2$,

$$\mathbb{P}(X_k = \omega_k \mid X_1 = \omega_1, \dots, X_{k-1} = \omega_{k-1}) = \rho_k|\omega_1, \dots, \omega_{k-1}(\omega_k),$$

pour tout $\omega_i \in \Omega_i$ tels que $\mathbb{P}(X_1 = \omega_1, \dots, X_{k-1} = \omega_{k-1}) > 0$.

Démonstration. L'unicité suit du Théorème d'unicité 5.1 puisque la collection

$$\mathcal{G} = \{\{X_1 = \omega_1, \ldots, X_k = \omega_k\} \,:\, k \geq 1, \omega_i \in \Omega_i, i = 1, \ldots, k\} \cup \{\varnothing\}$$

forme un π-système générant $\mathcal{F}$.

Nous allons montrer que l'existence de $\mathbb{P}$ découle de celle de la mesure de Lebesgue sur $[0,1]$. On commence par partitionner l'intervalle $[0,1)$ en $|\Omega_1|$ intervalles disjoints $(I_{\omega_1})_{\omega_1 \in \Omega_1}$, la longueur de I_{ω_1} étant égale à $\rho_1(\omega_1)$. Ce sont les intervalles de premier niveau.

On partitionne ensuite chaque intervalle de premier niveau I_{ω_1} en $|\Omega_2|$ intervalles disjoints de second niveau, $(I_{\omega_1,\omega_2})_{\omega_2 \in \Omega_2}$, avec I_{ω_1,ω_2} de longueur $\rho_1(\omega_1)\rho_{2|\omega_1}(\omega_2)$.

On poursuit de la même façon, partitionnant chaque intervalle $I_{\omega_1,\ldots,\omega_k}$ de niveau k en $|\Omega_{k+1}|$ intervalles disjoints $(I_{\omega_1,\ldots,\omega_{k+1}})_{\omega_{k+1} \in \Omega_{k+1}}$, de niveau $k+1$, tels que la longueur de $I_{\omega_1,\ldots,\omega_{k+1}}$ soit $\rho_1(\omega_1)\rho_{2|\omega_1}(\omega_2)\cdots\rho_{k+1|\omega_1,\ldots,\omega_k}(\omega_{k+1})$.

De cette manière, pour chaque $x \in [0,1)$ et chaque $k \geq 1$, il existe un unique intervalle de niveau k contenant x. En d'autres termes, il existe une unique suite $Z(x) = (Z_1(x), Z_2(x), \ldots) \in \Omega$ telle que $x \in I_{Z_1(x),\ldots,Z_k(x)}$ pour tout $k \geq 1$. L'application $Z : [0,1) \to \Omega$ est une variable aléatoire, puisque, pour tout $A = \{X_1 = \omega_1, \ldots, X_k = \omega_k\} \in \mathcal{G}$,

$$Z^{-1}(A) = \{x \,:\, Z_1(x) = \omega_1, \ldots, Z_k(x) = \omega_k\} = I_{\omega_1,\ldots,\omega_k} \in \mathcal{B}([0,1)),$$

et le Lemme 5.6 s'applique. Il suit donc du Théorème 5.3 que $\mathbb{P} = \lambda \circ Z^{-1}$ est une mesure de probabilité sur $(\Omega, \mathcal{F})$. Par construction, elle possède les propriétés désirées. □

Chapitre 6

Fonctions caractéristiques

Dans ce chapitre, nous allons très brièvement 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 arbitraires.

6.1 Définition et propriétés élémentaires

Définition 6.1. *La* ***fonction caractéristique*** *associée à une variable aléatoire X est la fonction $\phi_X : \mathbb{R} \to \mathbb{C}$ définie par*

$$\phi_X(t) = \mathbb{E}(e^{itX}).$$

Remarque 6.1. *Nous avons principalement travaillé avec des fonctions réelles jusqu'à présent. Toutefois tout ce qui a été dit reste vrai dans le cas complexe : il suffit de décomposer l'intégrant en sa partie réelle et sa partie imaginaire,*

$$\phi_X(t) = \mathbb{E}(\cos(tX)) + i\mathbb{E}(\sin(tX)).$$

Théorème 6.1. *ϕ est une fonction caractéristique si et seulement si elle possède les propriétés suivantes.*

1. *$\phi(0) = 1$, et $|\phi(t)| \leq 1$ pour tout t.*
2. *ϕ est uniformément continue sur $\mathbb{R}$.*
3. *ϕ est définie positive, c'est-à-dire*

$$\sum_{j,k} \phi(t_j - t_k) z_j \bar{z}_k \geq 0,$$

pour tout $t_1, \ldots, t_n$ réels, et tout $z_1, \ldots, z_n$ complexes.

Démonstration. Soit ϕ une fonction caractéristique. Alors $\phi(0) = \mathbb{E}(1) = 1$, et $|\phi(t)| \leq \mathbb{E}(|e^{itX}|) = 1$.

On a également

$$|\phi(t+s) - \phi(t)| = |\mathbb{E}(e^{i(t+s)X} - e^{itX})| \leq \mathbb{E}(|e^{itX}(e^{isX} - 1)|) = \mathbb{E}(|e^{isX} - 1|).$$

Soit $Y(s) = |e^{isX} - 1|$; manifestement $0 \leq Y \leq 2$ et $\lim_{s\to 0} Y(s) = 0$. Par conséquent, le Théorème de convergence dominée implique que $\lim_{s\to 0} \mathbb{E}(Y(s)) = 0$, et la continuité uniforme est établie.

Pour la positivité, il suffit d'observer que

$$\sum_{j,k} \phi(t_j - t_k) z_j \bar{z}_k = \mathbb{E}\Big(\sum_{j,k} z_j e^{\mathrm{i}t_j X}\, \bar{z}_k e^{-\mathrm{i}t_k X}\Big) = \mathbb{E}\Big(\big|\sum_j z_j e^{\mathrm{i}t_j X}\big|^2\Big) \geq 0.$$

Nous ne démontrerons pas la réciproque (Théorème de Bochner) ici. □

La fonction caractéristique permet de calculer les moments de la variable aléatoire associée.

Lemme 6.1. *Si $\mathbb{E}(|X|^k) < \infty$, alors*

$$\phi_X(t) = \sum_{j=0}^{k} \frac{\mathbb{E}(X^j)}{j!} (\mathrm{i}t)^j + o(t^k),$$

lorsque $t \to 0$. En particulier, $\phi_X^{(k)}(0) = \mathrm{i}^k \mathbb{E}(X^k)$.

Démonstration. Le théorème de Taylor (avec reste) implique que, pour tout $x \in \mathbb{R}$,

$$\cos(x) + \mathrm{i}\sin(x) = \sum_{\ell=0}^{k-1} \frac{(\mathrm{i}x)^\ell}{\ell!} + \frac{(\mathrm{i}x)^k}{k!}(\cos(\alpha_1 x) + \mathrm{i}\sin(\alpha_2 x)),$$

où $\alpha_1, \alpha_2 \in [-1, 1]$. Il suit que

$$e^{\mathrm{i}tX} = \sum_{\ell=0}^{k} \frac{(\mathrm{i}tX)^\ell}{\ell!} + \frac{(\mathrm{i}tX)^k}{k!}(\cos(A_1 tX) + \mathrm{i}\sin(A_2 tX) - 1),$$

où A_1, A_2 sont deux variables aléatoires telles que $|A_1|, |A_2| \leq 1$. En particulier,

$$|X^k(\cos(A_1 tX) + \mathrm{i}\sin(A_2 tX) - 1)| \leq 3|X|^k,$$

et le théorème de convergence dominée implique que

$$\lim_{t\to 0} \mathbb{E}(|X^k(\cos(A_1 tX) + \mathrm{i}\sin(A_2 tX) - 1)|) = 0.$$

□

Remarque 6.2. *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.*

Un des nombreux intérêts des fonctions caractéristiques est qu'elles fournissent un outil très efficace pour étudier les sommes de variables aléatoires indépendantes.

Proposition 6.1. *Soient X et Y deux variables aléatoires indépendantes. Alors*

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t).$$

Démonstration.

$$\phi_{X+Y}(t) = \mathbb{E}(e^{\mathrm{i}tX} e^{\mathrm{i}tY}) = \mathbb{E}(e^{\mathrm{i}tX})\mathbb{E}(e^{\mathrm{i}tY}) = \phi_X(t)\phi_Y(t).$$

La seconde identité suit de l'indépendance, après avoir décomposé chacune des exponentielles en sinus et cosinus, effectué la multiplication, et regroupé les termes. □

Le résultat suivant est également très utile.

Lemme 6.2. *Si $a, b \in \mathbb{R}$ et $Y = aX + b$, alors*

$$\phi_Y(t) = e^{itb}\phi_X(at).$$

Démonstration.

$$\phi_Y(t) = \mathbb{E}(e^{it(aX+b)}) = e^{itb}\mathbb{E}(e^{i(at)X}) = e^{itb}\phi_X(at).$$

□

On peut également définir une notion de fonction caractéristique conjointe pour une famille de variables aléatoires.

Définition 6.2. *La* ***fonction caractéristique conjointe*** *du vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ est définie par*

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}(e^{i\langle \mathbf{t}, \mathbf{X}\rangle}),$$

pour tout $\mathbf{t} = (t_1, \ldots, t_n) \in \mathbb{R}^n$.

Il est utile d'observer que $\phi_{(X_1,\ldots,X_n)}(t_1, \ldots, t_n) = \phi_{t_1X_1+\cdots+t_nX_n}(1)$.

La fonction caractéristique conjointe fournit une nouvelle caractérisation de l'indépendance.

Théorème 6.2. *Les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes si et seulement si*

$$\phi_{(X_1,\ldots,X_n)}(t_1, \ldots, t_n) = \prod_{j=1}^{n} \phi_{X_j}(t_j).$$

Démonstration. Si $X_1, \ldots, X_n$ sont indépendantes, alors le résultat suit de la Proposition 6.1. La réciproque suit (de la version à n variables) du Théorème d'inversion énoncé ci-dessous. □

6.2 Théorèmes d'inversion et de continuité

Le résultat fondamental suivant montre qu'une variable aléatoire est complètement caractérisée par sa fonction caractéristique : deux variables aléatoires possédant la même fonction caractéristique ont la même loi.

Théorème 6.3 (Théorème d'inversion)**.** *Soit X une variable aléatoire de fonction de répartition F_X et de fonction caractéristique ϕ_X. Alors,*

$$F_X(b) - F_X(a) = \lim_{T\to\infty} \int_{-T}^{T} \frac{e^{-iat} - e^{-ibt}}{2i\pi t}\phi_X(t)\,dt.$$

en chaque point de continuité de F_X.

Démonstration. On écrit simplement F et ϕ. On a

$$\begin{aligned}
\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} dt\, \frac{e^{-iat} - e^{-ibt}}{it} \int e^{itx}\, d\mathbb{P}_X & \\
&= \lim_{T\to\infty} \frac{1}{2\pi} \int d\mathbb{P}_X \int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{it}\, dt \\
&= \lim_{T\to\infty} \frac{1}{2\pi} \int d\mathbb{P}_X \int_{-T}^{T} \frac{\sin\left(t(x-a)\right) - \sin\left(t(x-b)\right)}{t}\, dt \\
&= \tfrac{1}{2} \int \{\text{signe}(x-a) - \text{signe}(x-b)\}\, d\mathbb{P}_X \\
&= F(b) - F(a),
\end{aligned}$$

pourvu que a et b soient des points de continuité de F. On a utilisé le Théorème de Fubini, et le Théorème de la convergence dominée pour prendre la limite $T \to \infty$. En effet, la fonction de Dirichlet

$$u(T, z) = \int_0^T \frac{\sin tz}{t} \, \mathrm{d}t$$

satisfait $\sup_{T,z} |u(T, z)| \leq C$ et [1]

$$\lim_{T\to\infty} u(T, z) = \begin{cases} \pi/2 & \text{si } z > 0 \\ -\pi/2 & \text{si } z < 0 \\ 0 & \text{si } z = 0. \end{cases}$$

□

Corollaire 6.1. *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, et donc $F_X = F_Y$, par continuité à droite des fonctions de répartition. □

Des résultats analogues sont également vrais pour les fonctions caractéristiques conjointes. Nous ne les énoncerons pas explicitement.

Les fonctions caractéristiques sont aussi très utiles pour étudier la convergence de variables aléatoires (nous reviendrons sur les différents modes de convergence au chapitre 7).

Définition 6.3. *On dit qu'une suite de fonction de répartition F_n converge vers une fonction de répartition F, $F_n \to F$, si $F(x) = \lim_{n\to\infty} F_n(x)$, en chaque point x où F est continue.*

Théorème 6.4 (Théorème de continuité de Lévy)**.** *Soient $F_1, F_2, \ldots$ une suite de fonctions de répartition, et $\phi_1, \phi_2, \ldots$ les fonctions caractéristiques associées.*

1. *Si $F_n \to F$, pour une certaine fonction de répartition F de fonction caractéristique ϕ, alors $\phi_n(t) \to \phi(t)$ pour tout t.*
2. *Si $\phi(t) = \lim_{n\to\infty} \phi_n(t)$ existe et est continue en $t = 0$, alors ϕ est la fonction caractéristique associée à une fonction de répartition F, et $F_n \to F$.*

Démonstration. Nous ne la ferons pas ici. □

1. Poser, pour $n \geq 1$, $u_n = \int_0^{\pi/2} \sin\big((2n-1)x\big)/\sin(x)\,\mathrm{d}x$ et $v_n = \int_0^{\pi/2} \sin(2nx)/x\,\mathrm{d}x$. Montrer que : (i) $u_{n+1} = u_n$, $\forall n \geq 1$ (observez que $\sin\big((2n+1)x\big) - \sin\big((2n-1)x\big) = 2\cos(2nx)\sin(x)$) ; (ii) $u_1 = \pi/2$; (iii) $\lim_{n\to\infty}(u_n - v_n) = 0$ (intégration par parties en observant que $1/x - 1/\sin(x)$ est continûment différentiable sur $[0, \pi/2]$) ; (iv) $\lim_{T\to\infty} u(T, 1) = \lim_{n\to\infty} v_n = \pi/2$.

6.3 Quelques exemples classiques

Loi de Bernoulli

Si X suit une loi de Bernoulli de paramètre p, alors

$$\phi_X(t) = e^{\mathrm{i}t\cdot 0}(1-p) + e^{\mathrm{i}t\cdot 1}p = 1 - p + pe^{\mathrm{i}t}.$$

Loi binomiale

Puisqu'une variable aléatoire X de loi binomiale de paramètres n et p possède la même distribution que la somme de n v.a. de Bernoulli de paramètre p, on a

$$\phi_X(t) = (1 - p + pe^{\mathrm{i}t})^n.$$

Loi exponentielle

Soit X une variable aléatoire de loi exponentielle de paramètre λ. Dans ce cas,

$$\phi_X(t) = \lambda \int_0^\infty e^{-\lambda x + \mathrm{i}tx}\,\mathrm{d}x.$$

On vérifie facilement (exercice), en utilisant le théorème de Cauchy, qu'on peut remplacer l'intégrale le long de l'axe réel positif par une intégrale le long de la demi droite $(\lambda + \mathrm{i}t)s$, $s \in \mathbb{R}^+$. On obtient donc

$$\lambda \int_0^\infty e^{-(\lambda - \mathrm{i}t)x}\,\mathrm{d}x = \lambda(\lambda + \mathrm{i}t)\int_0^\infty e^{-(\lambda^2+t^2)s}\mathrm{d}s = \frac{\lambda(\lambda+\mathrm{i}t)}{\lambda^2+t^2} = \frac{\lambda}{\lambda - \mathrm{i}t}.$$

Loi de Cauchy

Si X suit une loi de Cauchy,

$$\phi_X(t) = \frac{1}{\pi}\int_{-\infty}^\infty \frac{e^{\mathrm{i}tx}}{1+x^2}\,\mathrm{d}x.$$

Pour la calculer, 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 $t < 0$ (il faut prendre le demi-cercle dans le demi-plan inférieur), on obtient finalement que

$$\phi_X(t) = e^{-|t|}, \text{pour tout } t \in \mathbb{R}.$$

Loi normale

On sait par le Lemme 6.2 qu'il est suffisant de considérer le cas où X est une variable aléatoire normale standard. 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 voit que

$$\phi_X(t) = e^{-\frac{1}{2}t^2}.$$

On a vu qu'une variable aléatoire Y de loi $\mathcal{N}(\mu, \sigma^2)$ peut s'écrire $Y = \sigma X + \mu$. On en déduit que sa fonction caractéristique est donnée par

$$\phi_Y(t) = e^{-\frac{1}{2}\sigma^2 t^2 + \mathrm{i}\mu t}.$$

Vecteurs aléatoires gaussiens

Observons tout d'abord que si $\mathbf{X} = (X_1, \ldots, X_n)$ est un vecteur aléatoire gaussien dont les composantes sont des variables aléatoires indépendantes de loi $\mathcal{N}(0, \sigma_i^2)$, alors

$$\phi_{\mathbf{X}}(\mathbf{t}) = \prod_{i=1}^{n} \phi_{X_i}(t_i) = e^{-\frac{1}{2}\langle \mathbf{t}, \mathbf{D}\mathbf{t} \rangle},$$

où $D_{ii} = \sigma_i^2$ et $D_{ij} = 0$ si $i \neq j$.

Considérons à présent un vecteur aléatoire gaussien $\mathbf{X}$ de loi $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$. Pour $\mathbf{t} \in \mathbb{R}^n$, $Y = \langle \mathbf{t}, \mathbf{X} \rangle$ est une variable aléatoire normale, et un calcul élémentaire montre que son espérance est donnée par

$$\mathbb{E}(Y) = \langle \mathbf{t}, \mathbb{E}(\mathbf{X}) \rangle,$$

et sa variance par

$$\mathrm{Var}(Y) = \langle \mathbf{t}, \mathrm{Cov}(\mathbf{X}, \mathbf{X})\mathbf{t} \rangle.$$

Par conséquent, la fonction caractéristique conjointe du vecteur $\mathbf{X}$ est donnée par

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}(e^{\mathrm{i}\langle \mathbf{t}, \mathbf{X} \rangle}) = \phi_Y(1) = e^{-\frac{1}{2}\langle \mathbf{t}, \mathrm{Cov}(\mathbf{X}, \mathbf{X})\mathbf{t} \rangle + \mathrm{i}\langle \mathbf{t}, \mathbb{E}(\mathbf{X}) \rangle}.$$

Déterminons à présent la fonction caractéristique conjointe de $\mathbf{X}$ d'une autre manière. La matrice de covariance $\mathbf{C}$ étant symétrique, on peut trouver une matrice orthogonale $\mathbf{U}$ et une matrice diagonale $\mathbf{D}$ telles que $\mathbf{C} = {}^{\mathrm{t}}\mathbf{U}\mathbf{D}\mathbf{U}$. On a donc, en posant $\mathbf{Z} = \mathbf{U}(\mathbf{X} - \boldsymbol{\mu})$,

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}(e^{\mathrm{i}\langle \mathbf{t}, \mathbf{X} \rangle}) = \mathbb{E}(e^{\mathrm{i}\langle \mathbf{U}\mathbf{t}, \mathbf{Z} \rangle}) \, e^{\mathrm{i}\langle \mathbf{t}, \boldsymbol{\mu} \rangle} = \phi_{\mathbf{Z}}(\mathbf{U}\mathbf{t}) \, e^{\mathrm{i}\langle \mathbf{t}, \boldsymbol{\mu} \rangle}.$$

Or, le vecteur aléatoire $\mathbf{Z}$ est un vecteur gaussien de loi $\mathcal{N}(\mathbf{0}, \mathbf{D})$, et ses composantes sont donc indépendantes. L'observation ci-dessus implique ainsi que

$$\phi_{\mathbf{Z}}(\mathbf{U}\mathbf{t}) = e^{-\frac{1}{2}\langle \mathbf{U}\mathbf{t}, \mathbf{D}\mathbf{U}\mathbf{t} \rangle} = e^{-\frac{1}{2}\langle \mathbf{t}, {}^{\mathrm{t}}\mathbf{U}\mathbf{D}\mathbf{U}\mathbf{t} \rangle} = e^{-\frac{1}{2}\langle \mathbf{t}, \mathbf{C}\mathbf{t} \rangle}.$$

On a donc

$$\phi_{\mathbf{X}}(\mathbf{t}) = e^{-\frac{1}{2}\langle \mathbf{t}, \mathbf{C}\mathbf{t} \rangle + \mathrm{i}\langle \mathbf{t}, \boldsymbol{\mu} \rangle}.$$

On déduit de ces deux calculs que $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$ et que $\mathrm{Cov}(\mathbf{X}, \mathbf{X}) = \mathbf{C}$.

Une autre propriété très importante de des vecteurs gaussiens est la suivante.

Théorème 6.5. *Soit $\mathbf{X} = (X_1, \ldots, X_n)$ un vecteur aléatoire gaussien. Alors, les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes si et seulement si elles sont non-corrélées.*

Démonstration. Nous avons vu que $X_1, \ldots, X_n$ sont indépendants si et seulement si la matrice de covariance $\mathbf{C}$ est diagonale. Mais, puisque $\mathbf{C} = \mathrm{Cov}(\mathbf{X}, \mathbf{X})$, ceci a lieu si et seulement si $X_1, \ldots, X_n$ sont non corrélées. □

Exemple 6.1. Nous pouvons à présent donner un exemple de vecteur aléatoire dont chaque composante suit une loi normale, mais qui n'est pas gaussien. Soit X une variable aléatoire de loi $\mathcal{N}(0, 1)$, et ϵ une variable aléatoire discrète, indépendante de X et telle que $\mathbb{P}(\epsilon = 1) = \mathbb{P}(\epsilon = -1) = \frac{1}{2}$. On considère la variable aléatoire $Y = \epsilon X$. On vérifie aisément (exercice) que Y suit une loi $\mathcal{N}(0, 1)$. X et Y ne sont manifestement pas indépendants ; par contre,

$$\mathbb{E}(XY) = \mathbb{E}(\epsilon X^2) = \mathbb{E}(\epsilon)\mathbb{E}(X^2) = 0,$$

ce qui montre que X et Y sont non-corrélées. Par conséquent, le vecteur aléatoire (X, Y) n'est pas gaussien.

Chapitre 7

Théorèmes limites

Le but de ce chapitre est d'étudier un certain nombres de résultats classiques de théorie des probabilités : les lois des grands nombres (faible et forte), le théorème central limite, et la loi 0-1 de Kolmogorov. Nous verrons aussi plusieurs résultats techniques très utiles, en particulier les inégalité de Markov/Tchebychev, et les Lemmes de Borel-Cantelli.

Les théorèmes limites sont omniprésents en théorie des probabilités. Une raison de leur importance est le fait que, en un certain sens, ils permettent de transformer des événements de probabilité $p \in [0,1]$ en des événements de probabilité proche de 0 ou 1, et ce n'est que pour de tels événements qu'un énoncé probabiliste devient falsifiable.

7.1 Un point technique

La plupart des résultats de ce chapitre portent sur des suites infinies de variables aléatoires $X_1, X_2, X_3, \ldots$ de loi conjointe donnée. L'existence d'un espace de probabilité sur lequel une telle famille de variables aléatoire peut être définie n'est pas évidente, et nous allons brièvement discuter cette question à présent.

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité, et $\mathbf{X} = \{X_t\}_{t \in T}$ une famille de variables aléatoires sur Ω. Nous avons vu qu'à tout vecteur $\mathbf{t} = (t_1, \ldots, t_n)$ d'éléments de T de longueur finie, on peut associer la fonction de répartition conjointe $F_{\mathbf{t}}$ du vecteur aléatoire $(X_{t_k})_{k=1,\ldots,n}$. L'ensemble de toutes ces fonctions de répartition conjointes (pour tous les vecteurs $\mathbf{t}$ de longueur finie) forme ce que l'on appelle le système des **lois fini-dimensionnelles** de $\mathbf{X}$. Il est évident que ces fonctions de répartition conjointes satisfont aux deux **conditions de consistance de Kolmogorov** :

$$\lim_{x_{n+1} \to \infty} F_{(t_1,\ldots,t_n,t_{n+1})}(x_1, \ldots, x_n, x_{n+1}) = F_{(t_1,\ldots,t_n)}(x_1, \ldots, x_n), \tag{7.1}$$

$$F_{\pi \mathbf{t}}(\pi \mathbf{x}) = F_{\mathbf{t}}(\mathbf{x}), \tag{7.2}$$

où π est une permutation de $(1, 2, \ldots, n)$ et, pour tout n-vecteur $\mathbf{y} = (y_1, \ldots, y_n)$, $\pi \mathbf{y} = (y_{\pi(1)}, \ldots, y_{\pi(n)})$.

Le résultat suivant montre que ces deux propriétés caractérisent les systèmes de lois fini-dimensionnelles.

Théorème 7.1 (Théorème de consistance de Kolmogorov). *Soit T un ensemble arbitraire, et supposons qu'à chaque vecteur $\mathbf{t} = (t_1, \ldots, t_n)$ d'éléments de T de longueur finie il corresponde une fonction de répartition jointe $F_{\mathbf{t}}$. Si la collection $\{F_{\mathbf{t}}\}$ satisfait aux conditions de consistance de Kolmogorov, alors il existe un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et une collection $\mathbf{X} = \{X_t, t \in T\}$ de variables aléatoires sur cet espace telle que $\{F_{\mathbf{t}}\}$ soit le système des lois fini-dimensionnelles de $\mathbf{X}$*

Démonstration. Il s'agit d'un résultat classique de théorie de la mesure, qui sera démontré en Analyse III. En voici une esquisse. Observez que la procédure est fortement réminiscente de celle employée dans la Section 1.5 pour construire un espace de probabilité sur lequel décrire la répétition d'une infinité d'expériences identiques indépendantes.

Soit $\Omega = \mathbb{R}^T$; les points de Ω sont les collections $\mathbf{y} = (y_t)_{t\in T}$ de nombres réels. Soit $\mathcal{F} = \mathcal{B}^T$ la tribu engendrée par les ensembles de la forme $\times_{t\in T} B_t$, avec $B_t = \mathbb{R}$ pour tout $t \in T$ sauf un nombre fini. Un résultat fondamental de théorie de la mesure affirme qu'il existe une mesure de probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{F})$ telle que

$$\mathbb{P}(\{\mathbf{y} \in \Omega \,:\, y_{t_1} \leq x_1, y_{t_2} \leq x_2, \ldots, y_{t_n} \leq x_n\}) = F_{\mathbf{t}}(\mathbf{x}),$$

pour tout $\mathbf{t}$ et $\mathbf{x}$. L'espace $(\Omega, \mathcal{F}, \mathbb{P})$ est l'espace recherché. Il suffit de définir $X_t : \Omega \to \mathbb{R}$ par

$$X_t(\mathbf{y}) = y_t$$

pour obtenir la famille désirée $(X_t)_{t\in T}$. □

7.2 Quelques outils

7.2.1 Les lemmes de Borel-Cantelli

Soit $A_1, A_2 \ldots$, une suite infinie d'événements sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. L'événement « une infinité des A_k sont réalisés » peut s'écrire

$$\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \limsup_{n\to\infty} A_n.$$

Il est souvent important de savoir quand un tel événement est réalisé.

Théorème 7.2 (Lemmes de Borel-Cantelli). *Soit $A_1, A_2 \ldots$, une suite infinie d'événements sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, et $A = \limsup_{n\to\infty} A_n$ l'événement « une infinité des A_n sont réalisés ». Alors*

1. *$\mathbb{P}(A) = 0$ si $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$.*
2. *$\mathbb{P}(A) = 1$ si $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ et les événements $A_1, A_2, \ldots$ sont indépendants.*

Démonstration. 1. Puisque $\bigcup_{m=n}^{\infty} A_m$ est une suite décroissante en n, il suit du Lemme 5.1 que

$$\mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcup_{m=n}^{\infty} A_m\Big) \leq \lim_{n\to\infty} \sum_{m=n}^{\infty} \mathbb{P}(A_m) = 0.$$

2. On vérifie aisément que

$$A^{\mathsf{c}} = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m^{\mathsf{c}}.$$

Cependant,

$$\begin{aligned}
\mathbb{P}(\bigcap_{m=n}^{\infty} A_m^c) &= \lim_{N\to\infty} \mathbb{P}(\bigcap_{m=n}^{N} A_m^c) && \text{(Lemme 5.1)}\\
&= \lim_{N\to\infty} \prod_{m=n}^{N} (1-\mathbb{P}(A_m)) && \text{(indépendance)}\\
&\le \lim_{N\to\infty} \prod_{m=n}^{N} \exp(-\mathbb{P}(A_m)) && (1-x\le e^{-x})\\
&= \lim_{N\to\infty} \exp(-\sum_{m=n}^{N} \mathbb{P}(A_m))\\
&= 0
\end{aligned}$$

dès que $\sum_{n=1}^{\infty}\mathbb{P}(A_n)=\infty$. Manifestement $(\bigcap_{m=n}^{\infty} A_m^c)_{n\ge 1}$ est une suite croissante d'événements ; il suit donc du Lemme 5.1 que

$$\mathbb{P}(A^c) = \lim_{n\to\infty} \mathbb{P}(\bigcap_{m=n}^{\infty} A_m^c) = 0.$$

□

Remarque 7.1. *Sans l'hypothèse d'indépendance, la seconde partie peut être fausse : il suffit de considérer la suite d'événements $A_k = B$, pour tout $k \ge 1$. Dans ce cas, $\mathbb{P}(A) = \mathbb{P}(B)$. On peut toutefois remplacer cette condition par l'indépendance 2 à 2 (mais la preuve est alors moins simple).*

7.2.2 Quelques inégalités

Supposons que $X_1, \dots, X_n$ soient des variables aléatoires obtenues en répétant n fois la même expérience de façon indépendante. Si $\mathbb{E}(X_i) = \mu$ pour chaque i, on a vu que l'espérance de $(X_1 + \cdots + X_n)/n$ vaut également μ. Mais est-il possible d'affirmer que la moyenne des X_i a de fortes chances d'être proche de μ ? C'est précisément le contenu de la loi faible des grands nombres, dont on a déjà vu une version dans le Théorème 2.3. Avant d'en énoncer une autre version, démontrons une inégalité extrêmement utile, variante de (2.2).

Théorème 7.3. *Soit $\varphi : \mathbb{R} \to [0,\infty)$. Alors*

$$\mathbb{P}(\varphi(X) \ge a) \le \frac{\mathbb{E}(\varphi(X))}{a}, \qquad \forall a > 0.$$

Démonstration. Soit $A = \{\varphi(X) \ge a\}$. Trivialement,

$$\varphi(X) \ge a\mathbf{1}_A,$$

et donc, en prenant l'espérance,

$$\mathbb{E}(\varphi(X)) \ge a\mathbb{E}(\mathbf{1}_A) = a\mathbb{P}(A).$$

□

Corollaire 7.1. *Soit X une variable aléatoire.*

1. *(Inégalité de Markov) Si $\mathbb{E}(|X|)$ est bien défini, alors*

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}(|X|)}{a}, \qquad \forall a > 0 \ ;$$

2. *(Inégalité de Bienaymé-Tchebychev) Si X possède une variance, alors*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}, \qquad \forall a > 0 \ ;$$

3. *(Inégalité de Chernoff) Soit*

$$H(t) = \begin{cases} \log \mathbb{E}(e^{tX}) & \text{si } \mathbb{E}(e^{tX}) < \infty, \\ \infty & \text{sinon.} \end{cases}$$

Alors, pour tout $a \in \mathbb{R}$,

$$\mathbb{P}(X \geq a) \leq \exp(-\sup_{t \geq 0}\{ta - H(t)\}).$$

Démonstration. 1. Il suffit de prendre $\varphi(x) = |x|$ dans le Théorème 7.3.

2. Par le Théorème 7.3, avec $\varphi(x) = x^2$, appliqué à la variable aléatoire $Y = X - \mathbb{E}(X)$, on a

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) = \mathbb{P}(Y^2 \geq a^2) \leq \frac{\mathbb{E}(Y^2)}{a^2} = \frac{\mathrm{Var}(X)}{a^2}.$$

3. En appliquant le Théorème 7.3 avec $\varphi(x) = e^{tx}$, on obtient

$$\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq e^{-ta}\mathbb{E}(e^{tX}) = e^{-(ta - H(t))},$$

pour tout $t \geq 0$. □

7.3 Modes de convergence

Le but de ce chapitre est d'étudier le comportement asymptotiques de certaines variables aléatoires. Pour ce faire, nous allons avoir besoin d'une notion de convergence d'une suite de variables aléatoires. Il se trouve qu'il existe plusieurs notions de convergence naturelles, que nous allons brièvement décrire dans cette section.

Définition 7.1. *Soient $X_1, X_2, \ldots$ et X des variables aléatoires sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. On dit que*

1. $X_n \to X$ ***presque sûrement***, *noté* $X_n \overset{\text{p.s.}}{\to} X$, *si*
$$\mathbb{P}\Big(\Big\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big\}\Big) = 1.$$
2. $X_n \to X$ ***en moyenne d'ordre*** r *($r \geq 1$), noté* $X_n \overset{r}{\to} X$, *si* $\mathbb{E}(|X_n^r|) < \infty$, *pour tout* n, *et*
$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|^r) = 0.$$
3. $X_n \to X$ ***en probabilité***, *noté* $X_n \overset{\mathbb{P}}{\to} X$, *si*
$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0, \qquad \forall \epsilon > 0.$$
4. $X_n \to X$ ***en loi***, *noté* $X_n \overset{\mathcal{L}_{\mathbb{P}}}{\to} X$, *si*
$$\lim_{n\to\infty} \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x),$$
en chaque point x en lesquels $F_X(x) = \mathbb{P}(X \leq x)$ est continue.

Remarque 7.2. *Lorsque* $X_n \overset{1}{\to} X$, *on parle de* ***convergence en moyenne***.
Lorsque $X_n \overset{2}{\to} X$, *on parle de* ***convergence en moyenne quadratique***.

Notons le résultat suivant, qui montre quelles sont les implications entre ces différents modes de convergence.

Théorème 7.4. *Les implications suivantes sont vérifiées :*

$$\begin{array}{ccc} (X_n \overset{\text{p.s.}}{\to} X) & & \\ \Downarrow & & \\ (X_n \overset{\mathbb{P}}{\to} X) & \Rightarrow & (X_n \overset{\mathcal{L}_{\mathbb{P}}}{\to} X) \\ \Uparrow & & \\ (X_n \overset{s}{\to} X) & & \\ \Uparrow & & \\ (X_n \overset{r}{\to} X) & & \end{array}$$

pour tout $r > s \geq 1$. 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 théorème suivant contient quelques résultats de ce type qui se révèlent particulièrement utiles.

Théorème 7.5.
1. *Si* $X_n \overset{\mathcal{L}_{\mathbb{P}}}{\to} c$, *avec c une constante, alors* $X_n \overset{\mathbb{P}}{\to} c$.
2. *Si* $X_n \overset{\mathbb{P}}{\to} X$ *et* $\exists k$ *tel que* $\mathbb{P}(|X_n| \leq k) = 1$, *pour tout n, alors* $X_n \overset{r}{\to} X$, *pour tout* $r \geq 1$.
3. *Si* $\sum_n \mathbb{P}(|X_n - X| > \epsilon) < \infty$, *pour tout* $\epsilon > 0$, *alors* $X_n \overset{\text{p.s.}}{\to} X$.

Démonstration. 1. $\mathbb{P}(|X_n - c| > \epsilon) = \mathbb{P}(X_n < c - \epsilon) + \mathbb{P}(X_n > c + \epsilon) \to 0$, si $X_n \overset{\mathcal{L}_{\mathbb{P}}}{\to} c$.

2. Montrons tout d'abord que si $X_n \xrightarrow{\mathbb{P}} X$ et $\mathbb{P}(|X_n| \leq k) = 1$, alors $\mathbb{P}(|X| \leq k) = 1$. En effet, cela implique que $X_n \xrightarrow{\mathcal{L}_{\mathbb{P}}} X$ et donc que $\mathbb{P}(|X| \leq k) = \lim_{n\to\infty} \mathbb{P}(|X_n| \leq k) = 1$. Posons à présent $A_n(\epsilon) = \{|X_n - X| > \epsilon\}$. Alors

$$|X_n - X|^r \leq \epsilon^r \mathbf{1}_{A_n(\epsilon)^c} + (2k)^r \mathbf{1}_{A_n(\epsilon)}, \qquad \mathbb{P}\text{-p.s.}$$

En prenant l'espérance, on obtient

$$\mathbb{E}(|X_n - X|^r) \leq \epsilon^r + (2k)^r \mathbb{P}(A_n(\epsilon)) \to \epsilon^r,$$

lorsque $n \to \infty$. La conclusion suit puisque ϵ était arbitraire.

3. Soit $A_M = \bigcup_{N\geq 1} \bigcap_{n\geq N} \{|X_n - X| \leq \frac{1}{M}\}$. Par le Lemme 5.1,

$$\begin{aligned}
\mathbb{P}(\{\omega : X_n(\omega) \to X(\omega)\}) &= \mathbb{P}(\{\omega : \forall M \geq 1, \exists N \geq 1 \text{ t.q. } |X_n(\omega) - X(\omega)| \leq \tfrac{1}{M}, \forall n \geq N\}) \\
&= \mathbb{P}\Big(\bigcap_{M\geq 1} \bigcup_{N\geq 1} \bigcap_{n\geq N} \{|X_n - X| \leq \tfrac{1}{M}\}\Big) \\
&= \lim_{M\to\infty} \mathbb{P}(A_M).
\end{aligned}$$

À présent, il suit de l'hypothèse et du Lemme de Borel-Cantelli que, pour tout $M \geq 1$,

$$\begin{aligned}
\mathbb{P}(A_M^c) &= \mathbb{P}\Big(\bigcap_{N\geq 1} \bigcup_{n\geq N} \{|X_n - X| > \tfrac{1}{M}\}\Big) \\
&= \mathbb{P}(\{|X_n - X| > \tfrac{1}{M}\} \text{ pour une infinité de valeurs de } n) = 0,
\end{aligned}$$

ce qui implique bien que $\mathbb{P}(\{\omega : X_n(\omega) \to X(\omega)\}) = 1$.

□

7.4 La loi des grands nombres

7.4.1 Loi faible des grands nombres

Nous avons vu une version de la loi faible des grands nombres dans le Théorème 2.3. 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 7.6 (Loi faible des grands nombres). *Soient $X_1, X_2, \ldots$ des variables aléatoires i.i.d. d'espérance μ. Alors $S_n \xrightarrow{\mathbb{P}} \mu$.*

Démonstration. Le point 1. du Théorème 7.5 implique qu'il suffit de démontrer que $S_n \xrightarrow{\mathcal{L}_{\mathbb{P}}} \mu$. Pour ce faire, on observe tout d'abord que, par le Lemme 6.1,

$$\phi_X(t) = 1 + it\mu + o(t).$$

La Proposition 6.1 et le Lemme 6.2 impliquent alors que la fonction caractéristique de la variable aléatoire $S_n = \frac{1}{n}\sum_{i=1}^n X_i$ satisfait

$$\phi_{S_n}(t) = (\phi_X(t/n))^n = (1 + \frac{it\mu}{n} + o(\frac{t}{n}))^n \to e^{it\mu},$$

lorsque $n \to \infty$. Comme $e^{it\mu}$ est la fonction caractéristique de la variable aléatoire constante μ, la convergence en loi suit du Théorème de continuité 6.4. □

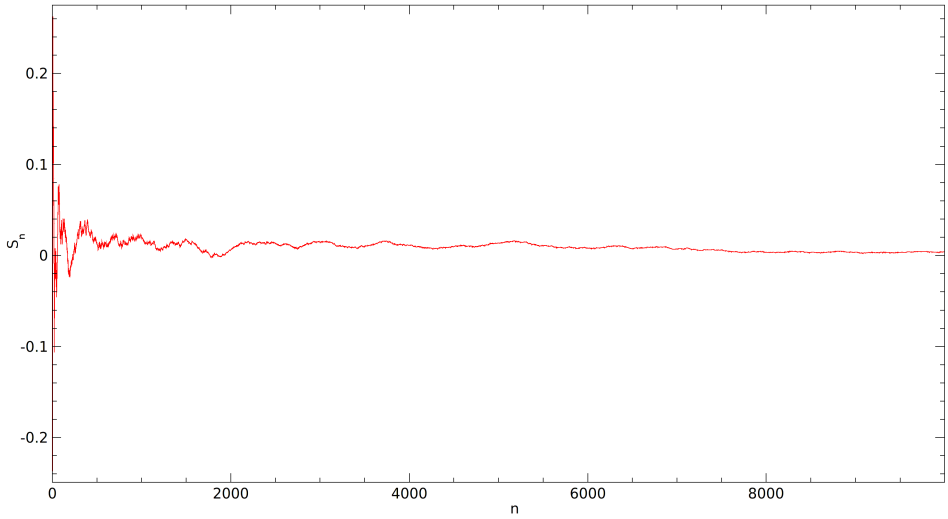


FIGURE 7.1: La moyenne empirique d'une famille de variables aléatoires de loi $\mathrm{U}(-1,1)$ (n allant de 1 à 10 000).

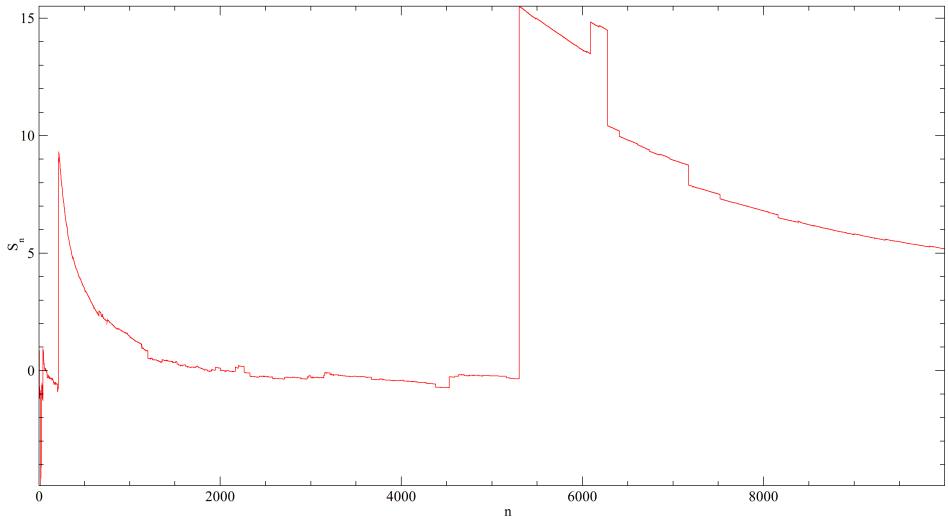


FIGURE 7.2: La moyenne empirique d'une famille de variables aléatoires suivant une loi de Cauchy (n allant de 1 à 10 000).

Remarque 7.3. *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 une loi de Cauchy. En effet, la fonction caractéristique de la somme de n variables aléatoires i.i.d. suivant une loi de Cauchy est donnée par*

$$\phi_{S_n}(t) = \left(\phi_X(t/n)\right)^n = e^{-|t|},$$

ce qui montre que S_n suit également une loi de Cauchy, et ne peut donc pas converger vers une constante ! La Figure 7.2 montre le comportement d'une réalisation de S_n pour n allant de 1 à 10 000 ; comparez ce comportement à celui de la Figure 7.1 qui montre le comportement correspondant pour des variables aléatoires uniformes !

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 7.1. Pour illustrer ce point, reprenons la discussion entamée dans l'Exemple 2.12, portant sur 10 000 jets d'une pièce équilibrée. Afin de travailler avec des variables centrées, on encode le résultat du $k^{\text{ème}}$ 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 H correspondante : $e^{H(t)} = \mathbb{E}(e^{tS_n}) = \mathbb{E}(\prod_{k=1}^n e^{tX_k/n}) = \prod_{k=1}^n \mathbb{E}(e^{tX_k/n}) = \cosh(t/n)^n$. On a donc

$$\mathbb{P}(S_n \geq x) \leq \inf_{t \geq 0} e^{(n \log\cosh(t/n) - tx)}.$$

Un petit calcul[1] montre que la fonction $f(t) = \log\cosh(t/n) - x\,t/n$ atteint son minimum en $t^* = \frac{n}{2}\log[(1+x)/(1-x)]$. En introduisant

$$I(x) = -f(t^*) = \frac{1}{2}\{(1+x)\log(1+x) + (1-x)\log(1-x)\},$$

et en utilisant la symétrie pour estimer $\mathbb{P}(S_n \leq -x)$, on a finalement

$$\mathbb{P}(|S_n| \geq x) \leq 2\,e^{-nI(x)}. \tag{7.3}$$

En posant $n = 10\ 000$ et $x = 0{,}1$, on trouve $I(0{,}1) \simeq 0{,}005$, et par conséquent

$$\mathbb{P}(S_{10000} \notin [-0{,}1, 0{,}1]) \leq 3{,}5 \cdot 10^{-22}.$$

Comparez ce résultat avec l'estimée de l'Exemple 2.12 !

Un résultat du type (7.3) est ce qu'on appelle une estimée de **grande déviation**. La théorie des grandes déviations est un domaine important de la théorie des probabilités, dont l'un des principaux artisans, S.R.S. Varadhan[2] et a été récompensée du prix Abe en 2007. ♢

7.4.2 La loi forte des grands nombres

La loi faible des grands nombres nous fournit des informations sur le comportement de S_n (pour n grand) lorsqu'on considère de nombreuses répétitions de l'expérience : pour tout grand n *fixé*, S_n est proche de μ pour la plupart des réalisations. Elle n'affirme cependant *pas* que, pour une réalisation ω donnée, la fonction $n \mapsto S_n(\omega)$ reste forcément proche de μ lorsque n augmente : elle laisse ouverte la possibilité qu'il existe $\epsilon > 0$ et une sous-suite $(n_k)_{k\geq 1}$, $n_k \to \infty$, telle que $|S_{n_k}(\omega) - \mu| > \epsilon$, pour tout $k \geq 1$. 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

$$\left|S_n - \mu\right| > \epsilon$$

sont réalisés.

Théorème 7.7. *Soit $X_1, X_2, \ldots$ une suite de variables aléatoires i.i.d. Alors, lorsque $n \to \infty$,*

$$\frac{1}{n}\sum_{i=1}^n X_i \overset{\text{p.s.}}{\to} \mu$$

pour une certaine constante μ, si et seulement si $\mathbb{E}(|X_1|) < \infty$. Dans ce cas, $\mu = \mathbb{E}(X_1)$.

1. Se rappeler que $\cosh(u) = 1/\sqrt{1-\tanh^2(u)}$ et que $\operatorname{argtanh}(u) = \frac{1}{2}\log\{(1+x)/(1-x)\}$.
2. Sathamangalam Ranga Iyengar Srinivasa Varadhan (1940, Chennai -), probabiliste américain d'origine indienne. Lauréat du prix Abel en 2007.

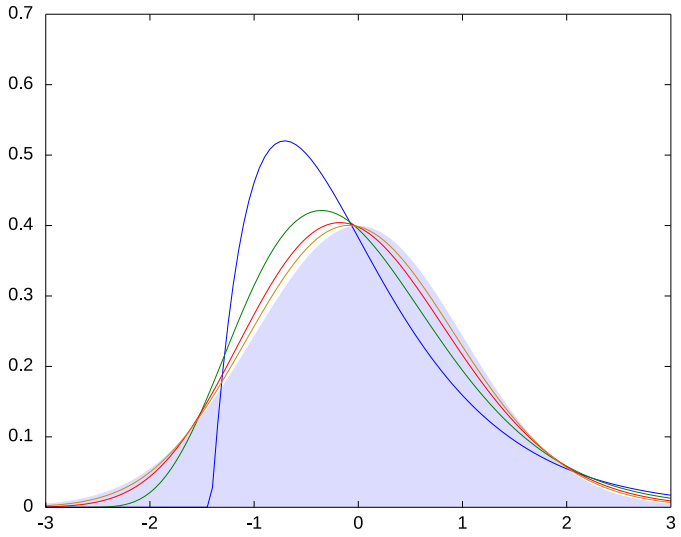


FIGURE 7.3: Convergence vers une loi normale pour une suite de variables aléatoires X_i de loi $\exp(1)$. Les courbes correspondent aux densités des variables $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - 1)$, pour $n = 2, 8, 32, 128$. La densité de la loi $\mathcal{N}(0,1)$ est aussi tracée.

Démonstration. Nous nous contenterons de démontrer la convergence, et ne le ferons que sous l'hypothèse que $\mathbb{E}(|X_1 - \mathbb{E}(X_1)|^4) < \infty$. Comme toujours, on peut supposer sans perte de généralité que $\mathbb{E}(X_1) = 0$. Dans ce cas, le Théorème 7.3 implique que $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ satisfait

$$\mathbb{P}(|S_n| > \epsilon) \leq \frac{\mathbb{E}(S_n^4)}{\epsilon^4}.$$

Puisque $\mathbb{E}(X_1) = 0$, on a

$$\mathbb{E}(S_n^4) = n^{-3}\mathbb{E}(X_1^4) + 3n^{-3}(n-1)\mathbb{E}(X_1^2)\mathbb{E}(X_2^2),$$

et il existe donc une constante C telle que, pour tout $\epsilon > 0$,

$$\sum_{n\geq 1}\mathbb{P}(|S_n| > \epsilon) \leq \frac{C}{\epsilon^4}\sum_{n\geq 1}\frac{1}{n^2} < \infty.$$

La convergence presque-sûre suit donc du point *3.* du Théorème 7.5. □

À présent que l'on sait que la moyenne empirique d'une suite de variables aléatoires indépendantes se concentre autour de son espérance, 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 $S_n - \mu$? La réponse à cette question, le Théorème Central Limite, est un des résultats majeurs de la théorie des probabilités, et est assez extraordinaire : il affirme que

1. $S_n - \mu$ est de l'ordre de $1/\sqrt{n}$.
2. La distribution de $\sigma(S_n - \mu)\sqrt{n}$ approche la même distribution, lorsque n devient grand, *quelle que soit la distribution des* X_i, tant que ceux-ci ont une variance σ^2 finie !

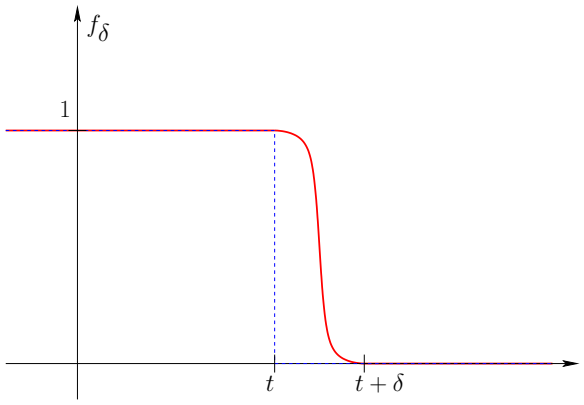


FIGURE 7.4: La fonction f_δ (en rouge) et l'indicatrice qu'elle approxime (traitillé).

7.5 Le Théorème Central Limite

Théorème 7.8 (Théorème Central Limite). *Soit $X_1, X_2, \ldots$ une suite de variables aléatoires i.i.d. telles que $\mathbb{E}(X_1) = \mu$ et $0 < \mathrm{Var}(X_1) = \sigma^2 < \infty$. Alors*

$$\frac{1}{\sqrt{n\sigma^2}} \sum_{k=1}^{n} (X_k - \mu) \xrightarrow{\mathcal{L}_{\mathbb{P}}} \mathcal{N}(0,1)\,.$$

Si, de plus, $\mathbb{E}(|X_1 - \mathbb{E}(X_1)|^3) < \infty$, alors

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}\left(\frac{1}{\sqrt{n\sigma^2}} \sum_{k=1}^{n} (X_k - \mu) \le x\right) - \Phi(x)\right| \le C \frac{\mathbb{E}(|X_1 - \mathbb{E}(X_1)|^3)}{\sigma^3 \sqrt{n}},$$

pour une certaine constante universelle $C < 0{,}4748$.

Remarque 7.4. *L'estimée explicite de l'erreur dans le théorème central limite donnée ci-dessus est appelée* ***inégalité de Berry[3]–Esséen[4]****. Elle joue un rôle très important lorsque l'on veut appliquer le théorème central limite dans la pratique.*

Démonstration. **Méthode directe.** On ne démontre que la seconde partie, et avec une estimation moins bonne de l'erreur. On peut supposer, sans perte de généralité, que $\mu = 0$ et $\sigma^2 = 1$ (sinon il suffit de considérer les variables aléatoires $\sigma^{-1}(X_i - \mu)$). Soit $Z_1, Z_2, \ldots$ une suite de variables aléatoires i.i.d. de loi $\mathcal{N}(0,1)$, indépendantes des variables aléatoires X_k. On pose

$$\widehat{S}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i, \qquad T_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i.$$

(Observez que T_n suit une loi $\mathcal{N}(0,1)$.) Soit $h : \mathbb{R} \to [0,1]$ une fonction de classe C^3, telle que $h(s) = 1$ si $s \le 0$, et $h(s) = 0$ si $s \ge 1$. Étant donné $t \in \mathbb{R}$ et $0 < \delta \le 1$, on définit une nouvelle fonction $f_\delta : \mathbb{R} \to [0,1]$ par (voir Fig. 7.4)

$$f_\delta(x) = h(\delta^{-1}(x - t)).$$

Par construction, $\mathbf{1}_{(-\infty,t]}(x) \le f_\delta(x)$, pour tout $x \in \mathbb{R}$, et donc

$$\mathbb{P}(\widehat{S}_n \le t) = \mathbb{E}(\mathbf{1}_{(-\infty,t]}(\widehat{S}_n)) \le \mathbb{E}(f_\delta(\widehat{S}_n)).$$

3. Andrew C. Berry (1???, -????,), mathématicien...
4. Carl-Gustav Esséen (1918, ??? - 2001, ???), mathématicien suédois.

Puisque $\Phi(t) = \mathbb{E}(\mathbf{1}_{(-\infty,t]}(T_n))$, on obtient donc

$$\mathbb{P}(\widehat{S}_n \le t) - \Phi(t) \le \mathbb{E}(f_\delta(\widehat{S}_n)) - \mathbb{E}(f_\delta(T_n)) + \mathbb{E}(f_\delta(T_n)) - \mathbb{E}(\mathbf{1}_{(-\infty,t]}(T_n)).$$

Manifestement, T_n suivant une loi $\mathcal{N}(0,1)$,

$$\mathbb{E}(f_\delta(T_n)) - \mathbb{E}(\mathbf{1}_{(-\infty,t]}(T_n)) = \frac{1}{\sqrt{2\pi}} \int_t^{t+\delta} h(\delta^{-1}(x-t)) e^{-x^2/2} \mathrm{d}x \le \frac{\delta}{\sqrt{2\pi}}.$$

Il reste donc à estimer $\mathbb{E}(f_\delta(\widehat{S}_n)) - \mathbb{E}(f_\delta(T_n))$. 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 :

$$\mathbb{E}(f_\delta(\widehat{S}_n)) - \mathbb{E}(f_\delta(T_n)) = \sum_{k=1}^{n} \{\mathbb{E}(f_\delta(U_k + \frac{X_k}{\sqrt{n}})) - \mathbb{E}(f_\delta(U_k + \frac{Z_k}{\sqrt{n}}))\},$$

où $U_k = (Z_1 + Z_2 + \cdots + Z_{k-1} + X_{k+1} + X_{k+2} + \cdots + X_n)/\sqrt{n}$. 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

$$f_\delta(U_k + (X_k/\sqrt{n})) = f_\delta(U_k) + \frac{X_k}{\sqrt{n}} f_\delta'(U_k) + \frac{X_k^2}{2n} f_\delta''(U_k) + \frac{X_k^3}{6n^{3/2}} f_\delta'''(Y),$$

avec $U_k \le Y \le U_k + (X_k/\sqrt{n})$. On traite de la même façon le terme $f_\delta(U_k + (Z_k/\sqrt{n}))$. On obtient ainsi

$$\mathbb{E}(f_\delta(U_k + \frac{X_k}{\sqrt{n}})) - \mathbb{E}(f_\delta(U_k + \frac{Z_k}{\sqrt{n}})) \le \frac{A}{6\delta^3 n^{3/2}} (\mathbb{E}(|X_k|^3) + \mathbb{E}(|Z_k|^3)),$$

où $A = \sup_{y\in\mathbb{R}} |h'''(y)| = \delta^3 \sup_{y\in\mathbb{R}} |f_\delta'''(y)|$. En choisissant $\delta = n^{-1/8}$, on obtient donc

$$\mathbb{P}(\widehat{S}_n \le t) - \Phi(t) \le C n^{-1/8}.$$

La borne inférieure est prouvée de façon similaire, en remplaçant la fonction f_δ par la fonction $g_\delta(x) = h(\delta^{-1}(x - t + \delta))$; observez que $g_\delta(x) \le \mathbf{1}_{(-\infty,t]}(x)$ pour tout x.

Méthode utilisant la fonction caractéristique. On ne démontre que la première affirmation. La preuve est presque identique à celle du Théorème 7.6. On peut à nouveau supposer, sans perte de généralité, que $\mu = 0$ et $\sigma^2 = 1$. Dans ce cas, il suit du Lemme 6.1 que

$$\phi_X(t) = 1 - \tfrac{1}{2}t^2 + o(t^2).$$

D'autre part, la Proposition 6.1 et le Lemme 6.2 impliquent que la fonction caractéristique de la variable aléatoire $\widehat{S}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ satisfait

$$\phi_{\widehat{S}_n}(t) = \{\phi_X(t/\sqrt{n})\}^n = \{1 - \frac{t^2}{2n} + o(\frac{t^2}{n})\}^n,$$

or cette dernière quantité converge vers $e^{-t^2/2}$, lorsque n tend vers l'infini. On reconnaît là la fonction caractéristique d'une variable aléatoire de loi $\mathcal{N}(0,1)$, et le résultat suit par conséquent du Théorème de continuité 6.4. □

Le Théorème Central Limite montre que, pour n grand, on a

$$\mathbb{P}\Big(\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} \in [a,b]\Big) \simeq \Phi(b) - \Phi(a),$$

ou encore

$$\mathbb{P}(\sum_{i=1}^n X_i \in [\widehat{a},\widehat{b}]) \simeq \Phi(\frac{\widehat{b} - n\mu}{\sqrt{n\sigma^2}}) - \Phi(\frac{\widehat{a} - n\mu}{\sqrt{n\sigma^2}}).$$

Exemple 7.2. 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 épreuve 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

$$\mathbb{P}(N \geq 50) = \mathbb{P}(N \in [50, 400]) \simeq \Phi(\infty) - \Phi(\frac{50-40}{\sqrt{36}}) \simeq 0{,}05.$$

Il y a environ 5% de chances d'obtenir au moins 50 pièces défectueuses.

À titre de comparaison, N suivant une loi binom$(400, 0{,}1)$, un calcul exact donne

$$\mathbb{P}(N \geq 50) = \sum_{k=50}^{400} \binom{400}{k} (0{,}1)^k (0{,}9)^{400-k} \simeq 0{,}06,$$

ce qui est assez proche de l'approximation précédente. ♢

7.6 La loi 0-1 de Kolmogorov

L'énoncé précis de ce résultat nécessite un peu de terminologie.

Définition 7.2. *Soit $X_1, X_2, \ldots$ une suite de variables aléatoires sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. Pour toute sous-collection $\{X_i, i \in I\}$, on note $\sigma(X_i, i \in I)$ la plus petite tribu telle que chaque X_i, $i \in I$, soit mesurable. $\sigma(X_i, i \in I)$ est appelée* **tribu engendrée** *par les variables aléatoires $X_i, i \in I$.*

$\sigma(X_i, i \in I)$ contient les événements que l'on peut définir à l'aide des X_i, $i \in I$.

Définition 7.3. *Soit $\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, \ldots)$. Alors, $\mathcal{T}_n \supseteq \mathcal{T}_{n+1} \supseteq \ldots$. La tribu $\mathcal{T}_\infty \overset{\text{df}}{=} \bigcap_n \mathcal{T}_n$ est appelée* **tribu asymptotique**. *Les éléments de cette tribu sont appelés* **événements asymptotiques**.

La tribu asymptotique contient des événements comme

$$\{(\sum_{i=1}^{n} X_i)_n \text{ converge}\}, \{\lim_n X_n \text{ existe}\}, \{\lim_n \frac{1}{n}(X_1 + \cdots + X_n) = 0\}, \ldots$$

Ceux-ci sont indépendants des valeurs prises par les X_i, $i \in I$, pour tout ensemble fini I.

Théorème 7.9 (loi 0-1 de Kolmogorov). *Si $X_1, X_2, \ldots$ sont des variables aléatoires indépendantes, alors tout événement $A \in \mathcal{T}_\infty$ satisfait $\mathbb{P}(A) \in \{0, 1\}$.*

Définition 7.4. *Une tribu dont tous les éléments sont de probabilité 0 ou 1 est dite* **triviale**.

Démonstration. Soit $A \in \mathcal{T}_\infty$. Puisque $A \in \mathcal{T}_n$, pour tout n, et que $\mathcal{T}_n$ est indépendant de $\sigma(X_1, X_2, \ldots, X_n)$, on en déduit que A est indépendant de $\bigcup_n \sigma(X_1, X_2, \ldots, X_n)$.

On vérifie aisément que la classe des événements indépendants de A forme une classe monotone. Puisque cette classe contient l'algèbre $\bigcup_n \sigma(X_1, X_2, \ldots, X_n)$, il suit du Théorème des classes monotones qu'elle contient également la tribu engendrée $\sigma(X_1, X_2, \ldots)$. Par conséquent, A est indépendant de $\sigma(X_1, X_2, \ldots)$.

Or, $A \in \sigma(X_1, X_2, \ldots)$. 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 7.5. *Une variable aléatoire mesurable par rapport à la tribu asymptotique $\mathcal{T}_\infty$ est dite* ***asymptotique****.*

Corollaire 7.2. *Soient $X_1, X_2, \ldots$ des variables aléatoires indépendantes, et Y une variable aléatoire asymptotique. Alors il existe $y \in \mathbb{R}$ tel que*

$$\mathbb{P}(Y = y) = 1.$$

Démonstration. Y est asymptotique si et seulement si

$$\{\omega \in \Omega : Y(\omega) \leq x\} \in \mathcal{T}_\infty,$$

pour tout $x \in \mathbb{R}$. La loi 0-1 de Kolmogorov implique la trivialité de $\mathcal{T}_\infty$. Par conséquent, la fonction de répartition de Y satisfait

$$F_Y(x) = \mathbb{P}(Y \leq x) \in \{0, 1\}.$$

Soit $y = \inf\{x : \mathbb{P}(Y \leq x) = 1\}$ (avec la convention que $\inf \varnothing = \infty$). On a donc $F_Y(x) = \mathbf{1}_{[y,\infty)}(x)$, ce qui implique que $Y = y$ presque sûrement.

□

Chapitre 8

Retour aux marches aléatoires

Dans ce chapitre, nous reprenons l'étude des marches aléatoires simples. Nous allons en particulier nous intéresser à des événements portant sur les trajectoires infinies du processus.

8.1 Compléments concernant la marche sur $\mathbb{Z}$

8.1.1 Ruine du joueur

Parmi les nombreuses interprétations de la marche aléatoire simple sur $\mathbb{Z}$, une des plus classiques est la suivante : a représente la fortune initiale d'un joueur jouant à un jeu lors duquel, à chaque étape, il fait une mise égale à 1 (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).

Sous 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 d'arriver à une fortune égale $N > a$. Quelle est la probabilité qu'il soit ruiné avant de réaliser son objectif ?

En notant A l'événement correspondant, on déduit de la propriété de Markov que

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

Par conséquent, la fonction $a \mapsto \mathbb{P}_a(A)$ est solution de l'équation aux différences finies suivante

$$\begin{cases} f(a) = p\,f(a+1) + q\,f(a-1), & 1 \le a \le N-1\\ f(0) = 1, f(N) = 0. \end{cases} \tag{8.1}$$

Lemme 8.1. *L'équation (8.1) possède une unique solution.*

Démonstration. Si f et g sont solutions de (8.1), alors $h = f - g$ est solution de

$$\begin{cases} h(x) = p\,h(x+1) + q\,h(x-1),\\ h(0) = h(N) = 0. \end{cases}$$

Soit $\bar{x} \in \{1, \dots, N-1\}$ tel que $|h(\bar{x})|$ soit maximum ; on suppose sans perte de généralité que $h(\bar{x}) \ge 0$ (sinon il suffit de considérer $g - f$). On a alors

$$h(\bar{x}+1) = \frac{1}{p}(h(\bar{x}) - q\,h(\bar{x}-1)) \ge \frac{1}{p}(h(\bar{x}) - q\,h(\bar{x})) = h(\bar{x}),$$

puisque $h(\bar{x})$ est maximum. En itérant cette procédure, on obtient que $h(N) = h(\bar{x})$. Comme $h(N) = 0$, ceci implique que $h \equiv 0$, et donc que $f = g$. □

Pour un jeu équitable, $p = q = \frac{1}{2}$. Dans ce cas, on vérifie que l'unique solution à (8.1) est donnée par [1]

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

Lorsque $p \neq q$, on vérifie aisément qu'elle est donnée par [2]

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

8.1.2 Visites et retours

Dans le chapitre 3, nous nous sommes intéressé au temps $\tau_0 = \min\{n \geq 1 : S_n = 0\}$ du premier retour au point de départ. Nous avons en particulier 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 $\tau_b = \min\{n \geq 1 : S_n = b\}$ de la première visite en $b \neq 0$.

Théorème 8.1. *On considère la marche symétrique sur $\mathbb{Z}$. Pour tout $b \in \mathbb{Z}$,*

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

Démonstration. On a déjà établi que $\mathbb{P}_0(\tau_0 < \infty) = 1$. Par symétrie, on peut supposer $b > 0$. En conditionnant sur X_1, on voit que la fonction $b \mapsto \mathbb{P}_0(\exists n \geq 0 : S_n = b)$ est solution de l'équation aux différences finies suivante :

$$\begin{cases} f(x) = \frac{1}{2}(f(x+1) + f(x-1)), & x > 0 \\ f(0) = 1. \end{cases}$$

Évidemment, les solutions de cette équation sont données par les fonctions de la forme $f(x) = 1 + \alpha x$, $\alpha \in \mathbb{R}$. Par conséquent, l'unique solution bornée est donnée par $f \equiv 1$. On en conclut

1. Une façon de trouver cette solution est d'observer que (8.1) peut être écrite, lorsque $p = q = \frac{1}{2}$, sous la forme $f(a+1) - f(a) = f(a) - f(a-1) = \delta$, $\forall 1 \leq a \leq N-1$, pour une certaine valeur de δ. Par conséquent, $f(a) = f(0) + (f(1) - f(0)) + \cdots + (f(a) - f(a-1)) = 1 + a\delta$. En particulier, $0 = f(N) = 1 + N\delta$, d'où l'on tire $\delta = -1/N$ et $f(a) = 1 - a/N$.

2. Trouver la solution lorsque $p \neq q$ conduit à des calculs plus pénibles, mais est conceptuellement très simple. On peut dans ce cas écrire (8.1) sous la forme suivante : pour tout $1 \leq a \leq N-1$,

$$\begin{pmatrix} f(a+1) \\ f(a) \end{pmatrix} = \begin{pmatrix} 1/p & -q/p \\ 1 & 0 \end{pmatrix} \begin{pmatrix} f(a) \\ f(a-1) \end{pmatrix} = \begin{pmatrix} 1/p & -q/p \\ 1 & 0 \end{pmatrix}^a \begin{pmatrix} f(1) \\ f(0) \end{pmatrix}. \tag{8.2}$$

Un calcul (diagonalisez la matrice !) montre que

$$\begin{pmatrix} 1/p & -q/p \\ 1 & 0 \end{pmatrix}^a = \frac{1}{p-q} \begin{pmatrix} p - (q/p)^a & q(q/p)^a - q \\ p - p(q/p)^a & p(q/p)^a - q \end{pmatrix}.$$

En particulier, comme $f(N) = 0$ et $f(0) = 1$, on tire de (8.2) avec $a = N-1$ que

$$f(1) = \frac{(q/p) - (q/p)^N}{1 - (q/p)^N},$$

et donc, en appliquant à nouveau (8.2),

$$f(a) = \frac{(q/p)^a - (q/p)^N}{1 - (q/p)^N}.$$

donc que, par symétrie,

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

□

La marche aléatoire simple symétrique sur $\mathbb{Z}$ visite donc presque-sûrement chaque sommet. En fait, il est facile de voir qu'elle le fait presque-sûrement infiniment souvent.

Théorème 8.2. *Soit N_b le nombre de visites de la marche aléatoire en $b \in \mathbb{Z}$. Alors, dans le cas symétrique,*

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

Démonstration. Soit $\tau_b^{(n)}$ le temps de la $n^{\text{ème}}$ visite en b (avec $\tau_b^{(n)} = \infty$ si $N_b < n$). Pour tout $k \in \mathbb{N}$,

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

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

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

et donc $\mathbb{P}_0(N_b = k) = 0$, pour tout $k \in \mathbb{N}$. La conclusion suit donc, puisque

$$\mathbb{P}_0(N_b < \infty) = \sum_{k \geq 1} \mathbb{P}_0(N_b = k) = 0.$$

□

On peut en fait facilement déterminer la loi de τ_b, même dans le cas asymétrique.

Théorème 8.3. *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. 8.1)

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

Manifestement, ces deux marches aléatoires ont même loi. Observez à présent que la première de ces marches satisfait $S_n = b > 0$ et $\tau_0 > n$ si et seulement si la marche renversée satisfait $R_n = b$ et $R_n - R_i = X_1 + \cdots + X_{n-i} > 0$ pour tout $1 \leq i < n$, ce qui signifie que la première visite de la marche renversée au point b a lieu au temps n. On a donc démontré le résultat suivant :

$$\mathbb{P}_0(S_n = b, \tau_0 > n) = \mathbb{P}_0(R_n = b, \max_{0 \leq i < n} R_i < b) = \mathbb{P}_0(S_n = b, \max_{0 \leq i < n} S_i < b) = \mathbb{P}_0(\tau_b = n).$$

La conclusion suit donc du Thèorème 3.1. □

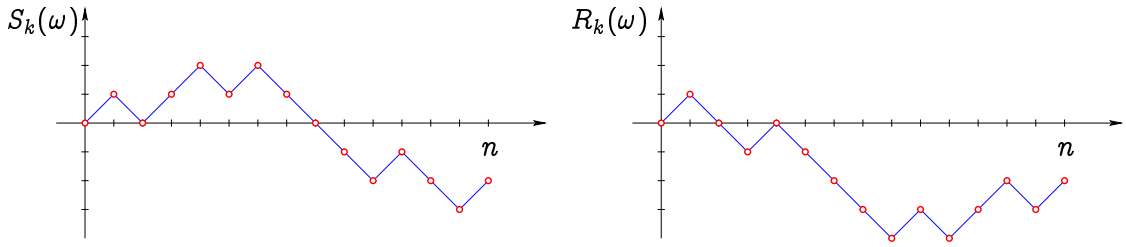


FIGURE 8.1: Une trajectoire et la trajectoire renversée.

Ce résultat a une conséquence assez surprenante.

Corollaire 8.1. *Dans le cas symétrique, le nombre moyen de visites de la marche (partant de 0) en un site $b \neq 0$ quelconque avant de retourner à l'origine est égal à 1.*

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

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

La conclusion suit donc du Théorème 8.1. □

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 » 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* !

8.2 Marche aléatoire simple sur $\mathbb{Z}^d$

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 [3] en 1905 afin d'expliquer le mouvement erratique des particules de pollen dans l'eau observé en 1827 par Brown [4], et de cette façon confirmer la théorie atomiste alors encore controversée en permettant à Perrin [5] de déterminer expérimentalement la constante d'Avogadro [6]).

Soit $X_1, X_2, \ldots$ une suite de vecteurs aléatoires i.i.d. prenant valeurs dans l'ensemble $\{\pm\vec{e}_i, i = 1, \ldots, d\}$ et de loi uniforme ; ici, $\vec{e}_i = (\delta_{ik})_{k=1,\ldots,d}$ est le vecteur unité de $\mathbb{R}^d$ dans la direction i. On appelle **marche aléatoire simple symétrique sur $\mathbb{Z}^d$ partant de $a \in \mathbb{Z}^d$** le processus

$$S_n = a + \sum_{i=1}^{n} X_i.$$

Comme précédemment, on note $\mathbb{P}_a$ la loi de la marche partant de a.

3. Albert Einstein (1879, Ulm – 1955, Princeton), physicien allemand, puis apatride (1896), suisse (1899), et enfin suisse-américain (1940). Prix Nobel de physique en 1921.
4. Robert Brown (1773, Montrose – 1858, Londres), botaniste britannique.
5. Jean Baptiste Perrin (1870, Lille – 1942, New York), physicien français. Prix Nobel de Physique en 1926.
6. Lorenzo Romano Amedeo Carlo Avogadro, Comte de Quaregna et Cerreto (1776, Turin – 1856, Turin). Physicien et chimiste italien.

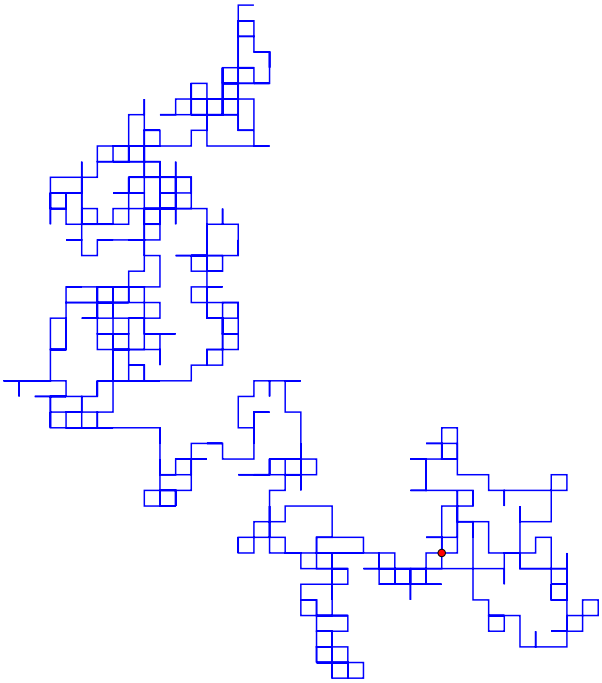


FIGURE 8.2: Les 1000 premiers pas d'une marche aléatoire simple symétrique sur $\mathbb{Z}^2$ partant du point rouge.

Ce processus décrit donc une particule se déplaçant aléatoirement de proche en proche sur le réseau $\mathbb{Z}^d$. Ce type de processus a été énormément étudié, et nous nous contenterons ici d'illustrer simplement quelques résultats élémentaires.

On vérifie aisément que les propriétés énoncées dans le Lemme 3.1 sont également vérifiées ici (la structure étant identique).

8.2.1 Probabilités de sortie

Le but de cette sous-section est de montrer que l'approche utilisée dans le cas unidimensionnel dans la Sous-section 8.1.1 s'étend sans autre à cette situation plus générale (Figure 8.3)).

Lemme 8.2. *Soit $\varnothing \neq D_1 \subset D_2 \subset \mathbb{Z}^d$. On note $T = \min\{n \geq 0 : S_n \notin D_2\}$ et $\tau = \min\{n \geq 0 : S_n \in D_1\}$. Alors la probabilité $\mathbb{P}_x(\tau < T)$ que la marche visite D_1 avant de quitter D_2 est donnée par l'unique solution de*

$$\begin{cases} \Delta_d f(x) = 0 & x \in D_2 \setminus D_1, \\ f(x) = 1 & x \in D_1, \\ f(x) = 0 & x \notin D_2, \end{cases}$$

où Δ_d est le ***Laplacien discret*** *sur $\mathbb{Z}^d$, défini par*

$$\Delta_d f(x) = \frac{1}{2d} \sum_{\substack{y \in \mathbb{Z}^d \\ |x-y|=1}} f(y) - f(x).$$

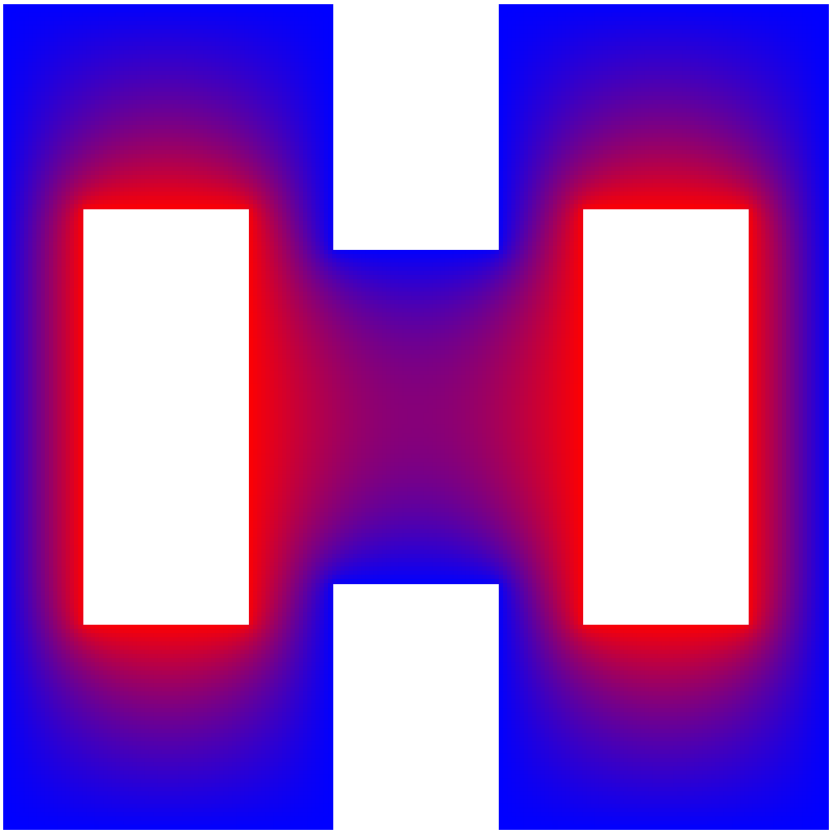


FIGURE 8.3: Probabilités de pénétrer dans un des deux trous avant de sortir du domaine : la couleur passe du bleu au rouge lorsque la probabilité passe de 0 à 1.

Démonstration. Par la propriété de Markov, on a, pour $x \in D_2 \setminus D_1$,

$$\begin{aligned}\mathbb{P}_x(\tau < T) &= \sum_{\substack{y\in\mathbb{Z}^d \\ |y-x|=1}} \mathbb{P}_x(\tau < T \,|\, S_1 = y)\mathbb{P}(S_1 = y) \\ &= \frac{1}{2d} \sum_{\substack{y\in\mathbb{Z}^d \\ |y-x|=1}} \mathbb{P}_y(\tau < T),\end{aligned}$$

et donc $\mathbb{P}_x(\tau < T)$ est bien solution de l'équation aux différences finies annoncée. Pour montrer que cette dernière possède une unique solution, on procède comme dans le cas unidimensionnel. Si f, g sont deux solutions de (8.2), alors $h = f - g$ est solution de la même équation, mais avec condition au bord $h(x) = 0$ pour tout $x \notin D_2 \setminus D_1$. Soit $z \in D_2 \setminus D_1$ un sommet où $|h|$ atteint son maximum. On a $\sum_{y:\,|y-z|=1}(h(y) - h(z)) = 0$. Tous les termes de la somme ayant le même signe, ceci implique que $h(y) = h(z)$, pour tout y voisin de z, et donc, en itérant, que $h \equiv$const. La condition au bord force alors $h \equiv 0$, ce qui est équivalent à $f \equiv g$. □

8.2.2 Récurrence et transience des marches aléatoires sur $\mathbb{Z}^d$

Finalement, nous allons nous intéresser à un problème classique : déterminer si la marche aléatoire simple est récurrente ou transiente. Nous avons déjà vu que dans le cas $d = 1$, la marche symétrique est récurrente-nulle. Le résultat suivant a été démontré par Pólya [7] en 1921 ; il montre que la dimension du réseau affecte crucialement le comportement de la marche aléatoire.

7. George Pólya (1887, Budapest – 1985, Palo Alto), mathématicien hongrois.

Théorème 8.4. *La marche aléatoire simple symétrique sur $\mathbb{Z}^d$ est récurrente si et seulement si $d \leq 2$.*

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. La première observation est le lemme suivant.

Lemme 8.3. *Soit N le nombre de retours de la marche aléatoire simple à l'origine. Alors*

$$S_n \text{ est récurrente} \iff \mathbb{E}_0(N) = \infty \iff \sum_{n\geq 1} \mathbb{P}_0(S_n = 0) = \infty.$$

Démonstration. Soit $r = \mathbb{P}_0(N \geq 1)$ la probabilité de retour à l'origine, et soit $\tau_0^{(n)}$ le temps du $n^{\text{ème}}$ retour en 0 (avec $\tau_0^{(n)} = \infty$ si $N < n$). Il suit de la propriété de Markov que, pour tout $n \geq 1$,

$$\begin{aligned}\mathbb{P}_0(N \geq n \,|\, N \geq n-1) &= \sum_{k \geq 2n-2} \mathbb{P}_0(N \geq n \,|\, \tau_0^{(n-1)} = k)\, \mathbb{P}_0(\tau_0^{(n-1)} = k \,|\, N \geq n-1) \\ &= r \sum_{k \geq 2n-2} \mathbb{P}_0(\tau_0^{(n-1)} = k \,|\, N \geq n-1) = r.\end{aligned}$$

Il suit donc que $\mathbb{P}_0(N \geq n) = r\,\mathbb{P}_0(N \geq n-1) = r^2\,\mathbb{P}_0(N \geq n-2) = \ldots = r^n$. Par conséquent,

$$\mathbb{E}_0(N) = \sum_{n\geq 1} \mathbb{P}_0(N \geq n) = \begin{cases} r/(1-r) & \text{si } r < 1 \\ \infty & \text{si } r = 1 \end{cases}$$

ce qui démontre la première équivalence. Puisque

$$\mathbb{E}_0(N) = \mathbb{E}_0\Big(\sum_{n\geq 1} \mathbf{1}_{\{S_n = 0\}}\Big) = \sum_{n\geq 1} \mathbb{P}_0(S_n = 0),$$

le lemme est démontré. □

En utilisant l'identité

$$\int_{[-\pi,\pi]^d} \frac{\mathrm{d}p}{(2\pi)^d} e^{\mathrm{i}\langle p, x\rangle} = \mathbf{1}_{\{x=0\}}, \qquad \forall x \in \mathbb{Z}^d,$$

on obtient

$$\mathbb{P}_0(S_n = 0) = \int_{[-\pi,\pi]^d} \frac{\mathrm{d}p}{(2\pi)^d} \mathbb{E}_0(e^{\mathrm{i}\langle p, S_n\rangle}),$$

et $\mathbb{E}_0(e^{\mathrm{i}\langle p, S_n\rangle}) = (\mathbb{E}(e^{\mathrm{i}\langle p, X_1\rangle}))^n = (\phi_{X_1}(p))^n$. Un calcul élémentaire montre que la fonction caractéristique de X_1 satisfait $\phi_{X_1}(p) = \frac{1}{d}\sum_{i=1}^d \cos(p_i)$, pour tout $p = (p_1, \ldots, p_d)$. Par conséquent, pour tout $0 < \lambda < 1$,

$$\begin{aligned}\sum_{n\geq 1} \lambda^n \mathbb{P}_0(S_n = 0) &= \int_{[-\pi,\pi]^d} \frac{\mathrm{d}p}{(2\pi)^d} \sum_{n\geq 1} (\lambda \phi_{X_1}(p))^n \\ &= \int_{[-\pi,\pi]^d} \frac{\mathrm{d}p}{(2\pi)^d} \frac{\lambda \phi_{X_1}(p)}{1 - \lambda \phi_{X_1}(p)}.\end{aligned}$$

On aimerait prendre la limite $\lambda \uparrow 1$ à présent, mais cela nécessite quelques précautions. Pour le membre de gauche, c'est facile : $\sum_{n\geq 1} \lambda^n \mathbf{1}_{\{S_n=0\}}$ est clairement une suite croissante de fonctions intégrables positives, et donc on peut permuter la limite et la somme en utilisant le Théorème

de la convergence monotone. En ce qui concerne le terme 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 la convergence monotone que

$$\lim_{\lambda\uparrow 1} \int_{[-1,1]^d} \frac{\mathrm{d}p}{(2\pi)^d} \frac{\lambda\phi_{X_1}(p)}{1-\lambda\phi_{X_1}(p)} = \int_{[-1,1]^d} \frac{\mathrm{d}p}{(2\pi)^d} \frac{\phi_{X_1}(p)}{1-\phi_{X_1}(p)}.$$

Pour traiter le reste, on observe que la suite de fonctions $\lambda\phi_{X_1}(p)/(1-\lambda\phi_{X_1}(p))$ converge ponctuellement et est uniformément bornée sur $[-\pi,\pi]^d \setminus [-1,1]^d$. Par conséquent, il suit du Théorème de convergence dominée que

$$\lim_{\lambda\uparrow 1} \int_{[-\pi,\pi]^d\setminus[-1,1]^d} \frac{\mathrm{d}p}{(2\pi)^d} \frac{\lambda\phi_{X_1}(p)}{1-\lambda\phi_{X_1}(p)} = \int_{[-\pi,\pi]^d\setminus[-1,1]^d} \frac{\mathrm{d}p}{(2\pi)^d} \frac{\phi_{X_1}(p)}{1-\phi_{X_1}(p)}.$$

On a donc finalement bien

$$\sum_{n\geq 1} \mathbb{P}_0(S_n = 0) = \int_{[-\pi,\pi]^d} \frac{\mathrm{d}p}{(2\pi)^d} \frac{\phi_{X_1}(p)}{1-\phi_{X_1}(p)}.$$

Le problème se réduit donc à l'analyse de la divergence de l'intégrande du membre de droite en $p = 0$. Par un développement de Taylor, on a que

$$\cos(x) = 1 - \tfrac{1}{2}x^2 + \tfrac{1}{24}x_0^4,$$

avec $0 \leq x_0 \leq x$. Par conséquent, pour tout $x \in [-1,1]$,

$$1 - \tfrac{1}{2}x^2 \leq \cos(x) \leq 1 - \tfrac{11}{24}x^2.$$

On en déduit que $\frac{1}{2d}\|p\|^2 \geq 1-\phi_{X_1}(p) \geq \frac{11}{24d}\|p\|^2$ au voisinage de 0. On voit donc que l'intégrande se comporte comme $\|p\|^{-2}$ au voisinage de 0. Par conséquent, l'intégrale converge si et seulement si $d > 2$. □

Remarque 8.1. *Le résultat précédent montre que lorsque $d \geq 3$, la probabilité π_d de retour au point de départ est inférieure à 1. Il est en fait possible de la déterminer. On peut montrer que $\pi_d = 1 - 1/u(d)$, où*

$$u(d) = \frac{d}{(2\pi)^d} \int_{-\pi}^{+\pi} \cdots \int_{-\pi}^{+\pi} \frac{\mathrm{d}x_1 \cdots \mathrm{d}x_d}{d - \cos x_1 - \cdots - \cos x_d}.$$

On obtient ainsi, par exemple : $\pi_3 \simeq 0{,}340$, $\pi_4 \simeq 0{,}193$, $\pi_5 \simeq 0{,}135$, etc.

Théorème 8.5. *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 $X_k = (X_k(1), X_k(2))$, $k \geq 1$, les incréments correspondants. On a déjà vu que S_n est récurrente, il suffit donc de montrer que $E_0(\tau_0) = \infty$.

On vérifie très facilement que le processus $\widetilde{S}_n = S_n(1) + S_n(2)$ est une marche aléatoire simple symétrique *unidimensionnelle* (il suffit de voir que $X_n(1) + X_n(2)$ est une variable aléatoire uniforme sur $\{-1,1\}$). Par conséquent, si on note $\widetilde{\tau}_0$ le temps de premier retour de $\widetilde{S}_n$, on a

$$\mathbb{E}_0(\tau_0) = \mathbb{E}_0(\inf\{n \geq 1 \,:\, S_n(1) = S_n(2) = 0\}) \geq \mathbb{E}_0(\inf\{n \geq 1 \,:\, \widetilde{S}_n = 0\}) = \mathbb{E}_0(\widetilde{\tau}_0) = \infty,$$

puisque la marche aléatoire simple symétrique sur $\mathbb{Z}$ est récurrente-nulle. □

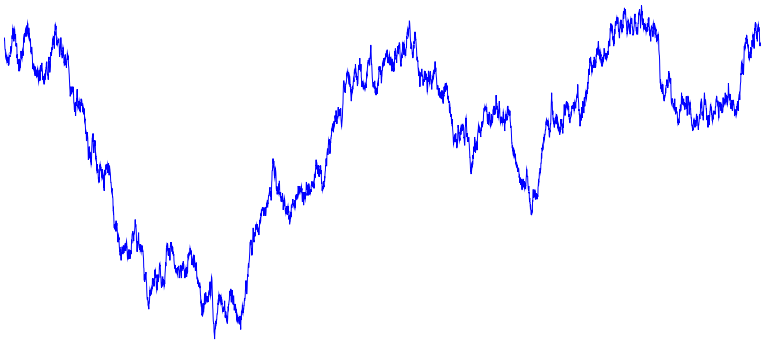


FIGURE 8.4: Partie d'une trajectoire du mouvement brownien en dimension 1.

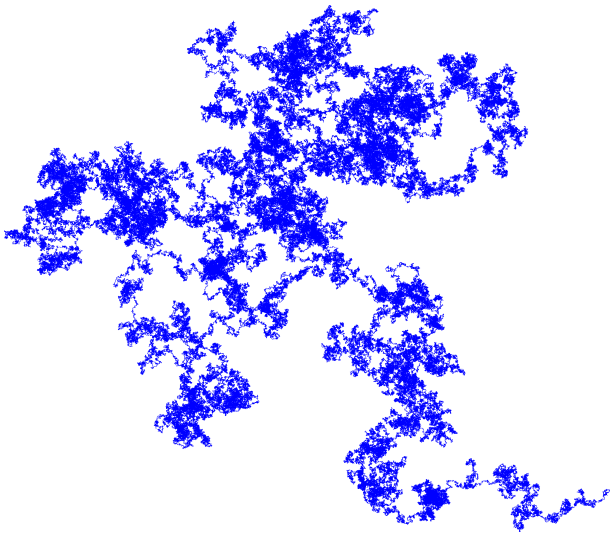


FIGURE 8.5: Partie d'une trajectoire du mouvement brownien en dimension 2 (tous les temps considérés sont superposés).

8.2.3 Convergence vers le mouvement brownien

On considère une marche aléatoire simple symétrique $(S_n)_{n\geq 0}$ sur $\mathbb{Z}$. Le théorème central limite implique que, pour tout $t \in \mathbb{R}_+$,

$$\frac{1}{\sqrt{N}} S_{[tN]} \overset{\mathcal{L}_{\mathbb{P}_0}}{\to} \mathcal{N}(0,t), \qquad N \to \infty.$$

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** [8]. Une partie d'une trajectoire de ce processus est donnée sur la Figure 8.4.

Similairement, 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 8.5, on a tracé une portion de trajectoire dans le cas bidimensionnel.

8. Norbert Wiener (1894, Columbia – 1964, Stockholm), mathématicien américain.

Chapitre 9

Les chaînes de Markov

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.

9.1 Définition et exemples

Soit $X_0, X_1, X_2, \dots$ 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 9.1. *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 $n \geq 1$ et tout $s_0, s_1, \dots, s_n \in S$.

S est appelé ***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 9.2. *Une chaîne de Markov X est* ***homogène*** *si*

$$\mathbb{P}(X_n = j \,|\, X_{n-1} = i) = \mathbb{P}(X_1 = j \,|\, X_0 = i),$$

pour tout $n \geq 1$ et tout $i, j \in S$.

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 l'évolution de la chaîne est caractérisée par la matrice $\mathbf{P} = (p(i,j))_{i,j \in S}$ définie par

$$p(i,j) = \mathbb{P}(X_1 = j \,|\, X_0 = i).$$

Définition 9.3. *La matrice $\mathbf{P}$ est appelée* ***matrice de transition*** *de la chaîne, et les probabilités $p(i,j)$ sont appelées* ***probabilités de transition*** *(de i à j).*

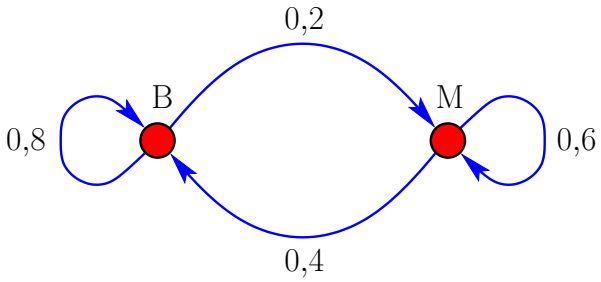


FIGURE 9.1: La représentation graphique de la chaîne de l'exemple 9.1.

Lemme 9.1. *Une matrice de transition est caractérisée par les deux propriétés suivantes :*

1. $p(i,j) \geq 0$, $\forall i,j \in S$;
2. $\sum_{j\in S} p(i,j) = 1$, $\forall i \in S$.

Une matrice possédant ces deux propriétés est appelée une ***matrice stochastique****.*

Démonstration. Exercice élémentaire. □

Définition 9.4. *Soit $\boldsymbol{\mu} = (\mu(i))_{i\in S}$ une mesure de probabilité sur S et $\mathbf{P}$ une matrice stochastique. La chaîne de Markov $(\mathbf{P}, \boldsymbol{\mu})$ est la chaîne de Markov (homogène dans le temps) de matrice de transition $\mathbf{P}$ et de loi initiale $\boldsymbol{\mu}$, c'est-à-dire telle que $\mathbb{P}(X_0 = i) = \mu(i)$, pour tout $i \in S$. On écrira simplement $X \sim (\mathbf{P}, \boldsymbol{\mu})$.*

Dans la suite, nous utiliserons les notations suivantes : la loi de la chaîne de Markov $(\mathbf{P}, \boldsymbol{\mu})$ sera notée $\mathbb{P}_{\boldsymbol{\mu}}$, et l'espérance correspondante $\mathbb{E}_{\boldsymbol{\mu}}$. En particulier, lorsque la loi initiale est concentrée sur un état $i \in S$, c'est-à-dire lorsque $\boldsymbol{\mu} = \boldsymbol{\delta}_i \equiv (\delta_{i,j})_{j\in S}$, nous écrirons simplement $\mathbb{P}_i$ et $\mathbb{E}_i$.

Exemple 9.1. Après une longue collecte de données, Robinson a conçu le modèle suivant pour précrire le temps qu'il fera sur son île :

$$S = \{\text{beau temps, mauvais temps}\}, \qquad \text{et} \qquad \mathbf{P} = \begin{pmatrix} 0{,}8 & 0{,}2 \\ 0{,}4 & 0{,}6 \end{pmatrix}.$$

La matrice $\mathbf{P}$ 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 9.1.

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 $Y_n = (X_n, X_{n-1})$. En effet, la connaissance du couple $Y_n = (X_n, X_{n-1})$ détermine X_n, et donc il ne reste plus qu'à prédire X_{n+1}, dont la probabilité est fonction uniquement de X_n et X_{n-1}. Les détails sont laissés en exercice. ♢

La matrice $\mathbf{P}$ 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

$$p_n(i,j) = \mathbb{P}_i(X_n = j).$$

Alors, pour tout $n \geq 1$,

$$\begin{aligned} p_n(i,j) &= \mathbb{P}_i(X_n = j) \\ &= \sum_{k \in S} \mathbb{P}_i(X_n = j, X_{n-1} = k) \\ &= \sum_{k \in S} \mathbb{P}_i(X_n = j \mid X_{n-1} = k)\, \mathbb{P}_i(X_{n-1} = k) \\ &= \sum_{k \in S} \mathbb{P}_k(X_1 = j)\, \mathbb{P}_i(X_{n-1} = k) \\ &= \sum_{k \in S} p(k,j) p_{n-1}(i,k). \end{aligned}$$

Cette relation est connue sous le nom d'**équation de Chapman-Kolmogorov**. On en déduit facilement le résultat fondamental suivant.

Théorème 9.1. *La **matrice de transition en n pas**, $\mathbf{P}_n = (p_n(i,j))_{i,j \in S}$, est donnée par la $n^{ème}$ puissance de la matrice de transition $\mathbf{P}$,*

$$\mathbf{P}_n = \mathbf{P}^n.$$

Démonstration. On peut réécrire l'équation de Chapman-Kolmogorov sous la forme

$$(\mathbf{P}_n)_{ij} = \sum_{k \in S} (\mathbf{P}_{n-1})_{ik} (\mathbf{P})_{kj} = (\mathbf{P}_{n-1}\mathbf{P})_{ij}.$$

En particulier, $\mathbf{P}_n = \mathbf{P}_{n-1}\mathbf{P} = \mathbf{P}_{n-2}\mathbf{P}^2 = \cdots = \mathbf{P}^n$. □

Il suit que l'on peut facilement exprimer la loi de la chaîne au temps n à partir de la loi de la chaîne au temps 0.

Corollaire 9.1. *Soit $X \sim (\mathbf{P}, \boldsymbol{\mu}_0)$. Alors, la loi de la chaîne au temps n, $\mu_n(i) = \mathbb{P}_{\boldsymbol{\mu}_0}(X_n = i)$, $i \in S$, est donnée par*

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_0 \mathbf{P}^n.$$

Démonstration.

$$\begin{aligned} \mu_n(i) = \mathbb{P}_{\boldsymbol{\mu}_0}(X_n = i) &= \sum_{j \in S} \mathbb{P}_{\boldsymbol{\mu}_0}(X_n = i \mid X_0 = j) \mathbb{P}_{\boldsymbol{\mu}_0}(X_0 = j) \\ &= \sum_{j \in S} p_n(j,i) \mu_0(j) = (\boldsymbol{\mu}_0 \mathbf{P}^n)_i. \end{aligned}$$

□

Nous nous intéresserons principalement à deux classes particulières, mais très importantes, de chaînes de Markov : les chaînes irréductibles et les chaînes absorbantes.

Définition 9.5. *Soit $\mathbf{P}$ 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 $n \geq 0$ tel que $p_n(i,j) > 0$.*
- *Un état $i \in S$ est **absorbant** si $p(i,i) = 1$.*
- *$\mathbf{P}$ est **irréductible** si, pour tout $i, j \in S$, on a $i \to j$.*
- *$\mathbf{P}$ 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 $\mathbf{P}$, on dira que X est irréductible, resp. absorbante, lorsque $\mathbf{P}$ est irréductible, resp. absorbante.

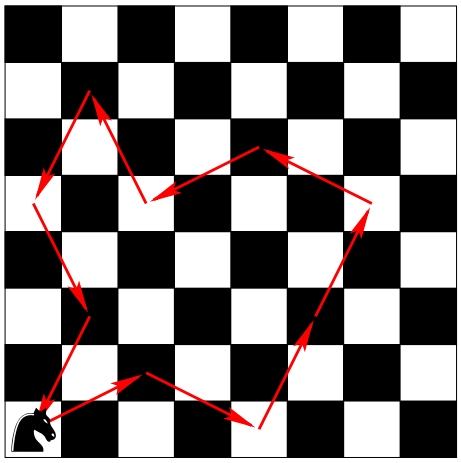


FIGURE 9.2: Quel est le nombre moyen de pas nécessaires pour que le cavalier se déplaçant au hasard sur l'échiquier se retrouve à son point de départ ?

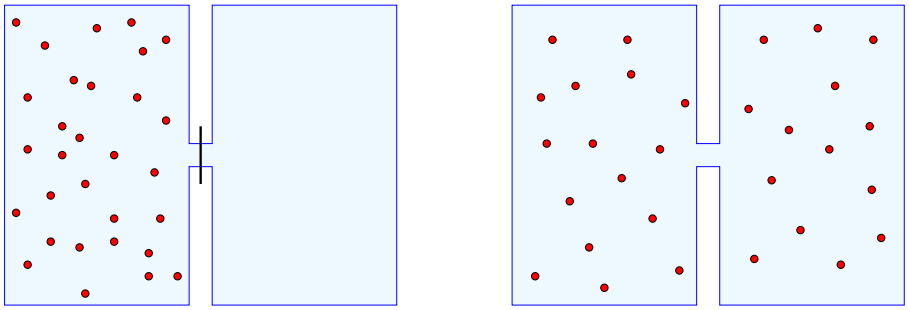


FIGURE 9.3: Au début de l'expérience, toutes les molécules du gaz sont confinées dans le récipient de gauche. Lorsque l'on retire la paroi séparant les deux récipients, les molécules se répartissent uniformément dans tout le volume disponible. Comment une telle irréversibilité peut-elle être compatible avec la réversibilité des équations d'évolution microscopiques ?

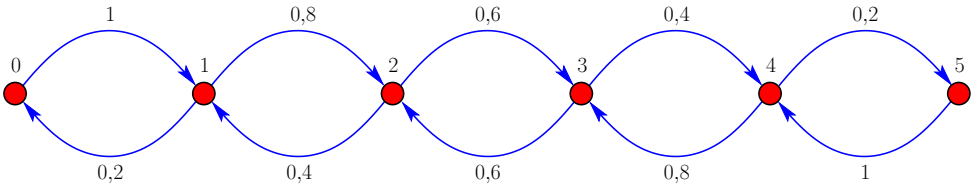


FIGURE 9.4: La représentation graphique du modèle des urnes d'Ehrenfest.

Par la suite, on notera $n(i,j) = \inf\{n \geq 1 \,:\, p_n(i,j) > 0\}$ le nombre minimal de pas permettant de passer de i à j avec probabilité positive ; en particulier, $n(i,j) < \infty$ si et seulement si $i \to j$.

Exemple 9.2. On positionne un cavalier sur une des cases d'un échiquier (Fig. 9.2). 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). 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 !), et on développera (Théorème 9.5 et exercices) des méthodes permettant de répondre très facilement à ce type de questions. ♢

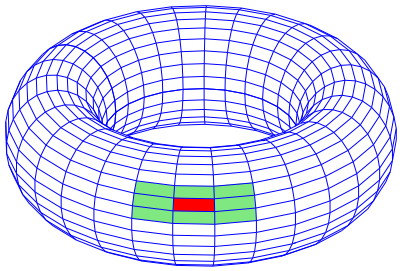


FIGURE 9.5: Une grille 30×30 enroulée en un tore. Chaque case possède 8 voisines.

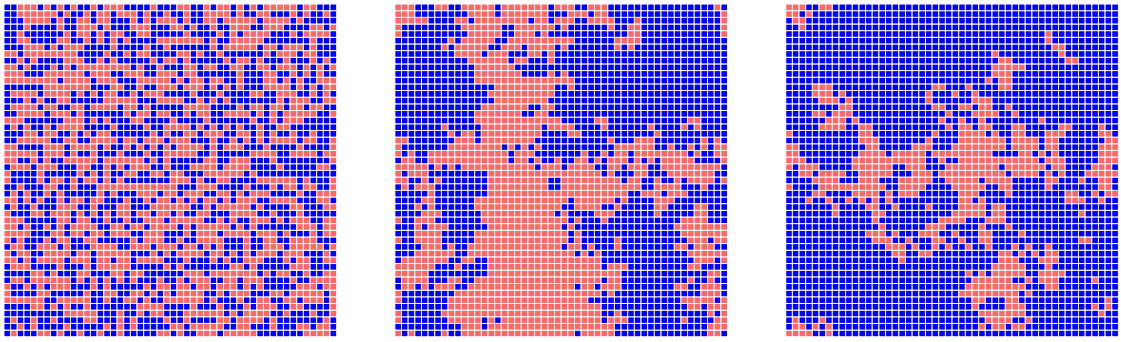


FIGURE 9.6: Le modèle du votant (Exemple 9.4), pour $k = 2$, sur une grille 50×50 (représentée « à plat »). Gauche : état initial ; milieu : après 1 000 000 de pas ; droite : après 10 000 000 de pas.

Exemple 9.3. **Le modèle des urnes d'Ehrenfest.** Ce modèle a été introduit par Paul et Tatiana Ehrenfest [1, 2] 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. 9.3). Leur modèle, très simplifié, de cette situation peut être formulé comme suit. On considère 2 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 9.4. X est clairement irréductible. ♢

Exemple 9.4 (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 (Fig. 9.5). La dynamique est la suivante : à chaque pas,

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.

1. Paul Ehrenfest (1880, Vienne – 1933, Amsterdam), physicien théoricien autrichien.
2. Tatiana Alexeyevna Afanaseva (1876, Kiev – 1964, Leiden), mathématicienne russe et danoise.

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).

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 au consensus.

La figure 9.6 montre l'état initial de la chaîne, et deux états ultérieurs. Nous démontrerons plus tard 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 ! ♢

Dans la suite de ce chapitre, nous allons étudier plus en détails les chaînes absorbantes et irréductibles.

9.2 Chaînes de Markov absorbantes

Soit $\mathbf{P}$ une matrice stochastique absorbante, et notons $\mathcal{A} = \{i \in S : p(i,i) = 1\}$ 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,

$$\mathbf{P} = \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}.$$

Si $|S| = m$ et $|\mathcal{A}| = r$, $\mathbf{Q}$ est donc une matrice $(m-r) \times (m-r)$, $\mathbf{R}$ une matrice $(m-r) \times r$, et $\mathbf{1}$ la matrice identité $r \times r$.

Lemme 9.2. *Soit $\mathbf{P}$ une matrice de transition sous sa forme canonique. Alors, pour tout $n \geq 1$,*

$$\mathbf{P}^n = \begin{pmatrix} \mathbf{Q}^n & (\mathbf{1} + \mathbf{Q} + \cdots + \mathbf{Q}^{n-1})\mathbf{R} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}.$$

Démonstration. On procède par récurrence.

$$\begin{aligned} \mathbf{P}^n = \mathbf{P}\mathbf{P}^{n-1} &= \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{1} \end{pmatrix} \begin{pmatrix} \mathbf{Q}^{n-1} & (\mathbf{1} + \mathbf{Q} + \cdots + \mathbf{Q}^{n-2})\mathbf{R} \\ \mathbf{0} & \mathbf{1} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{Q}^n & (\mathbf{1} + \mathbf{Q} + \cdots + \mathbf{Q}^{n-1})\mathbf{R} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}. \end{aligned}$$

□

Le résultat suivant montre qu'une chaîne de Markov absorbante finit toujours par se retrouver dans un état absorbant.

Proposition 9.1. *Soit $\mathbf{P}$ une matrice de transition mise sous forme canonique. Alors,*

$$\lim_{n \to \infty} \mathbf{Q}^n = \mathbf{0}.$$

Démonstration. Soient $i, j \notin \mathcal{A}$. On a

$$(\mathbf{Q}^n)_{ij} = \mathbb{P}_i(X_n = j) \leq \max_{i \in S} \mathbb{P}_i(X_n \notin \mathcal{A}).$$

Soient $M = \max_{j \in S} \min\{m : \mathbb{P}_j(X_m \in \mathcal{A}) > 0\}$ et

$$p = \min_{i \in S} \mathbb{P}_i(X_M \in \mathcal{A}) > 0.$$

On a alors,

$$\max_{j\in S} \mathbb{P}_j(X_M \notin \mathcal{A}) = 1 - p.$$

Par conséquent, on déduit de la propriété de Markov que, pour tout $i \in S$,

$$\begin{aligned}
\max_{i\in S} \mathbb{P}_i(X_n \notin \mathcal{A}) &= \max_{i\in S} \sum_{j\notin\mathcal{A}} \mathbb{P}_i(X_n \notin \mathcal{A} \,|\, X_M = j)\mathbb{P}_i(X_M = j) \\
&= \max_{i\in S} \sum_{j\notin\mathcal{A}} \mathbb{P}_j(X_{n-M} \notin \mathcal{A})\mathbb{P}_i(X_M = j) \\
&\leq \max_{j\in S} \mathbb{P}_j(X_{n-M} \notin \mathcal{A}) \max_{i\in S} \mathbb{P}_i(X_M \notin \mathcal{A}) \\
&= (1-p) \max_{j\notin\mathcal{A}} \mathbb{P}_j(X_{n-M} \notin \mathcal{A}) \\
&\leq \cdots \\
&\leq (1-p)^{\lfloor \frac{n}{M} \rfloor},
\end{aligned}$$

et le résultat suit en prenant la limite $n \to \infty$. □

Corollaire 9.2. *Soit $\mathbf{P}$ la matrice de transition d'une chaîne de Markov absorbante, sous forme canonique. Alors la matrice $\mathbf{1} - \mathbf{Q}$ est inversible et son inverse est donné par*

$$\mathbf{N} = (\mathbf{1} - \mathbf{Q})^{-1} = \mathbf{1} + \mathbf{Q} + \mathbf{Q}^2 + \cdots.$$

Démonstration. Soit v un vecteur tel que $(\mathbf{1} - \mathbf{Q})v = 0$. Alors,

$$\mathbf{Q}^n v = \mathbf{Q}^{n-1}\mathbf{Q}v = \mathbf{Q}^{n-1}v,$$

et donc $\mathbf{Q}^n v = v$, pour tout $n \geq 1$. On en déduit de la Proposition 9.1 que

$$v = \lim_{n\to\infty} \mathbf{Q}^n v = 0,$$

ce qui montre que la matrice $\mathbf{1} - \mathbf{Q}$ n'admet pas 0 comme valeur propre et est donc inversible. À présent, il suffit d'observer que

$$(\mathbf{1} - \mathbf{Q})(\mathbf{1} + \mathbf{Q} + \mathbf{Q}^2 + \cdots + \mathbf{Q}^n) = \mathbf{1} - \mathbf{Q}^{n+1},$$

et donc

$$\mathbf{1} + \mathbf{Q} + \mathbf{Q}^2 + \cdots + \mathbf{Q}^n = \mathbf{N}(\mathbf{1} - \mathbf{Q}^{n+1}),$$

ce qui implique que

$$\mathbf{N} = \lim_{n\to\infty} \sum_{i=0}^{n} \mathbf{Q}^i.$$

□

Définition 9.6. *La matrice $\mathbf{N}$ 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 9.2. *Soit $\mathbf{N}$ la matrice fondamentale de la chaîne et $\tau_{\mathcal{A}} = \min\{n \geq 0 : X_n \in \mathcal{A}\}$. Alors,*

1. $\mathbb{E}_i(\sum_{k\geq 0} \mathbf{1}_{\{X_k=j\}}) = \mathbf{N}_{ij}$, *pour tout* $i, j \notin \mathcal{A}$;
2. $\mathbb{E}_i(\tau_{\mathcal{A}}) = \sum_{j\notin\mathcal{A}} \mathbf{N}_{ij}$, *pour tout* $i \notin \mathcal{A}$;
3. $\mathbb{P}_i(X_{\tau_{\mathcal{A}}} = j) = (\mathbf{NR})_{ij}$, *pour tout* $i \notin \mathcal{A}, j \in \mathcal{A}$.

Démonstration. *1.* Soient i, j deux états non-absorbants. Alors,

$$\mathbb{E}_i(\sum_{n\geq 0} \mathbf{1}_{\{X_n=j\}}) = \sum_{n\geq 0} \mathbb{P}_i(X_n = j) = \sum_{n\geq 0} (\mathbf{P}^n)_{ij} = \sum_{n\geq 0} (\mathbf{Q}^n)_{ij} = \mathbf{N}_{ij}.$$

2. Il suffit d'observer que, par le point précédent,

$$\mathbb{E}_i(\tau_{\mathcal{A}}) = \mathbb{E}_i(\sum_{n\geq 0} \mathbf{1}_{\{X_n\notin\mathcal{A}\}}) = \sum_{j\notin\mathcal{A}} \mathbb{E}_i(\sum_{n\geq 0} \mathbf{1}_{\{X_n=j\}}) = \sum_{j\notin\mathcal{A}} \mathbf{N}_{ij}.$$

3. On a, pour tout $i \notin \mathcal{A}$ et $j \in \mathcal{A}$,

$$\begin{aligned}
\mathbb{P}_i(X_{\tau_{\mathcal{A}}} = j) &= \sum_{n\geq 1} \mathbb{P}_i(X_n = j, X_{n-1} \notin \mathcal{A}) \\
&= \sum_{n\geq 1} \sum_{k\notin\mathcal{A}} \mathbb{P}_i(X_n = j, X_{n-1} = k) \\
&= \sum_{n\geq 1} \sum_{k\notin\mathcal{A}} \mathbb{P}(X_n = j \mid X_{n-1} = k)\mathbb{P}_i(X_{n-1} = k) \\
&= \sum_{n\geq 1} \sum_{k\notin\mathcal{A}} \mathbf{R}_{kj}(\mathbf{Q}^{n-1})_{ik} \\
&= \sum_{n\geq 1} (\mathbf{Q}^{n-1}\mathbf{R})_{ij} \\
&= (\mathbf{NR})_{ij}.
\end{aligned}$$

□

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 9.7. *Soit f une fonction définie sur S et $\mathbf{P} = (p(i,j))_{i,j\in S}$ une matrice stochastique. On dit que f est une* ***fonction $\mathbf{P}$-harmonique*** *si*

$$f(i) = \sum_{j\in S} p(i,j)f(j), \qquad \forall i \in S,$$

c'est à dire, sous forme vectorielle, $\mathbf{f} = \mathbf{Pf}$, où $\mathbf{f} = (f(i))_{i\in S}$.

Théorème 9.3. *Soient $(X_n)_{n\geq 0}$ une chaîne de Markov absorbante de matrice de transition $\mathbf{P}$, $\tau_{\mathcal{A}}$ le temps d'absorption, et $\mathcal{A}$ l'ensemble des états absorbants. Alors, pour toute fonction f $\mathbf{P}$-harmonique et tout $i \in S$,*

$$f(i) = \sum_{j\in\mathcal{A}} f(j)\mathbb{P}_i(X_{\tau_{\mathcal{A}}} = j).$$

Démonstration. Puisque f est $\mathbf{P}$-harmonique,

$$\mathbf{f} = \mathbf{P}^n \mathbf{f}, \qquad \forall n \geq 1,$$

et, par conséquent,

$$\mathbf{f} = \lim_{n\to\infty} \mathbf{P}^n \mathbf{f} = \begin{pmatrix} \mathbf{0} & \mathbf{NR} \\ \mathbf{0} & \mathbf{1} \end{pmatrix} \mathbf{f},$$

et on conclut à l'aide du Théorème 9.2. □

Ce théorème peut se révéler particulièrement utile, dans certaines circonstances.

Exemple 9.5. Retournons au modèle introduit dans l'Exemple 9.4. On considère une grille $n \times n$, et k couleurs, notées $\{1, \dots, k\}$. On note $\mathbf{P}$ la matrice de transition associée. La fonction f donnant la fraction de cases de couleur 1 dans la configuration est $\mathbf{P}$-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,

$$\sum_{j\in S} p(i,j)(N_1(j) - N_1(i)) = \sum_{\substack{(x,y) \\ \text{voisins}}} \frac{1}{8n^2} \left(\mathbf{1}_{\{i(x)\neq 1, i(y)=1\}} - \mathbf{1}_{\{i(x)=1, i(y)\neq 1\}} \right),$$

où $i(x)$ est la couleur de la case x dans la configuration i. La dernière somme est nulle, puisque chaque contribution positive dûe 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 $\mathbf{P}$-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 précédent implique donc que

$$\rho = f(i_0) = \sum_{\ell=1}^{k} f(a_\ell) \mathbb{P}_{i_0}(X_{\tau_{\mathcal{A}}} = a_\ell).$$

Or, $f(a_1) = 1$ (puisque toutes les cases de a_1 sont de couleur 1), et $f(a_\ell) = 0$ pour $\ell = 2, \dots, k$. 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 difficile à obtenir directement à partir du point 3 du Théorème 9.2. ♢

Une remarque s'impose avant de conclure cette section. Considérons la chaîne de Markov de la Figure 9.7 (gauche). 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 vert). 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 représentée sur la Figure 9.7 (droite), et celle-ci est absorbante.

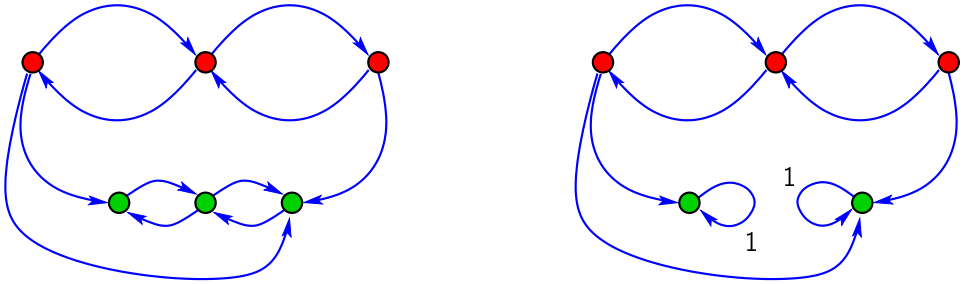


FIGURE 9.7: Une fois la chaîne entrée dans un des états représentés en vert, elle ne peut plus retourner vers les états représentés en rouge (gauche). Ce type de chaîne peut être étudié de la même façon que les chaînes absorbantes, en rendant les points d'entrée de la sous-chaîne absorbants (droite).

9.3 Chaînes de Markov irréductibles

Dans cette section, nous allons nous intéresser au cas des chaînes de Markov irréductibles. La terminologie suivante va se révéler utile.

Définition 9.8.
- ▷ *Un état $i \in S$ est* ***récurrent*** *si $\mathbb{P}_i(\exists n \geq 1, X_n = i) = 1$. 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 9.3. *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 en particulier la conséquence immédiate suivante : pour tout état transient $j \in S$, $\lim_{n\to\infty} \mathbb{P}_i(X_n = j) = 0$ pour tout $i \in S$.

Démonstration. De façon similaire à ce que l'on a fait dans le cas des marches aléatoires, on introduit les fonctions génératrices

$$\mathbb{G}_{ij}(s) = \sum_n s^n p_n(i,j), \qquad \mathbb{H}_{ij}(s) = \sum_n s^n h_n(i,j),$$

où $h_n(i,j) = \mathbb{P}_i(X_1 \neq j, X_2 \neq j, \ldots, X_{n-1} \neq j, X_n = j)$. Notons que $\mathbb{H}_{ij}(1) = \mathbb{P}_i(\exists n \geq 1, X_n = j)$. En procédant exactement comme dans le Lemme 4.2, 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 $\mathbb{H}_{ij}(1) \leq 1$. □

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 9.4. *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, $\rho_i = \mathbb{E}_i(T_i)$ avec*

$$T_i = \min\{n \geq 1 \,:\, X_n = i\},$$

est fini pour tout $i \in S$. 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 9.3

$$1 = \lim_{n\to\infty} \sum_{j\in S} p_n(i,j) = \sum_{j\in S} \lim_{n\to\infty} p_n(i,j) = 0,$$

puisque $\lim_{n\to\infty} p_n(i,j) = 0$ dès que j est transient.

On se souvient que $n(i,j) = \min\{n \geq 1 \,:\, p_n(i,j) > 0\}$. Montrons à présent que si $i \to j$ et $j \to i$, et que i est récurrent, alors j est également récurrent. Puisque $n(i,j) < \infty$ et $n(j,i) < \infty$, on a

$$\begin{aligned}\sum_{n\geq 1} p_n(j,j) &\geq \sum_{n\geq n(j,i)+n(i,j)+1} p_{n(j,i)}(j,i)\, p_{n-n(j,i)-n(i,j)}(i,i)\, p_{n(i,j)}(i,j) \\ &= p_{n(j,i)}(j,i)\, p_{n(i,j)}(i,j) \sum_{n\geq 1} p_n(i,i) = \infty,\end{aligned}$$

et la première affirmation est démontrée.

Pour montrer la seconde, on note que l'irréductibilité de la chaîne et la finitude de S impliquent que $n(i) = \max_{j\in S} n(j,i) < \infty$, et $p = \min_{j\in S} p_{n(j,i)}(j,i) > 0$. On a alors, avec les notations $M = \lfloor n/n(i) \rfloor$ et $k_0 = i$,

$$\begin{aligned}\mathbb{P}_i(T_i \geq n) &\leq \sum_{k_1\neq i,\dots,k_M\neq i} \prod_{\ell=1}^{M} \mathbb{P}_{k_{\ell-1}}(X_{n(k_{\ell-1},i)} = k_\ell) \\ &= \sum_{k_1\neq i,\dots,k_{M-1}\neq i} \prod_{\ell=1}^{M-1} \mathbb{P}_{k_{\ell-1}}(X_{n(k_{\ell-1},i)} = k_\ell) \\ &\qquad\qquad \times \mathbb{P}_{k_{M-1}}(X_{n(k_{M-1},i)} \neq i) \\ &\leq (1-p) \sum_{k_1\neq i,\dots,k_{M-1}\neq i} \prod_{\ell=1}^{M-1} \mathbb{P}_{k_{\ell-1}}(X_{n(k_{\ell-1},i)} = k_\ell) \\ &\leq \cdots \leq (1-p)^M.\end{aligned}$$

Par conséquent, on a bien $\rho_i = \mathbb{E}_i(T_i) = \sum_{n\geq 1} \mathbb{P}_i(T_i \geq n) < \infty$. □

Lemme 9.5. *Soit X une chaîne de Markov irréductible sur un espace d'états S fini. Alors, pour tout $i, j \in S$,*

$$\mathbb{P}_j(\exists n \geq 1, X_n = i) = 1.$$

Démonstration. Soient $i \neq j$ deux états. Manifestement, si $X_0 = i$ et $X_{n(i,j)} = j$, alors, $X_k \neq i$, pour tout $1 \leq k \leq n(i,j)$ (sinon $n(i,j)$ ne serait pas minimal). On a donc

$$\begin{aligned}\mathbb{P}_i(X_n \neq i, \forall n \geq 1) &\geq \mathbb{P}_i(X_n \neq i, \forall n \geq 1, X_{n(i,j)} = j) \\ &= \mathbb{P}_i(X_n \neq i, \forall n > n(i,j), X_{n(i,j)} = j) \\ &= p_{n(i,j)}(i,j)\, \mathbb{P}_j(X_n \neq i, \forall n \geq 1).\end{aligned}$$

On sait du Lemme 9.4 que X est récurrente, et par conséquent, le membre de gauche est nul. Puisque $p_{n(i,j)}(i,j) > 0$ par construction, on conclut que $\mathbb{P}_j(X_n \neq i, \forall n > 1) = 0$. □

9.3.1 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 qu'une telle convergence ait lieu, c'est-à-dire que, pour un certain $i \in S$, il existe un vecteur $\boldsymbol{\pi}$ 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 9.9. *Un vecteur $\boldsymbol{\pi} = (\pi(i))_{i\in S}$ est appelé* ***distribution stationnaire*** *associée à la matrice de transition X si*

1. *$\pi(j) \geq 0$ pour tout $j \in S$, et $\sum_{j\in S} \pi(j) = 1$;*
2. *$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$.*

La raison derrière cette terminologie est la suivante : si $X \sim (\mathbf{P}, \boldsymbol{\pi})$, alors il suit du Corollaire 9.1 que les probabilités d'occupation au temps n sont données par

$$\boldsymbol{\pi}\mathbf{P}^n = (\boldsymbol{\pi}\mathbf{P})\mathbf{P}^{n-1} = \boldsymbol{\pi}\mathbf{P}^{n-1} = \cdots = \boldsymbol{\pi}.$$

On voit donc que la distribution $\boldsymbol{\pi}$ 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 et une seule distribution stationnaire. Pour ce faire, introduisons, pour chaque $k \in S$, le vecteur $\boldsymbol{\gamma}^k = (\gamma^k(i))_{i\in S}$ 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 9.4. *Soit $\mathbf{P}$ irréductible sur S fini. Alors, pour tout $k \in S$,*

(i) $\gamma^k(k) = 1$;

(ii) $\boldsymbol{\gamma}^k\mathbf{P} = \boldsymbol{\gamma}^k$;

(iii) $\sum_{i\in S} \gamma^k(i) = \rho_k$;

(iv) $0 < \gamma^k(i) < \infty$, pour tout $i \in S$.

En particulier, pour tout $k \in S$, le vecteur $\boldsymbol{\pi} = \boldsymbol{\gamma}^k / \sum_{i\in S} \gamma^k(i) = \boldsymbol{\gamma}^k/\rho_k$ est une distribution stationnaire.

Démonstration. *(i)* suit immédiatement de la définition.
(ii) D'une part, il suit du Lemme 9.4 qu'avec probabilité 1, $T_k < \infty$ et donc $X_0 = X_{T_k} = k$. D'autre part, l'événement $\{T_k \geq n\} = \{X_1 \neq k, X_2 \neq k, \ldots, X_{n-1} \neq k\}$ ne dépendant que de $X_1, \ldots, X_{n-1}$, la propriété de Markov au temps $n-1$ (et le fait que les deux membres sont nuls lorsque $i = k$) donne

$$\mathbb{P}_k(X_{n-1} = i, X_n = j, T_k \geq n) = \mathbb{P}_k(X_{n-1} = i, T_k \geq n)\, p(i,j), \qquad \forall i, j \in S.$$

On peut donc écrire

$$\begin{aligned}
\gamma^k(j) &= \mathbb{E}_k\Big(\sum_{n=0}^{T_k-1} \mathbf{1}_{\{X_n=j\}}\Big) = \mathbb{E}_k\Big(\sum_{n=1}^{T_k} \mathbf{1}_{\{X_n=j\}}\Big) \\
&= \mathbb{E}_k\Big(\sum_{n=1}^{\infty} \mathbf{1}_{\{X_n=j, T_k \geq n\}}\Big) = \sum_{n=1}^{\infty} \mathbb{P}_k(X_n = j, T_k \geq n) \\
&= \sum_{i\in S} \sum_{n=1}^{\infty} \mathbb{P}_k(X_{n-1} = i, X_n = j, T_k \geq n) \\
&= \sum_{i\in S} p(i,j) \sum_{n=1}^{\infty} \mathbb{P}_k(X_{n-1} = i, T_k \geq n) \\
&= \sum_{i\in S} p(i,j)\, \mathbb{E}_k\Big(\sum_{n=1}^{\infty} \mathbf{1}_{\{X_{n-1}=i, T_k \geq n\}}\Big) \\
&= \sum_{i\in S} p(i,j)\, \mathbb{E}_k\Big(\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n=i, T_k-1 \geq n\}}\Big) \\
&= \sum_{i\in S} p(i,j)\, \mathbb{E}_k\Big(\sum_{n=0}^{T_k-1} \mathbf{1}_{\{X_n=i\}}\Big) = \sum_{i\in S} p(i,j)\, \gamma^k(i)\,.
\end{aligned}$$

(iii) suit directement de la définition :

$$\sum_{i\in S} \gamma^k(i) = \sum_{i\in S} \mathbb{E}_k\Big(\sum_{n=0}^{T_k-1} \mathbf{1}_{\{X_n=i\}}\Big) = \mathbb{E}_k\Big(\sum_{n=0}^{T_k-1} \sum_{i\in S} \mathbf{1}_{\{X_n=i\}}\Big) = \mathbb{E}_k(T_k) = \rho_k.$$

(iv) D'une part, il suit du point précédent que $\gamma^k(i) \leq \sum_{j\in S} \gamma^k(j) = \rho_k < \infty$. 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

$$\gamma^k(i) = \sum_{j\in S} \gamma^k(j)\, p_{n(k,i)}(j,i) \geq \gamma^k(k)\, p_{n(k,i)}(k,i) = p_{n(k,i)}(k,i) > 0. \tag{9.1}$$

□

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 9.5. *Soit $\mathbf{P}$ une matrice stochastique irréductible sur un espace des états S fini. Alors, $\mathbf{P}$ possède une unique distribution stationnaire $\boldsymbol{\pi}$. De plus,*

$$\pi(i) = \frac{1}{\rho_i}, \qquad \forall i \in S,$$

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. Soit $\boldsymbol{\lambda}$ un vecteur non-nul satisfaisant $\lambda(i) \geq 0$ pour tout $i \in S$ et $\boldsymbol{\lambda} = \boldsymbol{\lambda}\mathbf{P}$. L'argument utilisé en (9.1) implique que $\lambda(i) > 0$, pour tout $i \in S$; on peut donc supposer sans perte de généralité que $\lambda(k) = 1$. Alors, pour tout $j \in S$,

$$\begin{aligned}
\lambda(j) &= \sum_{i_1 \in S} \lambda(i_1) p(i_1, j) = \sum_{i_1 \in S \setminus \{k\}} \lambda(i_1) p(i_1, j) + \lambda(k) p(k, j) \\
&= \sum_{i_1, i_2 \in S \setminus \{k\}} \lambda(i_2) p(i_2, i_1) p(i_1, j) + \Big(p(k, j) + \sum_{i_1 \in S \setminus \{k\}} p(k, i_1) p(i_1, j) \Big) \\
&= \dots \\
&= \sum_{i_1, \dots, i_n \in S \setminus \{k\}} \lambda(i_n) p(i_n, i_{n-1}) \cdots p(i_1, j) \\
&\quad + \Big(p(k, j) + \sum_{i_1 \in S \setminus \{k\}} p(k, i_1) p(i_1, j) + \cdots + \sum_{i_1, \dots, i_{n-1} \in S \setminus \{k\}} p(k, i_{n-1}) \cdots p(i_2, i_1) p(i_1, j) \Big) \\
&\geq \mathbb{P}_k(X_1 = j, T_k \geq 1) + \mathbb{P}_k(X_2 = j, T_k \geq 2) + \cdots + \mathbb{P}_k(X_n = j, T_k \geq n).
\end{aligned}$$

Cette dernière expression convergeant vers $\gamma^k(j)$ lorsque $n \to \infty$, on en déduit que $\lambda(j) \geq \gamma^k(j)$, pour tout $j \in S$. Par conséquent, le vecteur $\boldsymbol{\mu} = \boldsymbol{\lambda} - \boldsymbol{\gamma}^k$ satisfait $\mu(i) \geq 0$ pour tout $i \in S$.

Par irréductibilité, pour chaque $i \in S$, $n(i, k) < \infty$. Comme $\boldsymbol{\mu}\mathbf{P} = \boldsymbol{\lambda}\mathbf{P} - \boldsymbol{\gamma}^k\mathbf{P} = \boldsymbol{\lambda} - \boldsymbol{\gamma}^k = \boldsymbol{\mu}$, on en conclut que

$$0 = \mu(k) = \sum_{j \in S} \mu(j) p_{n(i,k)}(j, k) \geq \mu(i) p_n(i, k),$$

ce qui implique $\mu(i) = 0$.

Passons à la seconde affirmation. La première partie et le théorème 9.4 impliquent que $\pi(k) = \gamma^k(k) / \sum_{i \in S} \gamma^k(i) = 1/\rho_k$. La conclusion suit, puisque le choix de l'état k est arbitraire. □

9.3.2 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 $\mathbf{P} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, 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 pas. Le problème ici est que la chaîne de Markov X a un comportement périodique.

Définition 9.10.
- ▷ *Le nombre* $d(i) = \operatorname{pgcd}\{n : p_n(i, i) > 0\}$ *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 9.4 montre qu'une chaîne de Markov X sur S est ergodique si et seulement si elle est irréductible et apériodique.

Lemme 9.6. *Soit X une chaîne de Markov irréductible et apériodique. Alors, il existe $N < \infty$ tel que, pour tout $i, j \in S$,*

$$p_n(i, j) > 0, \qquad \forall n \geq N.$$

Démonstration. Soit $j \in S$. Par apériodicité, il existe une suite de temps $t_1, t_2, \dots, t_\ell$ ayant 1 pour plus grand diviseur commun, et tels que $p_{t_k}(j, j) > 0$, pour tout $1 \leq k \leq \ell$. On peut

alors montrer qu'il suit du Théorème de Bézout [3] qu'il existe un entier $M = M(j)$ tel que tout nombre entier $m \geq M(j)$ peut se décomposer comme $m = \sum_{k=1}^{\ell} a_k t_k$, pour une suite $a_1, \ldots, a_\ell$ d'entiers positifs. Par conséquent, on a

$$p_m(j,j) \geq \prod_{k=1}^{\ell} (p_{t_k}(j,j))^{a_k} > 0, \qquad \forall m \geq M(j).$$

Soit $i \in S$, $i \neq j$. Par irréductibilité, $n(i,j) < \infty$, et donc

$$p_m(i,j) \geq p_{n(i,j)}(i,j) p_{m-n(i,j)}(j,j) > 0, \qquad \forall m \geq M(j) + n(i,j) \equiv M'(i,j).$$

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 9.6. *Soit $\mathbf{P}$ irréductible et apériodique sur un espace d'états S fini et $\boldsymbol{\mu}$ une mesure de probabilité sur S. Alors,*

$$\lim_{n \to \infty} \boldsymbol{\mu} \mathbf{P}^n = \boldsymbol{\pi},$$

où $\boldsymbol{\pi}$ est l'unique distribution stationnaire associée à $\mathbf{P}$.

En particulier, si $X \sim (\mathbf{P}, \boldsymbol{\mu})$, avec $\mathbf{P}$ comme ci-dessus, les Théorèmes 9.5 et 9.6 impliquent que $\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\mu}}(X_n = i) = 1/\rho_i$, pour tout $i \in S$.

Remarque 9.1. *On peut vérifier (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 $Y^{(r)}$, $0 \leq r < d$, définies par $Y_n^{(r)} = X_{nd+r}$ sont apériodiques, et qu'on peut donc leur appliquer le théorème.*

Démonstration. Soient X_n et Y_n 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, pour tout $i, j, k, l \in S$, il suit de l'indépendance de X et Y que

$$\begin{aligned} p_n((i,j),(k,l)) &= \mathbb{P}(Z_n = (k,l) \,|\, Z_0 = (i,j)) \\ &= \mathbb{P}(X_n = k \,|\, X_0 = i)\mathbb{P}(Y_n = l \,|\, Y_0 = j) = p_n(i,k) p_n(j,l), \end{aligned}$$

et, les chaînes X et Y étant irréductibles et apériodiques, il existe N tel que

$$p_n(i,k) p_n(j,l) > 0,$$

pour tout $n \geq N$ (voir le Lemme 9.6).

Notons $\mathbb{P}_{(i,j)}$ la loi de la chaîne Z partant de $Z_0 = (i,j)$. Fixons $s \in S$, et introduisons $T = \min\{n \geq 1 : Z_n = (s,s)\}$. Z étant irréductible, $\mathbb{P}_{(i,j)}(T < \infty) = 1$, pour tout $i, j \in S$. L'observation cruciale est que, pour tout $m \geq 0$, les lois de X_{T+m} et Y_{T+m} sont identiques, puisqu'elles ne dépendent que de s et m, et de la matrice de transition commune de X et Y. On peut donc écrire

$$\begin{aligned} p_n(i,k) &= \mathbb{P}_{(i,j)}(X_n = k) \\ &= \mathbb{P}_{(i,j)}(X_n = k, T \leq n) + \mathbb{P}_{(i,j)}(X_n = k, T > n) \\ &= \mathbb{P}_{(i,j)}(Y_n = k, T \leq n) + \mathbb{P}_{(i,j)}(X_n = k, T > n) \\ &\leq \mathbb{P}_{(i,j)}(Y_n = k) + \mathbb{P}_{(i,j)}(T > n) \\ &= p_n(j,k) + \mathbb{P}_{(i,j)}(T > n). \end{aligned}$$

3. Théorème de Bézout : si $x_1, \ldots, x_m \in \mathbb{N}^*$ sont tels que $\operatorname{pgcd}(x_1, \ldots, x_m) = d$, alors, pour tout $n \geq 0$, $\exists a_1, \ldots, a_m \in \mathbb{Z}$ tels que $a_1 x_1 + \cdots + a_m x_m = nd$. De plus, si $n \geq x_1 \cdots x_m$, $a_1, \ldots, a_m$ peuvent être choisis tous positifs.

On obtient donc

$$|p_n(i,k) - p_n(j,k)| \leq \mathbb{P}_{(i,j)}(T > n) \overset{n\to\infty}{\longrightarrow} 0,$$

pour tout $i, j, k \in S$. On en déduit que

$$\pi_k - p_n(j,k) = \sum_{i\in S} \pi_i (p_n(i,k) - p_n(j,k)) \overset{n\to\infty}{\longrightarrow} 0,$$

et donc

$$\begin{aligned}\lim_{n\to\infty} \left|\sum_{j\in S} \mu(j) p_n(j,i) - \pi(i)\right| &= \lim_{n\to\infty} \left|\sum_{j\in S} \mu(j)(p_n(j,i) - \pi(i))\right| \\ &\leq \sum_{j\in S} \mu(j) \lim_{n\to\infty} |p_n(j,i) - \pi(i)| = 0.\end{aligned}$$

□

9.3.3 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 l'état stationnaire), 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 X_n, $-\infty < n < \infty$, une chaîne de Markov irréductible, telle que la loi de X_n soit donnée par π pour tout $n \in \mathbb{Z}$. On définit la **chaîne renversée** Y par

$$Y_n = X_{-n}, n \in \mathbb{Z}.$$

Définition 9.11. *La chaîne X est **réversible** (à l'équilibre) si les matrices de transition de X et Y sont identiques.*

Théorème 9.7. *X est réversible si et seulement si la **condition d'équilibre local** est satisfaite :*

$$\pi(i)\, p(i,j) = \pi(j)\, p(j,i), \qquad \forall i, j \in S.$$

Démonstration.

$$\begin{aligned}\mathbb{P}(Y_{n+1} = j \,|\, Y_n = i) &= \mathbb{P}(X_{-n-1} = j \,|\, X_{-n} = i) \\ &= \mathbb{P}(X_{-n} = i \,|\, X_{-n-1} = j) \frac{\mathbb{P}(X_{-n-1} = j)}{\mathbb{P}(X_{-n} = i)} \\ &= p(j,i) \frac{\pi(j)}{\pi(i)}.\end{aligned}$$

□

Une façon d'interpréter cette formule 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 en sort 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 9.8. *Soit X une chaîne irréductible. S'il existe $\boldsymbol{\pi}$ tel que*

$$0 \leq \pi(i) \leq 1, \quad \sum_{i \in S} \pi(i) = 1, \quad \pi(i)\, p(i,j) = \pi(j)\, p(j,i) \text{ pour tout } i, j \in S,$$

alors la chaîne est réversible (à l'équilibre) et de distribution stationnaire $\boldsymbol{\pi}$.

Démonstration. Par la propriété d'équilibre local,

$$\sum_{j \in S} \pi(j)\, p(j,i) = \sum_{j \in S} \pi(i)\, p(i,j) = \pi(i),$$

ce qui montre que $\boldsymbol{\pi}$ est la distribution stationnaire de la chaîne. □

Ce dernier théorème permet, dans certaines situations, de déterminer beaucoup plus simplement la distribution stationnaire : si l'on parvient à trouver une distribution de probabilité sur S satisfaisant la condition d'équilibre local pour une chaîne irréductible, on est assuré que cette solution est bien la mesure stationnaire de la chaîne.

Exemple 9.6. Il est intuitivement clair que la chaîne de Markov du modèle d'Ehrenfest devrait être réversible à l'équilibre. Il est donc naturel d'essayer de trouver une distribution sur S satisfaisant la condition d'équilibre local. Dans le cas présent, cela revient à trouver un vecteur $\mathbf{m} = (m(0), \ldots, m(N))$ tel que, pour tout $0 \leq i \leq N-1$,

$$\frac{m(i+1)}{m(i)} = \frac{p(i,i+1)}{p(i+1,i)} = \frac{(N-i)/N}{(i+1)/N} = \frac{N-i}{i+1},$$

et $\sum_i m(i) = 1$. La mesure stationnaire est donc donnée par

$$m(k) = 2^{-N} \binom{N}{k}.$$

En particulier, on peut à présent aisément utiliser le Théorème 9.6 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 donné par

$$\mathbb{E}_N(T_N) = \frac{1}{m(N)} = \frac{2^N}{\binom{N}{N}} = 2^N \simeq 2^{10^{23}} \text{ secondes} \simeq 2^{10^{23}} \text{ âge de l'univers.}$$

D'un autre côté, le temps moyen de récurrence de l'état dans lequel chacune des deux urnes contient la moitié des boules est de

$$\mathbb{E}_{N/2}(T_{N/2}) = \frac{1}{m(N/2)} = \frac{2^N}{\binom{N}{N/2}} \simeq \sqrt{\tfrac{1}{2}\pi N} \simeq 40 \text{ secondes.}$$

Ceci résout 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 (bien plus grand que l'âge de l'univers !), que cela n'aura jamais lieu en pratique. ♢

Chapitre 10

Modèle de percolation

Dans ce chapitre, nous allons introduire un autre processus très important en théorie des probabilités : le modèle de **percolation**. Contrairement à la marche aléatoire et aux chaînes de Markov, 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 de $\mathbb{Z}^2$.

10.1 Définition

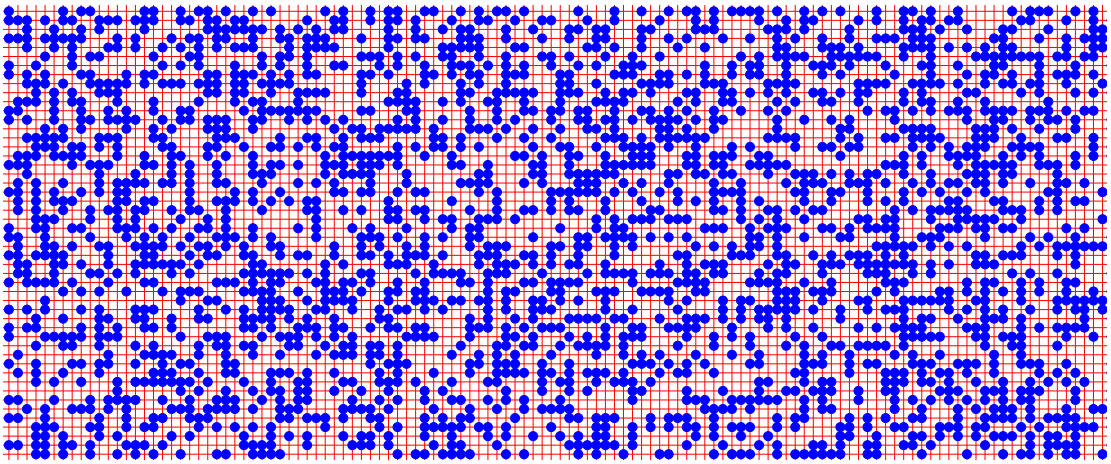
Soit $p \in [0,1]$, et soit $(X_i)_{i\in\mathbb{Z}^2}$ une famille de variables aléatoires indépendantes suivant une loi de Bernoulli de paramètre p, indicées par les sommets de $\mathbb{Z}^2$. On note $\mathbb{P}_p$ la loi de ce champ.

Un sommet i est dit **occupé** si $X_i = 1$ et **vide** si $X_i = 0$. On centre en chaque sommet occupé un disque de diamètre $1 < \rho < \sqrt{2}$. Deux sommets sont dits **connectés** s'ils appartiennent à la même composante de l'union de ces disques. Les composantes connexes maximales de sommets de $\mathbb{Z}^2$ sont appelées **amas**. Étant donné un sommet $x \in \mathbb{Z}^2$, on note $C(x)$ l'amas contenant x. On a représenté sur la Figure 10.1 trois réalisations de ce processus pour des valeurs diverses de p.

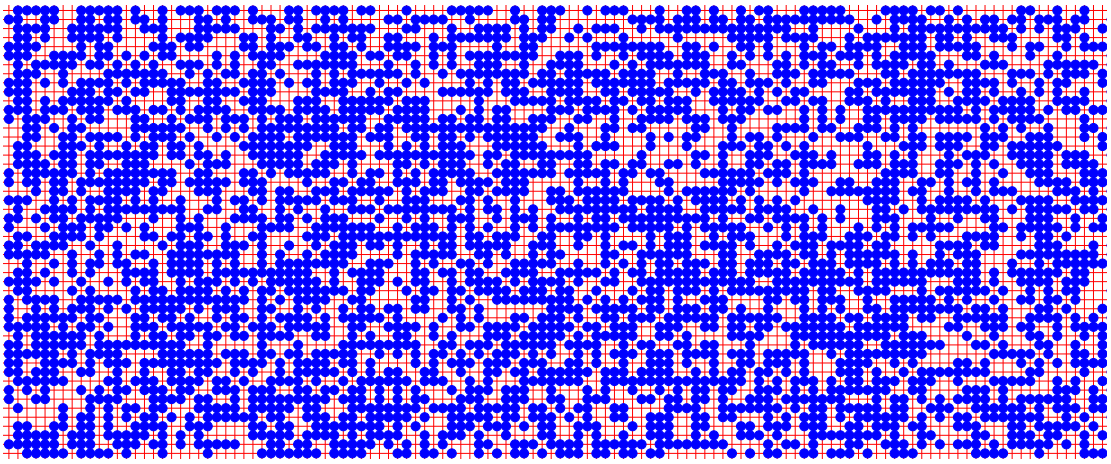
L'interprétation originelle de ce processus est comme modèle d'un matériau poreux. Un tel matériau contient un grand nombre de trous microscopiques. La question de base que l'on se pose alors est si cette porosité locale induit une porosité globale : si l'on plonge une pierre poreuse dans de l'eau, quelle est la probabilité que le centre de la pierre soit mouillé ? Dans le modèle de percolation, les trous correspondent aux disques placés sur les sommets occupés. La question de base peut alors se reformuler dans les termes suivants : existe-t-il un amas infini (l'eau pourrait alors se propager infiniment loin à travers ce dernier) ? Il y a de nombreuses autres interprétations bien entendu : comme modèle d'épidémie, de feu de forêt, etc. Ce modèle est devenu l'exemple classique pour modéliser des milieux aléatoires.

10.2 Transition de phase

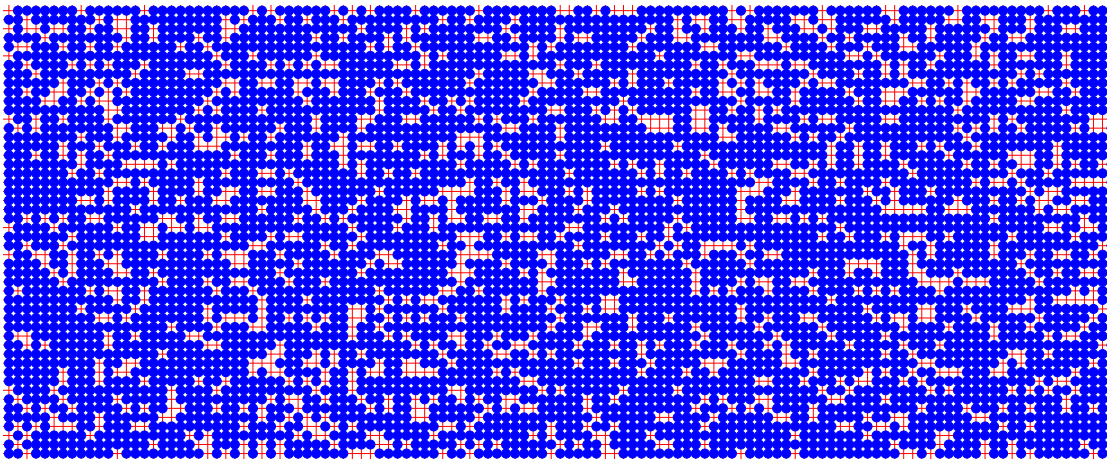
Soit $\Theta(p) = \mathbb{P}_p(|C(0)| = \infty)$, où $|A|$ représente la cardinalité de l'ensemble $A \subset \mathbb{Z}^2$. $\Theta(p)$ est donc la probabilité que de l'eau injectée à l'infini parvienne jusqu'à l'origine. Le résultat suivant est fondamental.



$p = 0{,}4$



$p = 0{,}6$



$p = 0{,}8$

FIGURE 10.1: Trois réalisations du processus de percolation.

Théorème 10.1. *1. Il existe $0 < p_c < 1$ tel que*

$$\Theta(p) = 0 \qquad \forall p < p_c,$$
$$\Theta(p) > 0 \qquad \forall p > p_c.$$

2. La probabilité qu'il existe (au moins) un amas infini est égale à 1 si $\Theta(p) > 0$, et 0 sinon.

Remarque 10.1. *1. Un argument de théorie ergodique permet de montrer qu'avec probabilité 1, il n'y a jamais plus d'un amas infini. Nous ne le ferons pas ici.*

2. La valeur exacte de p_c est inconnue, mais des simulations numériques montrent que $p_c \simeq 0{,}5928$.

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 l'état d'un nombre fini de sommets n'a pas d'influence sur sa réalisation. Par conséquent, il suit de la loi 0-1 de Kolmogorov que la probabilité qu'il existe au moins un amas infini a probabilité 0 ou 1. Or,

$$\mathbb{P}_p(\bigcup_{i\in\mathbb{Z}^2}\{|C(i)| = \infty\}) \geq \mathbb{P}_p(\{|C(0)| = \infty\}) > 0,$$

si $\Theta(p) > 0$, et donc $\mathbb{P}_p(\bigcup_{i\in\mathbb{Z}^2}\{|C(i)| = \infty\}) = 1$. Réciproquement,

$$\mathbb{P}_p(\bigcup_{i\in\mathbb{Z}^2}\{|C(i)| = \infty\}) \leq \sum_{i\in\mathbb{Z}^2}\mathbb{P}_p(\{|C(i)| = \infty\}),$$

et donc $\mathbb{P}_p(\bigcup_{i\in\mathbb{Z}^2}\{|C(i)| = \infty\}) = 0$ dès que $\mathbb{P}_p(\{|C(i)| = \infty\}) = \mathbb{P}_p(\{|C(0)| = \infty\}) = \Theta(p) = 0$.

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) $\Theta(p)$ est une fonction croissante de p.

(i) On appelle **chemin de longueur** n dans $\mathbb{Z}^2$ une suite $\gamma = (i_1, i_2, \ldots, i_n)$ de sommets tous distincts et tels que $\|i_k - i_{k-1}\|_2 = 1$, $k = 2, \ldots, n$. Soit $\mathcal{N}(n)$ l'ensemble des chemins de longueur n commençant en $i_1 = 0$, et $N(n)$ la cardinalité de cet ensemble. On vérifie facilement que $N(n) \leq 4^n$. En effet, lorsque l'on construit un tel chemin sommet par sommet, on a au plus 4 choix à chaque étape.

Soit à présent $N_o(n)$ le nombre de chemins de longueur n composés uniquement de sommets occupés (**chemins occupés**). Étant donné $\gamma \in \mathcal{N}(n)$, la probabilité que les sommets le constituant soient tous occupés est exactement donnée par $\prod_{i\in\gamma}\mathbb{P}_p(X_i = 1) = p^n$. Par conséquent,

$$\mathbb{E}_p(N_o(n)) = \mathbb{E}_p(\sum_{\gamma\in\mathcal{N}(n)}\mathbf{1}_{\{\gamma \text{ occupé}\}}) = \sum_{\gamma\in\mathcal{N}(n)}\mathbb{P}_p(\gamma \text{ occupé}) = p^n N(n) \leq (4p)^n.$$

Lorsque l'événement $\{|C(0)| = \infty\}$ est réalisé, il existe de tels chemins occupés de toutes les longueurs. On obtient donc, pour tout $n \geq 1$,

$$\mathbb{P}_p(|C(0)| = \infty) \leq \mathbb{P}_p(N_o(n) \geq 1) \leq \mathbb{E}_p(N_o(n)) \leq (4p)^n.$$

En laissant $n \to \infty$, on voit que $\mathbb{P}_p(|C(0)| = \infty) = 0$ dès que $p < \frac{1}{4}$.

(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]. On appelle **$\star$-circuit de longueur** n une suite

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.

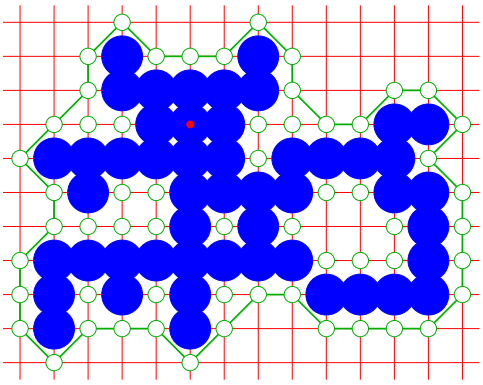


FIGURE 10.2: Lorsque $X_0 = 1$ mais que l'amas contenant l'origine est fini, il y a toujours un $\star$-circuit vide entourant l'origine (le point rouge). On a représenté par des cercles verts les sommets qui sont nécessairement vides si $C(0)$ est l'amas représenté en bleu.

de sommets $i_1, \dots, i_n$ tous distincts tels que $\|i_k - i_{k-1}\|_2 \leq \sqrt{2}$, $k = 2, \dots, n$, et $\|i_1 - i_n\|_2 \leq \sqrt{2}$. L'observation cruciale est que lorsque l'origine est occupée, mais que l'amas contenant l'origine est fini, il existe un $\star$-circuit composé entièrement de sommets vides (**$\star$-circuit vide**) et entourant l'origine (*cf.* Figure 10.2). Notons $N^\star(n)$ le nombre de $\star$-circuits de longueur n entourant l'origine. On peut à nouveau facilement borner leur nombre. En effet, le nombre de $\star$-circuits de longueur n contenant un sommet donné est inférieur à 8^n, puisqu'il y a exactement 8 sommets à distance au plus $\sqrt{2}$ d'un sommet donné. D'autre part, un $\star$-circuit de longueur n entourant l'origine intersecte nécessairement l'ensemble des sommets de coordonnées $(0, y)$ avec $0 < y < \frac{1}{2}n$. On obtient donc que $N^\star(n) \leq \frac{1}{2}n8^n$.

Soit $N^\star_{\mathrm{v}}(n)$ le nombre de tels $\star$-circuits entièrement composés de sommets vides. En procédant de la même façon qu'auparavant, la probabilité que tous les sommets d'un $\star$-circuit de longueur n donné soient vides est $(1-p)^n$. Par conséquent,

$$\mathbb{E}_p(N^\star_{\mathrm{v}}(n)) = (1-p)^n N^\star(n) < \tfrac{1}{2}n(8(1-p))^n.$$

Comme on l'a vu, lorsque $X_0 = 1$ et $|C(0)| < \infty$, il exite un entier n tel que $N^\star_{\mathrm{v}}(n) \geq 1$. À présent,

$$\begin{aligned}\mathbb{P}_p(X_0 = 1, |C(0)| < \infty) &\leq \mathbb{P}_p\Big(\bigcup_{n\geq 4}\{N^\star_{\mathrm{v}}(n) \geq 1\}\Big) \leq \sum_{n\geq 4}\mathbb{P}_p(N^\star_{\mathrm{v}}(n) \geq 1)\\ &\leq \sum_{n\geq 4}\mathbb{E}_p(N^\star_{\mathrm{v}}(n)) \leq \sum_{n\geq 4}\tfrac{1}{2}n(8(1-p))^n,\end{aligned}$$

et cette dernière quantité tend vers 0 lorsque $p \to 1$. Par conséquent, $\mathbb{P}_p(|C(0)| = \infty) = 1 - (1-p) - \mathbb{P}_p(X_0 = 1, |C(0)| < \infty) > 0$ pour tout p suffisamment proche de 1.

(iii) Il reste à montrer que $\Theta(p)$ est une fonction croissante de p. Soit $(Y_i)_{i\in\mathbb{Z}^2}$ une famille de variables aléatoires i.i.d. de loi uniforme sur $[0,1]$; on note $\widehat{\mathbb{P}}$ la loi de ce processus. Pour $p \in [0,1]$, on définit les variables aléatoires $(X_i^p)_{i\in\mathbb{Z}^2}$ par

$$X_i^p = \begin{cases} 1 & \text{si } Y_i \leq p,\\ 0 & \text{si } Y_i > p.\end{cases}$$

Un peu de réflexion montre que la loi de la famille $(X_i^p)_{i\in\mathbb{Z}^2}$ est précisément $\mathbb{P}_p$. L'intérêt de cette construction est que l'on peut définir *simultanément* tous les processus $(X_i^p)_{i\in\mathbb{Z}^2}$ pour $p \in [0,1]$ sur cet 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 $p' \geq p$, puisque $Y_i \leq p \implies Y_i \leq p'$, pour tout $p' \geq p$, et donc chaque sommet occupé dans X^p est nécessairement également occupé dans $X^{p'}$. On a donc, pour $0 \leq p \leq p' \leq 1$,

$$\begin{aligned}\Theta(p) &= \mathbb{P}_p(|C(0)| = \infty) = \widehat{\mathbb{P}}(|C(0)| = \infty \text{ dans } X^p) \\ &\leq \widehat{\mathbb{P}}(|C(0)| = \infty \text{ dans } X^{p'}) = \mathbb{P}_{p'}(|C(0)| = \infty) = \Theta(p').\end{aligned}$$

□

Chapitre 11

Le processus de Poisson

Nous allons à présent introduire un processus de nature différente, dont le domaine d'applicabilité est très important : le processus de Poisson. Dans le cadre qui va nous intéresser ici, celui-ci décrit la répartition aléatoire et uniforme de points sur la droite réelle positive. Il peut servir à modéliser par exemple : les appels téléphoniques arrivant dans une centrale, l'arrivée de particules sur un compteur Geiger, les temps d'arrivée de clients à une caisse, les temps d'occurrence de sinistres à dédommager par une compagnie d'assurance, etc.

11.1 Définition et propriétés élémentaires

Il y a trois façons naturelles de décrire un tel processus (*cf.* Fig 11.1) :

- ▷ On peut, tout d'abord, encoder une réalisation d'un tel processus par une collection $0 < T_1(\omega) < T_2(\omega) < \cdots$ de nombres réels positifs, correspondant à la position des points sur $\mathbb{R}^+$. Il est pratique de poser également $T_0 = 0$.
- ▷ Une seconde façon de coder une réalisation revient à donner, pour chaque intervalle de la forme $I = (t, t+s]$, le nombre de points $N_I(\omega)$ contenus dans l'intervalle. Si l'on utilise la notation simplifiée $N_t = N_{(0,t]}$, on aura alors $N_{(t,t+s]} = N_{t+s} - N_t$. La relation entre les variables aléatoires T_n et N_t est donc simplement

$$N_t(\omega) = \sup\{n \geq 0 \,:\, T_n(\omega) \leq t\}\,, \qquad T_n(\omega) = \inf\{t \geq 0 \,:\, N_t(\omega) \geq n\}\,.$$

- ▷ Une troisième façon naturelle d'encoder cette information est de considérer la suite $X_1(\omega), X_2(\omega), X_3(\omega), \ldots$ de nombres réels positifs correspondant aux distances successives entre deux points. La relation entre ces variables et les T_n est donnée par

$$X_k = T_k - T_{k-1}, \qquad T_k = \sum_{i=1}^{k} X_i.$$

Définition 11.1. *Soient $X_1, X_2, \ldots$ une suite de variables aléatoires satisfaisant $\mathbb{P}(X_k > 0) = 1$ pour tout $k \geq 1$. Soit $T_0 = 0$ et $T_n = \sum_{i=1}^{n} X_i$. Finalement, posons $N_t = \sup\{n \geq 0 \,:\, T_n \leq t\}$. Le processus $(N_t)_{t\geq 0}$ est appelé* **processus de comptage**.

Remarque 11.1. *On supposera toujours par la suite qu'un processus de comptage satisfait presque-sûrement $T_n \to \infty$ lorsque $n \to \infty$.*

On appelle souvent les variables aléatoires X_n les **temps d'attente** ou **durées de vie** du processus $(N_t)_{t\geq 1}$.

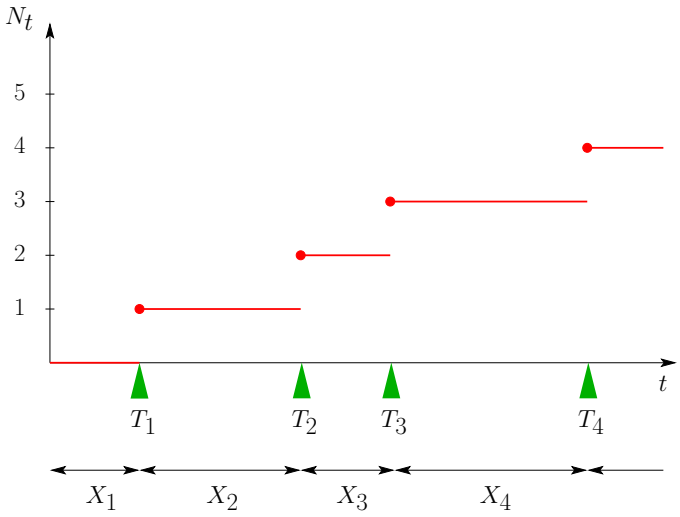


FIGURE 11.1: Une réalisation d'un processus de Poisson.

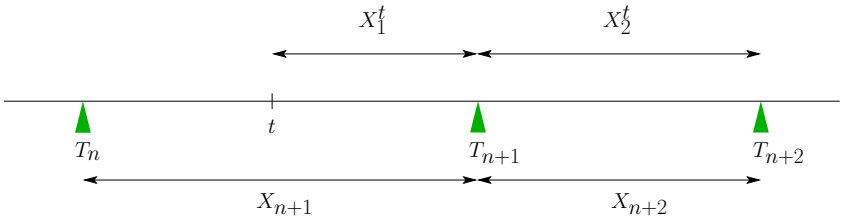


FIGURE 11.2: Définition des variables aléatoires X_k^t (lorsque $N_t = n$).

Le cas le plus simple, mais très important, est celui où les temps d'attente forment un processus i.i.d. : prenons l'exemple d'une lampe dont l'ampoule est changée instantanément dès qu'elle est défaillante. Dans ce cas, les durées de vie correspondent précisément à la durée pendant laquelle l'ampoule fonctionne. À chaque fois qu'une ampoule cesse de fonctionner et qu'elle est remplacée par une ampoule neuve, le système se retrouve dans le même état. On dit qu'il y a renouvellement.

Définition 11.2. *Un processus de comptage pour lequel les temps d'attente sont i.i.d. est appelé* ***processus de renouvellement****.*

Le processus de Poisson est l'exemple le plus important de processus de renouvellement.

Définition 11.3. *Un* ***processus de Poisson*** *d'****intensité*** λ *est un processus de renouvellement dont les durées de vie suivent une loi* $\exp(\lambda)$*. On note* $\mathbb{P}_\lambda$ *la loi du processus de Poisson d'intensité* λ*.*

Vérifions à présent deux propriétés tout à fait remarquables du processus de Poisson.

Théorème 11.1. *Soit* $(N_t)_{t\geq 0}$ *un processus de Poisson d'intensité* λ*. Alors, pour tout* $t, s \geq 0$*,*

1. $N_{t+s} - N_t$ *suit la même loi que* N_s*.*
2. $\forall\ 0 = t_0 < t_1 < t_2 < \cdots < t_n$*, les variables aléatoires* $(N_{t_{i+1}} - N_{t_i})_{i=1,\ldots,n-1}$ *sont indépendantes.*

Démonstration. Soit $t > 0$ fixé. Notons $X_1^t = T_{N_t+1} - t$ le temps restant après t jusqu'au point suivant du processus, $X_k^t = X_{N_t+k}$, $k \geq 2$, et $T_k^t = X_1^t + \cdots + X_k^t$, $k \geq 1$. Évidemment,

$$N_{t+s} - N_t = n \iff T_n^t \leq s < T_{n+1}^t.$$

Observons à présent que l'indépendance de X_{n+1} et de T_n implique que, pour tout $x > 0$,

$$\begin{aligned}
\mathbb{P}_\lambda(X_1^t > x \,|\, N_t = n)\mathbb{P}_\lambda(N_t = n) &= \mathbb{P}_\lambda(X_1^t > x, T_n \leq t < T_{n+1}) \\
&= \mathbb{P}_\lambda(T_n \leq t, X_{n+1} > t + x - T_n) \\
&= \int_0^t \mathrm{d}y \int_{t+x-y}^\infty \mathrm{d}z\, f_{(T_n, X_{n+1})}(y, z) \\
&= \int_0^t \mathrm{d}y\, f_{T_n}(y) \int_{t+x-y}^\infty \mathrm{d}z\, f_{X_{n+1}}(z) \\
&= \int_0^t \mathbb{P}_\lambda(X_{n+1} > t + x - y) f_{T_n}(y) \mathrm{d}y \\
&= e^{-\lambda x} \int_0^t \mathbb{P}_\lambda(X_{n+1} > t - y) f_{T_n}(y) \mathrm{d}y \\
&= e^{-\lambda x}\, \mathbb{P}_\lambda(T_n \leq t, X_{n+1} > t - T_n) \\
&= e^{-\lambda x}\, \mathbb{P}_\lambda(N_t = n). \qquad (11.1)
\end{aligned}$$

En procédant de la même façon, on voit que

$$\begin{aligned}
&\mathbb{P}_\lambda(X_1^t > x_1, X_2^t > x_2, \ldots, X_k^t > x_k \,|\, N_t = n) \\
&\qquad = \mathbb{P}_\lambda(T_n \leq t, X_{n+1} > t + x_1 - T_n, X_{n+2} > x_2, \ldots, X_{n+k} > x_k) / \mathbb{P}_\lambda(N_t = n) \\
&\qquad = \prod_{\ell=2}^k \mathbb{P}_\lambda(X_{n+\ell} > x_\ell)\, \mathbb{P}_\lambda(T_n \leq t, X_{n+1} > t + x_1 - T_n) / \mathbb{P}_\lambda(N_t = n) \\
&\qquad = e^{-\lambda(x_1 + \cdots + x_k)},
\end{aligned}$$

la seconde identité suivant de l'indépendance de $T_n, X_{n+1}, \ldots, X_{n+k}$, et la dernière identité de (11.1). On en déduit que, conditionnellement à $N_t = n$, les variables aléatoires X_k^t, $k \geq 1$, sont des variables aléatoires i.i.d. de loi $\exp(\lambda)$. Par conséquent, la loi conjointe des variables aléatoires T_k^t, $k \geq 1$, sous $\mathbb{P}_\lambda(\,\cdot \,|\, N_t = n)$ coïncide avec celle des variables aléatoires T_k, $k \geq 1$, sous $\mathbb{P}_\lambda$. On a donc

$$\begin{aligned}
\mathbb{P}_\lambda(N_{t+s} - N_t = k) &= \sum_{n \geq 0} \mathbb{P}_\lambda(N_{t+s} = n + k \,|\, N_t = n)\mathbb{P}_\lambda(N_t = n) \\
&= \sum_{n \geq 0} \mathbb{P}_\lambda(T_k^t \leq s < T_{k+1}^t \,|\, N_t = n)\mathbb{P}_\lambda(N_t = n) \\
&= \sum_{n \geq 0} \mathbb{P}_\lambda(T_k \leq s < T_{k+1})\mathbb{P}_\lambda(N_t = n) \\
&= \mathbb{P}_\lambda(T_k \leq s < T_{k+1}) \\
&= \mathbb{P}_\lambda(N_s = k).
\end{aligned}$$

Passons à la seconde affirmation. Les arguments ci-dessus montrent que la loi conjointe des variables aléatoires $N_s - N_t = \max\{n \geq 0 : T_n^t \leq s - t\}$ sous $\mathbb{P}_\lambda(\,\cdot \,|\, N_t = \ell)$ coïncide avec celle des variables aléatoires $N_{s-t} = \max\{n \geq 0 : T_n \leq s - t\}$ sous $\mathbb{P}_\lambda$. Posons $m_i = k_1 + \cdots + k_i$,

$i \geq 1$. On a alors

$$\begin{aligned}\mathbb{P}_\lambda(N_{t_{i+1}} - N_{t_i} = k_i, i = 0, \dots, n-1) \\ &= \mathbb{P}_\lambda(N_{t_{i+1}} - N_{t_i} = k_i, i = 1, \dots, n-1 \,|\, N_{t_1} = k_0)\, \mathbb{P}_\lambda(N_{t_1} = k_0) \\ &= \mathbb{P}_\lambda(N_{t_{i+1}} - N_{t_1} = m_i, i = 1, \dots, n-1 \,|\, N_{t_1} = k_0)\, \mathbb{P}_\lambda(N_{t_1} = k_0) \\ &= \mathbb{P}_\lambda(N_{t_{i+1}-t_1} = m_i, i = 1, \dots, n-1)\, \mathbb{P}_\lambda(N_{t_1} = k_0) \\ &= \mathbb{P}_\lambda(N_{t_2-t_1} = k_1, N_{t_{i+1}-t_1} - N_{t_2-t_1} = m_i - m_1, i = 2, \dots, n-1) \\ &\qquad \times \mathbb{P}_\lambda(N_{t_1} = k_0).\end{aligned}$$

De la même façon, puisque $t_{i+1} - t_1 - (t_2 - t_1) = t_{i+1} - t_2$,

$$\begin{aligned}\mathbb{P}_\lambda(N_{t_2-t_1} = k_1, N_{t_{i+1}-t_1} - N_{t_2-t_1} = m_i - m_1, i = 2, \dots, n-1) \\ &= \mathbb{P}_\lambda(N_{t_{i+1}-t_1} - N_{t_2-t_1} = m_i - m_1, i = 2, \dots, n-1 \,|\, N_{t_2-t_1} = k_1) \\ &\qquad \times \mathbb{P}_\lambda(N_{t_2-t_1} = k_1) \\ &= \mathbb{P}_\lambda(N_{t_{i+1}-t_2} = m_i - m_1, i = 2, \dots, n-1)\, \mathbb{P}_\lambda(N_{t_2-t_1} = k_1). \qquad (11.2)\end{aligned}$$

Mais

$$\begin{aligned}\mathbb{P}_\lambda(N_{t_{i+1}-t_2} = m_i - m_1, i = 2, \dots, n-1) \\ &= \mathbb{P}_\lambda(N_{t_3-t_2} = k_2, N_{t_{i+1}-t_2} - N_{t_3-t_2} = m_i - m_2, i = 3, \dots, n-1),\end{aligned}$$

et l'on peut donc répéter la procédure (11.2), pour obtenir finalement

$$\mathbb{P}_\lambda(N_{t_{i+1}} - N_{t_i} = k_i, i = 1, \dots, n-1) = \prod_{i=0}^{n-1} \mathbb{P}_\lambda(N_{t_{i+1}-t_i} = k_i) = \prod_{i=0}^{n-1} \mathbb{P}_\lambda(N_{t_{i+1}} - N_{t_i} = k_i),$$

la dernière identité résultant de la première partie du théorème. □

Lemme 11.1. *Soit $(N_t)_{t\geq 0}$ un processus de Poisson d'intensité λ. Alors, T_n suit une loi* gamma(λ, n),

$$f_{T_n}(x) = \frac{1}{(n-1)!}\lambda^n x^{n-1} e^{-\lambda x} \mathbf{1}_{[0,\infty)}(x).$$

Démonstration. T_n est une somme de n variables aléatoires i.i.d. de loi exp(λ). Manifestement, T_1 suit une loi exp(λ), et celle-ci coïncide avec la loi gamma$(\lambda, 1)$. On procède par récurrence. Supposons l'énoncé vrai pour T_n. On a alors, pour $x \geq 0$,

$$\begin{aligned}f_{T_{n+1}}(x) = f_{T_n + X_{n+1}}(x) &= \int_{-\infty}^{\infty} f_{T_n}(u) f_{X_{n+1}}(x-u) \mathrm{d}u \\ &= \int_0^x \frac{1}{(n-1)!}\lambda^n u^{n-1} e^{-\lambda u}\, \lambda e^{-\lambda(x-u)} \mathrm{d}u \\ &= \frac{\lambda^{n+1}}{(n-1)!} e^{-\lambda x} \int_0^x u^{n-1} \mathrm{d}u \\ &= \frac{\lambda^{n+1}}{n!} e^{-\lambda x} x^n,\end{aligned}$$

et le lemme est démontré. □

Théorème 11.2. *Soit $(N_t)_{t\geq 0}$ un processus de Poisson d'intensité λ. Alors, pour tout $t \geq s \geq 0$, $N_t - N_s$ suit une loi* poisson$(\lambda(t-s))$.

Démonstration. Il suit du Théorème 11.1 qu'il suffit de considérer le cas $s = 0$. Puisque $N_t = n \iff T_n \leq t < T_{n+1}$, on a immédiatement

$$\begin{aligned}\mathbb{P}_\lambda(N_t = n) &= \mathbb{P}_\lambda(T_n \leq t < T_{n+1}) = \mathbb{P}_\lambda(T_{n+1} > t) - \mathbb{P}_\lambda(T_n > t) \\ &= \frac{\lambda^n}{n!} \int_t^\infty (x^n \lambda e^{-\lambda x} - n x^{n-1} e^{-\lambda x}) \mathrm{d}x \\ &= \frac{\lambda^n}{n!} \int_t^\infty \frac{\mathrm{d}}{\mathrm{d}x}(-x^n e^{-\lambda x}) \mathrm{d}x \\ &= \frac{\lambda^n}{n!} t^n e^{-\lambda t}.\end{aligned}$$

□

Définition 11.4. *On appelle* **accroissements** *d'un processus stochastique $(Z_t)_{t\geq 0}$ les différences $Z_t - Z_s$ entre les valeurs prises par le processus en deux temps $0 \leq s < t$.*

Un processus $(Z_t)_{t\geq 0}$ est à **accroissements stationnaires** *si, pour tout $s, t \geq 0$, $Z_{t+s} - Z_t$ a même loi que $Z_s - Z_0$.*

Un processus $(Z_t)_{t\geq 0}$ est à **accroissements indépendants** *si, pour tout choix de $0 = t_0 \leq t_1 \leq t_2 \leq \cdots \leq t_n < \infty$, les variables aléatoires $Z_{t_k} - Z_{t_{k-1}}$ sont indépendantes.*

Les Théorèmes 11.1 et 11.2 montrent que les accroissements $N_{t+s} - N_t$ d'un processus de Poisson d'intensité λ sont stationnaires, indépendants et suivent une loi de Poisson de paramètre λs. Nous allons montrer que ces propriétés caractérisent ce processus. Ceci fournit donc une définition alternative du processus de Poisson.

Théorème 11.3. *Un processus de comptage $(N_t)_{t\geq 0}$ est un processus de Poisson d'intensité λ si et seulement si ses accroissements $N_{t+s} - N_t$ sont stationnaires et indépendants, et suivent une loi* poisson(λs).

Remarque 11.2. *En fait, on peut montrer assez facilement qu'un processus de comptage $(N_t)_{t\geq 0}$ est un processus de Poisson (d'intensité non spécifiée) si et seulement si ses accroissements $N_{t+s} - N_t$ sont stationnaires et indépendants. Cela montre que ce processus va correctement modéliser toutes les situations où ces deux hypothèses sont approximativement vérifiées.*

Démonstration. On a déjà montré que le processus de Poisson possède les propriétés énoncées. Montrons donc que ces propriétés caractérisent ce processus.

Fixons $0 \leq s_1 < t_1 \leq s_2 < t_2 \leq \cdots \leq s_n < t_n$. En observant que $T_1 \in (s_1, t_1], \ldots, T_n \in (s_n, t_n]$ si et seulement si

- $N_{s_i} - N_{t_{i-1}} = 0$, $1 \leq i \leq n$, (avec $t_0 = 0$),
- $N_{t_i} - N_{s_i} = 1$, $1 \leq i < n$,
- $N_{t_n} - N_{s_n} \geq 1$,

et en utilisant les hypothèses sur les accroissements, on obtient

$$\begin{aligned}
&\mathbb{P}(T_1 \in (s_1,t_1],\dots,T_n\in(s_n,t_n]) \\
&\quad= \prod_{i=1}^{n}\mathbb{P}(N_{s_i}-N_{t_{i-1}}=0)\prod_{i=1}^{n-1}\mathbb{P}(N_{t_i}-N_{s_i}=1)\mathbb{P}(N_{t_n}-N_{s_n}\geq 1)\\
&\quad= \prod_{i=1}^{n}e^{-\lambda(s_i-t_{i-1})}\prod_{i=1}^{n-1}\lambda(t_i-s_i)e^{-\lambda(t_i-s_i)}(1-e^{-\lambda(t_n-s_n)})\\
&\quad= \lambda^{n-1}(e^{-\lambda s_n}-e^{-\lambda t_n})\prod_{i=1}^{n-1}(t_i-s_i)\\
&\quad= \int_{s_1}^{t_1}\cdots\int_{s_n}^{t_n}\lambda^n e^{-\lambda u_n}\mathrm{d}u_n\cdots\mathrm{d}u_1.
\end{aligned}$$

La loi conjointe de $(T_1,\dots,T_n)$ possède donc la densité

$$f_{(T_1,\dots,T_n)}(u_1,\dots,u_n)=\begin{cases}\lambda^n e^{-\lambda u_n} & \text{si } 0<u_1<\cdots<u_n,\\ 0 & \text{sinon.}\end{cases}$$

Déterminons à présent la densité de la loi conjointe de $(X_1,\dots,X_n)$. La fonction de répartition conjointe est donnée par

$$\begin{aligned}
\mathbb{P}(X_1\leq x_1,\dots,X_n\leq x_n) &= \mathbb{P}(T_1\leq x_1, T_2-T_1\leq x_2,\dots,T_n-T_{n-1}\leq x_n)\\
&= \int_0^{x_1}\int_{u_1}^{u_1+x_2}\cdots\int_{u_{n-1}}^{u_{n-1}+x_n} f_{(T_1,\dots,T_n)}(u_1,\dots,u_n)\mathrm{d}u_n\cdots\mathrm{d}u_1,
\end{aligned}$$

et la densité conjointe est donc donnée par

$$\begin{aligned}
\frac{\partial^n}{\partial x_1\cdots\partial x_n}\mathbb{P}(X_1\leq x_1,\dots,X_n\leq x_n) &= f_{(T_1,\dots,T_n)}(x_1,x_1+x_2,\dots,x_1+\cdots+x_n)\\
&= \lambda^n e^{-\lambda(x_1+\cdots+x_n)}.
\end{aligned}$$

On reconnaît la densité conjointe de n variables aléatoires i.i.d. de loi $\exp(\lambda)$. □

Nous allons voir à présent une troisième définition du processus, de nature plus dynamique.

Lemme 11.2. *Soit $(N_t)_{t\geq 0}$ un processus de Poisson d'intensité λ, et $0<t_1<\cdots<t_k$. Alors, pour $0\leq n_1\leq\cdots\leq n_k$ des entiers, on a, lorsque $\epsilon\downarrow 0$,*

$$\begin{aligned}
\mathbb{P}_\lambda(N_{t_k+\epsilon}-N_{t_k}=0\,|\,N_{t_j}=n_j,1\leq j\leq k) &= 1-\lambda\epsilon+o(\epsilon),\\
\mathbb{P}_\lambda(N_{t_k+\epsilon}-N_{t_k}=1\,|\,N_{t_j}=n_j,1\leq j\leq k) &= \lambda\epsilon+o(\epsilon),\\
\mathbb{P}_\lambda(N_{t_k+\epsilon}-N_{t_k}\geq 2\,|\,N_{t_j}=n_j,1\leq j\leq k) &= o(\epsilon).
\end{aligned}\tag{11.3}$$

Démonstration. Posons $n_0=0$. Puisque $\{N_{t_j}=n_j,1\leq j\leq k\}=\{N_{t_j}-N_{t_{j-1}}=n_j-n_{j-1},1\leq j\leq k\}$, il suit de l'indépendance et de la stationnarité des accroissements qu'il suffit de montrer que

$$\mathbb{P}_\lambda(N_\epsilon=0)=1-\lambda\epsilon+o(\epsilon)\quad\text{et}\quad\mathbb{P}_\lambda(N_\epsilon=1)=\lambda\epsilon+o(\epsilon).$$

Or, ceci est une conséquence immédiate du fait que N_ϵ suit une loi poisson$(\lambda\epsilon)$: on a, par exemple,

$$\mathbb{P}_\lambda(N_\epsilon=1)=e^{-\lambda\epsilon}\frac{(\lambda\epsilon)^1}{1!}=(1-\lambda\epsilon+o(\epsilon))\lambda\epsilon=\lambda\epsilon+o(\epsilon).$$

□

Nous allons voir maintenant que cette propriété caractérise le processus de Poisson d'intensité λ. Ceci fournit une troisième définition du processus.

Théorème 11.4. *Un processus de comptage est un processus de Poisson d'intensité λ si et seulement s'il satisfait* (11.3).

Démonstration. On a déjà montré que le processus de Poisson d'intensité λ possède les propriétés énoncées. Montrons donc que ces propriétés caractérisent ce processus.

Notons $A = \{N_{t_j} = n_j, 1 \le j \le k\}$, et posons, pour $t \ge 0$, $p_n(t) = \mathbb{P}(N_{t_k+t} - N_{t_k} = n \,|\, A)$. Il suffit de montrer que

$$p_n(t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \qquad n = 0, 1, \ldots,$$

puisque le résultat suivra alors du Théorème 11.3. En utilisant l'inégalité $|\mathbb{P}(B) - \mathbb{P}(C)| \le \mathbb{P}(B \triangle C)$, on obtient que [1]

$$|p_n(t) - p_n(s)| \le \mathbb{P}(N_{t_k+s} \ne N_{t_k+t})/\mathbb{P}(A),$$

ce qui montre que $p_n(t)$ est une fonction continue de t, puisque $\lim_{s\to t}\mathbb{P}(N_s \ne N_t) = 0$.

Pour simplifier les notations, posons $D_t = N_{t_k+t} - N_{t_k}$. Observons que $D_{t+\epsilon} = n \implies D_t = m$, pour un $m \le n$. On a donc

$$\begin{aligned} p_n(t+\epsilon) = p_n(t)\,&\mathbb{P}(D_{t+\epsilon} - D_t = 0 \,|\, A, D_t = n) \\ &+ \mathbf{1}_{\{n\ge 1\}}\, p_{n-1}(t)\,\mathbb{P}(D_{t+\epsilon} - D_t = 1 \,|\, A, D_t = n-1) \\ &+ \mathbf{1}_{\{n\ge 2\}} \sum_{m=0}^{n-2} p_m(t)\,\mathbb{P}(D_{t+\epsilon} - D_t = n-m \,|\, A, D_t = m). \end{aligned}$$

Par (11.3), on obtient

$$p_n(t+\epsilon) = p_n(t)(1-\lambda\epsilon) + \mathbf{1}_{\{n\ge 1\}}\, p_{n-1}(t)\lambda\epsilon + o(\epsilon).$$

En divisant par ϵ et en prenant la limite $\epsilon \downarrow 0$, on obtient [2]

$$p_n'(t) = -\lambda p_n(t) + \mathbf{1}_{\{n\ge 1\}}\, \lambda p_{n-1}(t), \tag{11.4}$$

avec condition au bord $p_n(0) = \delta_{n0}$.

Il reste à intégrer (11.4). Pour $n = 0$, on a

$$p_0'(t) = -\lambda p_0(t),$$

et donc $p_0(t) = e^{-\lambda t}$. En insérant cette solution dans l'équation pour $p_1(t)$, on trouve

$$p_1'(t) = -\lambda p_1(t) + \lambda e^{-\lambda t},$$

et donc $p_1(t) = \lambda t e^{-\lambda t}$. Par induction, on obtient donc bien

$$p_n(t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!},$$

pour chaque $n \ge 0$. □

1. En effet, $\{N_{t_k+s} = n, A\} = \{N_{t_k+s} = n, N_{t_k+t} = n, A\} \cup \{N_{t_k+s} = n, N_{t_k+t} \ne n, A\}$ et par conséquent $\{N_{t_k+s} = n, A\} \setminus \{N_{t_k+t} = n, A\} = \{N_{t_k+s} = n, N_{t_k+t} \ne n, A\}$, et de même avec s et t interchangés.

2. Il y a une petite subtilité ici : *a priori*, la dérivée dans le membre de gauche de (11.4) n'est qu'une dérivée à droite. Afin de montrer qu'il s'agit réellement d'une dérivée, il suffit d'observer que le membre de droite est continu. En effet, pour montrer qu'une fonction continue $f(t)$ avec dérivée à droite $f^+(t)$ continue pour tout $t \ge 0$, est nécessairement dérivable pour chaque $t > 0$, il suffit de prouver que $F(t) = f(t) - f(0) - \int_0^t f^+(s)\mathrm{d}s \equiv 0$. Supposons que ce ne soit pas le cas, et que (disons) $F(t_0) < 0$. Alors $G(t) = F(t) - tF(t_0)/t_0$ satisfait $G(0) = G(t_0) = 0$ et, puisque $F^+ \equiv 0$, $G^+(t) > 0$, ce qui implique que G doit posséder un maximum strictement positif en un point $s_0 \in (0, t_0)$. Mais $G^+(s_0) \le 0$, puisque s_0 est un maximum, ce qui est une contradiction.

11.2 Autres propriétés

11.2.1 Le paradoxe de l'autobus

Nous avons déjà rencontré ce paradoxe dans les résultats de la section précédente, mais n'avons pas encore explicité son caractère surprenant (au premier abord). On considère une lampe dont on change immédiatement l'ampoule lorsque celle-ci est défaillante ; la durée de vie d'une ampoule est supposée suivre une loi $\exp(\lambda)$. Si l'on considère un temps arbitraire $t > 0$, cet instant se trouvera presque-sûrement entre deux pannes. On a vu que la variable aléatoire X_1^t représentant le temps séparant t de la prochaine panne suit une loi $\exp(\lambda)$; en particulier, le temps moyen jusqu'à la prochaine panne est donné par $1/\lambda$. On peut de la même façon déterminer la loi du temps écoulé entre la panne précédente et t, $S^t = t - T_{N_t}$. Bien sûr, si $s > t$, $\mathbb{P}_\lambda(S^t > s) = 0$, puisque $T_0 = 0$. Pour $s \leq t$, on trouve

$$\mathbb{P}_\lambda(S^t \geq s) = \mathbb{P}_\lambda(N_t - N_{t-s} = 0) = \mathbb{P}_\lambda(N_s = 0) = e^{-\lambda s}.$$

Par conséquent, S^t a même loi que $\min(X, t)$, où X est une variable de loi $\exp(\lambda)$. Si l'on s'intéresse au comportement de la lampe après un temps long, la loi de S^t est bien entendu très bien approximée par une loi $\exp(\lambda)$.

En particulier, on voit que, pour t grand, le temps moyen entre les deux pannes est très proche de $2/\lambda$, alors que la durée de vie moyenne d'une ampoule est de $1/\lambda$. C'est le paradoxe de l'autobus. Celui-ci est traditionnellement présenté comme suit : les différences entre les temps de passage successifs d'un autobus passant par un arrêt donné suivent une loi exponentielle, de moyenne 5 minutes. Un individu arrive à l'arrêt pour prendre le bus. Le temps moyen qui s'écoule entre le passage du bus précédent son arrivée et le passage du bus suivant est (approximativement) de 10 minutes, bien que les bus passent en moyenne toutes les 5 minutes !

L'explication de ce « paradoxe » est la suivante : la distribution des longueurs d'intervalle n'est pas triviale, certains seront beaucoup plus longs que la moyenne, d'autres beaucoup plus courts. En faisant une observation « au hasard », on a donc davantage de chance de tomber dans un long intervalle plutôt que dans un court. On biaise ainsi la loi de la taille de l'intervalle observé vers les plus grandes tailles.

11.2.2 Processus de Poisson et statistiques d'ordre

Soit $t > 0$. Nous allons étudier la loi de T_1 conditionnellement à $N_t = 1$. Dans ce cas, on a bien entendu $T_1 \leq t$, et donc, pour $s \in (0, t]$,

$$\begin{aligned}\mathbb{P}_\lambda(T_1 \leq s \mid N_t = 1) &= \frac{\mathbb{P}_\lambda(T_1 \leq s, N_t = 1)}{\mathbb{P}_\lambda(N_t = 1)} = \frac{\mathbb{P}_\lambda(N_s = 1, N_t - N_s = 0)}{\mathbb{P}_\lambda(N_t = 1)} \\ &= \frac{(\lambda s e^{-\lambda s})(e^{-\lambda(t-s)})}{\lambda t e^{-\lambda t}} = \frac{s}{t}.\end{aligned}$$

T_1 suit donc une loi uniforme sur $(0, t]$, conditionnellement à $N_t = 1$. Ainsi, savoir qu'un événement a eu lieu avant le temps t ne nous fournit aucune information sur l'instant auquel il a été réalisé. De plus, la loi conditionnelle est indépendante de l'intensité λ du processus.

Nous allons à présent généraliser ce résultat, en déterminant la loi de $T_1, \ldots, T_n$, conditionnellement à $N_t = n$. Soient $0 < t_1 < \cdots < t_n < t$. On a, pour tout $\epsilon > 0$ suffisamment

petit,

$$\begin{aligned}
&\mathbb{P}_\lambda(T_k \in (t_k-\epsilon, t_k+\epsilon), 1 \le k \le n \,|\, N_t = n) \\
&\qquad = \frac{\mathbb{P}_\lambda(T_k \in (t_k-\epsilon, t_k+\epsilon), 1 \le k \le n, N_t = n)}{\mathbb{P}_\lambda(N_t = n)} \\
&\qquad = \frac{e^{-\lambda(t_1-\epsilon)} 2\epsilon\lambda e^{-2\epsilon\lambda} e^{-\lambda(t_2-t_1-2\epsilon)} \cdots 2\epsilon\lambda e^{-2\epsilon\lambda} e^{-\lambda(t-t_n-\epsilon)}}{(\lambda^n t^n/n!) e^{-\lambda t}} \\
&\qquad = (2\epsilon/t)^n n!,
\end{aligned}$$

puisque l'événement $\{T_k \in (t_k-\epsilon, t_k+\epsilon), 1 \le k \le n, N_t = n\}$ est réalisé si et seulement si $N_{t_1-\epsilon} = 0$, $N_{t_k+\epsilon} - N_{t_k-\epsilon} = 1$, $k = 1, \ldots, n$, $N_{t_{k+1}-\epsilon} - N_{t_k+\epsilon} = 0$, $k = 1, \ldots, n-1$, et $N_t - N_{t_n+\epsilon} = 0$. Par conséquent, la densité conjointe de $T_1, \ldots, T_n$, conditionnellement à $N_t = n$ est donnée par

$$\lim_{\epsilon \downarrow 0} \frac{1}{(2\epsilon)^n} \mathbb{P}_\lambda(T_k \in (t_k-\epsilon, t_k+\epsilon), 1 \le k \le n \,|\, N_t = n) = n!\, t^{-n},$$

si $0 < t_1 < \cdots < t_n < t$, et 0 sinon. C'est ce qu'on appelle la loi conjointe des **statistiques d'ordre** de n variables aléatoires indépendantes de loi uniforme sur $(0, t]$. Elle revient à tirer au hasard, indépendamment, n points uniformément sur l'intervalle $(0, t]$, puis à les ordonner du plus petit au plus grand.

11.2.3 Superposition et amincissement

Le processus de Poisson possède deux autres propriétés remarquables : (i) la « superposition » de deux processus de Poisson indépendants donne à nouveau un processus de Poisson, dont l'intensité est la somme de celles des deux processus originaux, et (ii) tout processus de Poisson d'intensité λ peut être décomposé en deux processus de Poisson *indépendants* d'intensités λ_1 et $\lambda - \lambda_1$.

Théorème 11.5. *Soient $\lambda_1, \lambda_2 > 0$, et $\lambda = \lambda_1 + \lambda_2$. Soient $(N_t^{(1)})_{t\ge 0}$ et $(N_t^{(2)})_{t\ge 0}$ deux processus de Poisson indépendants d'intensités λ_1 et λ_2. Alors, le processus défini par*

$$N_t = N_t^{(1)} + N_t^{(2)}$$

est un processus de Poisson d'intensité λ.

Démonstration. On utilise la caractérisation du processus de Poisson du Théorème 11.3. Pour tout $0 < s < t$ et $n \ge 0$, l'indépendance des processus $N_t^{(1)}$ et $N_t^{(2)}$ implique que

$$\begin{aligned}
\mathbb{P}(N_t - N_s = n) &= \mathbb{P}(N_t^{(1)} + N_t^{(2)} - N_s^{(1)} - N_s^{(2)} = n) \\
&= \sum_{k=0}^{n} \mathbb{P}(N_t^{(1)} - N_s^{(1)} = n-k) \mathbb{P}(N_t^{(2)} - N_s^{(2)} = k) \\
&= \sum_{k=0}^{n} \frac{(\lambda_1(t-s))^{n-k}}{(n-k)!} e^{-\lambda_1(t-s)} \frac{(\lambda_2(t-s))^k}{k!} e^{-\lambda_2(t-s)} \\
&= \frac{(t-s)^n}{n!} e^{-\lambda(t-s)} \sum_{k=0}^{n} \binom{n}{k} \lambda_1^{n-k} \lambda_2^k = \frac{(t-s)^n}{n!} \lambda^n e^{-\lambda(t-s)},
\end{aligned}$$

ce qui montre que les accroissements $N_t - N_s$ de $(N_t)_{t\ge 0}$ sont stationnaires et suivent une loi de Poisson de paramètre $\lambda(t-s)$. Il reste à vérifier qu'ils sont indépendants. Nous ne le ferons que

pour deux intervalles, le cas général se traitant de la même manière. Soient donc $0 < s \leq t < u$, et $n, m \geq 0$. Écrivons $\Delta^{(i)} = N_u^{(i)} - N_t^{(i)}$, $i = 1, 2$ et $\Delta = N_u - N_t$. On a

$$\begin{aligned}
&\mathbb{P}(\Delta = n, N_s = m) \\
&\quad = \mathbb{P}(\Delta^{(1)} + \Delta^{(2)} = n, N_s^{(1)} + N_s^{(2)} = m) \\
&\quad = \sum_{k=0}^{n} \sum_{\ell=0}^{m} \mathbb{P}(\Delta^{(1)} = n-k, N_s^{(1)} = m-\ell)\mathbb{P}(\Delta^{(2)} = k, N_s^{(2)} = \ell) \\
&\quad = \sum_{k=0}^{n} \sum_{\ell=0}^{m} \mathbb{P}(\Delta^{(1)} = n-k)\mathbb{P}(N_s^{(1)} = m-\ell)\mathbb{P}(\Delta^{(2)} = k)\mathbb{P}(N_s^{(2)} = \ell) \\
&\quad = \sum_{k=0}^{n} \sum_{\ell=0}^{m} \mathbb{P}(\Delta^{(1)} = n-k, \Delta^{(2)} = k)\mathbb{P}(N_s^{(1)} = m-\ell, N_s^{(2)} = \ell) \\
&\quad = \mathbb{P}(\Delta = n)\mathbb{P}(N_s = m).
\end{aligned}$$

□

Définition 11.5. *On dit que le processus $(N_t)_{t\geq 0}$ ci-dessus est la **superposition** des processus $(N_t^{(1)})_{t\geq 0}$ et $(N_t^{(2)})_{t\geq 0}$.*

Théorème 11.6. *Soit $(N_t)_{t\geq 0}$ un processus de Poisson d'intensité λ, et soit $p \in (0,1)$. On peint chaque point du processus en rouge ou en bleu, de façon indépendante, avec probabilité p et $1-p$ respectivement. Alors, les points rouges et bleus définissent deux processus de Poisson indépendants d'intensités λp et $\lambda(1-p)$ respectivement.*

Démonstration. Soit $0 < s < t$ et $k \geq 0$. On a

$$\begin{aligned}
\mathbb{P}_\lambda(N_t^{(1)} - N_s^{(1)} = k) &= \sum_{n=k}^{\infty} \mathbb{P}_\lambda(N_t - N_s = n) \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{n=k}^{\infty} \frac{\lambda^n (t-s)^n}{n!} e^{-\lambda(t-s)} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{(\lambda p\,(t-s))^k}{k!} e^{-\lambda(t-s)} \sum_{n\geq k} \frac{(\lambda(t-s)(1-p))^{n-k}}{(n-k)!} \\
&= \frac{(\lambda p\,(t-s))^k}{k!} e^{-\lambda p\,(t-s)}.
\end{aligned}$$

On montre de la même façon que $N_t^{(2)} - N_s^{(2)}$ est poisson$(\lambda(1-p))$.

Soient $0 \leq s_1 < t_1 \leq s_2 < t_2 \leq \cdots \leq s_n < t_n$. Alors, en notant $\Delta_i = N_{t_i} - N_{s_i}$ et

$\Delta_i^{(j)} = N_{t_i}^{(j)} - N_{s_i}^{(j)}$ $(j = 1, 2)$, on a

$$\begin{aligned}
\mathbb{P}_\lambda(\Delta_i^{(1)} &= n_i, \Delta_i^{(2)} = m_i, 1 \le i \le n) \\
&= \mathbb{P}_\lambda(\Delta_i^{(1)} = n_i, 1 \le i \le n \,|\, \Delta_i = m_i + n_i, 1 \le i \le n)\mathbb{P}_\lambda(\Delta_i = m_i + n_i, 1 \le i \le n) \\
&= \prod_{i=1}^n \binom{n_i + m_i}{n_i} p^{n_i}(1-p)^{m_i} \prod_{i=1}^n \mathbb{P}(\Delta_i = m_i + n_i) \\
&= \prod_{i=1}^n \binom{n_i + m_i}{n_i} p^{n_i}(1-p)^{m_i} \prod_{i=1}^n \frac{(\lambda(t_i - s_i))^{n_i+m_i}}{(n_i+m_i)!} e^{-\lambda(t_i - s_i)} \\
&= \prod_{i=1}^n \frac{(\lambda p\,(t_i - s_i))^{n_i}}{n_i!} e^{-\lambda p(t_i - s_i)} \frac{(\lambda(1-p)\,(t_i - s_i))^{m_i}}{m_i!} e^{-\lambda(1-p)(t_i - s_i)} \\
&= \prod_{i=1}^n \mathbb{P}_\lambda(\Delta_i^{(1)} = n_i)\, \mathbb{P}_\lambda(\Delta_i^{(2)} = m_i),
\end{aligned}$$

et les processus $(N_t^{(1)})_{t\ge 0}$ et $(N_t^{(2)})_{t\ge 0}$ sont donc à accroissements indépendants, et sont indépendants l'un de l'autre. □

Définition 11.6. *On dit que les processus $(N_t^{(1)})_{t\ge 0}$ et $(N_t^{(2)})_{t\ge 0}$ ci-dessus sont des* ***amincissements*** *du processus $(N_t)_{t\ge 0}$.*

Remarque 11.3. *Bien entendu, on peut itérer les procédures de superposition et d'amincissement. Les résultats ci-dessus restent donc valides pour un nombre fini arbitraire de processus $(N_t^{(i)})_{t\ge 0}$.*

Exemple 11.1. On considère deux caissières, servant chacune une infinité de clients. On suppose que les temps de service de chaque caissière sont i.i.d. de loi $\exp(\lambda_1)$ et $\exp(\lambda_2)$ respectivement. On désire déterminer la probabilité que la première caissière ait fini de s'occuper de son $n^{\text{ème}}$ client avant que la seconde ait fini de s'occuper de son $m^{\text{ème}}$ client, c'est-à-dire

$$\mathbb{P}(T_n^{(1)} < T_m^{(2)}).$$

Une approche revient à utiliser le fait que ces deux variables aléatoires sont indépendantes et de lois $\text{gamma}(\lambda_1, n)$ et $\text{gamma}(\lambda_2, m)$ respectivement, et faire un calcul laborieux. Nous allons à la place utiliser les résultats de cette section. Soit $(N_t)_{t\ge 0}$ un processus de Poisson de paramètre $\lambda = \lambda_1 + \lambda_2$. On a vu que les processus $(N_t^{(1)})_{t\ge 0}$ et $(N_t^{(2)})_{t\ge 0}$ peuvent être obtenus en coloriant les points de $(N_t)_{t\ge 0}$ indépendamment en rouge et en bleu, avec probabilité λ_1/λ et λ_2/λ respectivement. Par conséquent, $T_n^{(1)} < T_m^{(2)}$ si et seulement si au moins n points parmi les $n + m - 1$ premiers points de N_t sont coloriés en rouge. On a donc

$$\mathbb{P}(T_n^{(1)} < T_m^{(2)}) = \sum_{k=n}^{n+m-1} \binom{n+m-1}{k} \Big(\frac{\lambda_1}{\lambda_1+\lambda_2}\Big)^k \Big(\frac{\lambda_2}{\lambda_1+\lambda_2}\Big)^{n+m-1-k}.$$

11.2.4 Processus de Poisson non homogène

Il est souvent peu réaliste de supposer que la fréquence d'apparition des points est constante. Par exemple, si on veut modéliser les arrivées de clients dans un supermarché, ou de voitures sur une autoroute, ou de requêtes sur un serveur web, il est clair que la fréquence de ces événements va dépendre de l'heure de la journée, du jour de la semaine, de la saison, etc. Afin de modéliser ce type de situations, on va permettre à l'intensité $\lambda(t)$ du processus de Poisson de varier au

cours du temps. Il est possible de définir ce processus pour des fonctions $\lambda(t)$ très générales (il suffit que $\lambda(t)$ soit intégrable) ; nous supposerons ici pour simplifier que $\lambda(t)$ est continue par morceaux.

Définition 11.7. *Un processus de comptage à accroissements indépendants $(N_t)_{t\geq 0}$ est un* ***processus de Poisson non homogène*** *de* ***fonction de densité*** *$\lambda(t) \geq 0$, $t \geq 0$, si*

1. $\mathbb{P}(N_{t+\epsilon} - N_t = 1) = \lambda(t)\epsilon + o(\epsilon)$;
2. $\mathbb{P}(N_{t+\epsilon} - N_t \geq 2) = o(\epsilon)$.

Manifestement, un tel processus n'est *pas* à accroissements stationnaires (sauf lorsque $\lambda(t) \equiv \lambda$ est constante, auquel cas il se réduit à un processus de Poisson d'intensité λ).

Théorème 11.7. *Soit $(N_t)_{t\geq 0}$ un processus de Poisson non homogène de fonction de densité $\lambda(t)$. Alors, pour tout $t \geq s \geq 0$, $N_t - N_s$ suit une loi* $\mathsf{poisson}(m(t) - m(s))$, *où*

$$m(u) = \int_0^u \lambda(v)\mathrm{d}v.$$

Définition 11.8. *La fonction $m(t)$ dans le Théorème 11.7 est appelée* ***fonction de valeur moyenne*** *du processus.*

Démonstration. La preuve est semblable à celle du Théorème 11.4, et nous ne ferons que l'esquisser. Notons

$$p_n(s,t) = \mathbb{P}(N_t - N_s = n), \qquad n = 0, 1, 2, \ldots$$

Par indépendance des accroissements, on peut écrire

$$\begin{aligned} p_n(s, t+\epsilon) &= \mathbb{P}(N_t - N_s = n, N_{t+\epsilon} - N_t = 0) \\ &\qquad + \mathbf{1}_{\{n\geq 1\}}\mathbb{P}(N_t - N_s = n-1, N_{t+\epsilon} - N_t = 1) + o(\epsilon) \\ &= p_n(s,t)(1 - \lambda(t)\epsilon + o(\epsilon)) + \mathbf{1}_{\{n\geq 1\}}p_{n-1}(s,t)(\lambda(t)\epsilon + o(\epsilon)) + o(\epsilon). \end{aligned}$$

Il suit que

$$\frac{\partial}{\partial t}p_n(s,t) = \lambda(t)\left(\mathbf{1}_{\{n\geq 1\}}p_{n-1}(s,t) - p_n(s,t)\right),$$

avec condition au bord $p_n(s,s) = \delta_{n0}$, pour tout $s \geq 0, n \in \mathbb{N}$.

Lorsque $n = 0$, cette équation est simplement

$$\frac{\partial}{\partial t}p_0(s,t) = -\lambda(t)p_0(s,t),$$

dont la solution est

$$p_0(s,t) = \exp\Big(-\int_s^t \lambda(u)\mathrm{d}u\Big) = e^{-(m(t)-m(s))}, \qquad s, t \geq 0.$$

En substituant ce résultat dans l'équation pour $n = 1$, on obtient

$$\frac{\partial}{\partial t}p_1(s,t) = \lambda(t)\left(e^{-(m(t)-m(s))} - p_1(s,t)\right),$$

qui peut être réécrit comme

$$\frac{\partial}{\partial t}p_1(s,t) = \left(e^{-(m(t)-m(s))} - p_1(s,t)\right)\frac{\partial}{\partial t}(m(t) - m(s)).$$

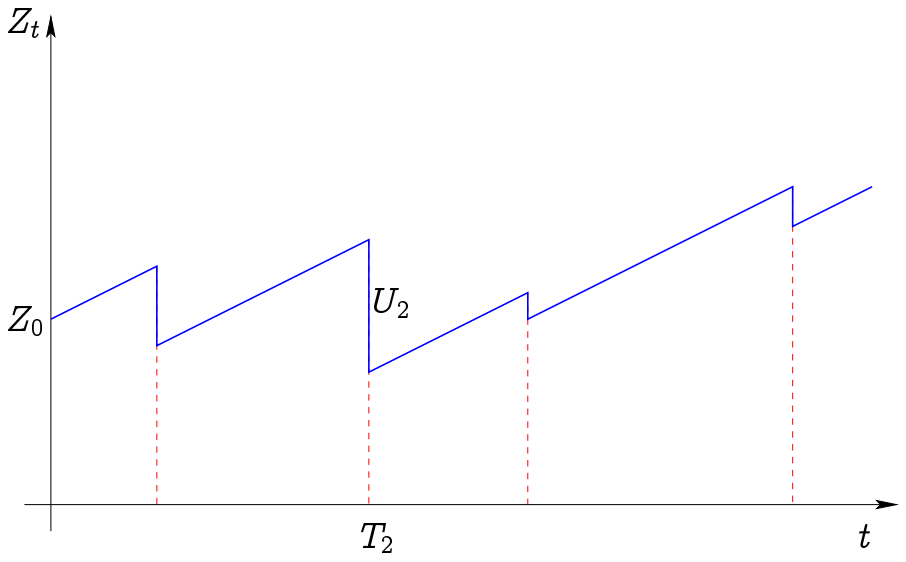


FIGURE 11.3: Évolution des réserves d'une compagnie d'assurance.

On voit alors facilement que la solution est donnée par

$$p_1(s,t) = e^{-(m(t)-m(s))}(m(t)-m(s)), \qquad s,t \geq 0.$$

Par récurrence, on montre ensuite que

$$p_n(s,t) = e^{-(m(t)-m(s))}\frac{1}{n!}(m(t)-m(s))^n, \qquad s,t \geq 0, n \in \mathbb{N}.$$

□

11.2.5 Processus de Poisson composé

Le processus de Poisson est utilisé comme base pour construire de nombreux autres processus. Nous allons en voir un exemple dans cette sous-section : le processus de Poisson composé.

Définition 11.9. *Soient $(N_t)_{t\geq 0}$ un processus de Poisson d'intensité λ, et $U_1, U_2, \ldots$ des variables aléatoires i.i.d. indépendantes du processus de Poisson. Le processus stochastique*

$$Y_t = \sum_{k=1}^{N_t} U_k, \qquad t \geq 0$$

(avec la convention que $Y_t = 0$ si $N_t = 0$) est appelé ***processus de Poisson composé****.*

Exemple 11.2. Voici un modèle très simple pour les réserves d'une compagnie d'assurances.

On considère que des sinistres se produisent aux instants T_n d'un processus de Poisson homogène et que le $n^{\text{ème}}$ sinistre coûte à la compagnie d'assurance une somme U_n. Si c est le taux des primes par unité de temps, le bilan de la compagnie à l'instant t est donc $Z_t = Z_0 + ct - Y_t$, où Z_0 est son capital initial. Soit $W = \inf\{t \geq 0 : Z_t < 0\}$ le premier instant où les réserves de la compagnie deviennent négatives. Le problème est alors de trouver la probabilité de ruine, c'est à dire $\mathbb{P}(W < \infty \mid Z_0 = x)$.

Diverses propriétés du processus de Poisson composé seront étudiées en exercices.

11.2.6 Processus de Poisson spatial

Le processus de Poisson introduit précédemment était restreint à $[0, \infty)$. Il est en fait possible de l'étendre à des espaces beaucoup plus généraux. Nous esquissons à présent le cas de $\mathbb{R}^d$.

Une réalisation d'un processus de Poisson sur $\mathbb{R}^d$ est un sous-ensemble aléatoire dénombrable Π de $\mathbb{R}^d$. La loi de Π sera caractérisée via la collection de variables aléatoires $(N(B))_{B\in\mathcal{B}(\mathbb{R}^d)}$ indicées par les boréliens de $\mathbb{R}^d$, la variable $N(B)$ correspondant au nombre de points de Π se trouvant dans B.

On note $|A|$ le volume (c'est-à-dire la mesure de Lebesgue) d'un borélien A.

Définition 11.10. *Le sous-ensemble aléatoire dénombrable Π de $\mathbb{R}^d$ est un* ***processus de Poisson d'intensité*** *λ si*

- *$N(B)$ suit une loi de Poisson de paramètre $\lambda|B|$, pour tout $B \in \mathcal{B}(\mathbb{R}^d)$;*
- *$N(B_1), \dots, N(B_n)$ sont indépendantes lorsque $B_1, \dots, B_n$ sont disjoints.*

On peut également considérer des processus de Poisson inhomogènes (c'est-à-dire, d'intensité variable), mais nous ne le ferons pas ici.

Un grand nombre des résultats établis plus haut pour le processus de Poisson sur $[0, \infty)$ s'étendent à ce cadre-ci : en particulier, les propriétés d'amincissement et de superposition admettent des généralisations naturelles. Dans ce bref aperçu, nous nous contenterons de démontrer une propriété importante, qui montre que le processus de Poisson sur $\mathbb{R}^d$ modélise bien une « distribution aléatoire uniforme » de points dans $\mathbb{R}^d$. Elle est également très utile pour la simulation de tels processus.

Théorème 11.8. *Soit Π un processus de Poisson d'intensité λ sur $\mathbb{R}^d$, et soit A un ouvert de $\mathbb{R}^d$ de volume fini. Alors, conditionnellement à $N(A) = n$, les n points de Π se trouvant dans A suivent la même loi que n points choisis indépendamment avec la mesure uniforme sur A.*

Démonstration. Notons $B_\epsilon(x) = \{y \in A : \|y - x\|_\infty < \epsilon/2\}$. Soient $x_1, \dots, x_n$ des points distincts de A. Étant donnée une réalisation du processus de Poisson avec $N(A) = n$, on numérote au hasard de façon uniforme les n points dans A : $X_1, \dots, X_n$. Alors, pour $\epsilon > 0$ suffisamment petit,

$$\begin{aligned}
&\mathbb{P}(X_i \in B_\epsilon(x_i),\, i = 1, \dots, n \mid N(A) = n) \\
&\qquad = \frac{1}{n!}\mathbb{P}(N(B_\epsilon(x_i)) = 1,\, i = 1, \dots, n \mid N(A) = n) \\
&\qquad = \frac{1}{n!}\frac{\mathbb{P}(N(A \setminus \bigcup_{j=1}^n B_\epsilon(x_j)) = 0)\prod_{i=1}^n \mathbb{P}(N(B_\epsilon(x_i)) = 1)}{\mathbb{P}(N(A) = n)} \\
&\qquad = \frac{1}{n!}\frac{e^{-\lambda(|A| - n\epsilon^d)}\prod_{i=1}^n \lambda\epsilon^d e^{-\lambda\epsilon^d}}{(\lambda|A|)^n e^{-\lambda|A|}/n!} \\
&\qquad = \epsilon^{nd}|A|^{-n}.
\end{aligned}$$

Par conséquent, conditionnellement à $N(A) = n$, la densité conjointe de $(X_1, \dots, X_n)$ en $(x_1, \dots, x_n)$ est donnée par

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon^{nd}}\,\mathbb{P}(X_i \in B_\epsilon(x_i),\, i = 1, \dots n \,|\, N(A) = n) = |A|^{-n},$$

et coïncide donc bien avec la densité conjointe de n points tirés indépendamment uniformément dans A. □

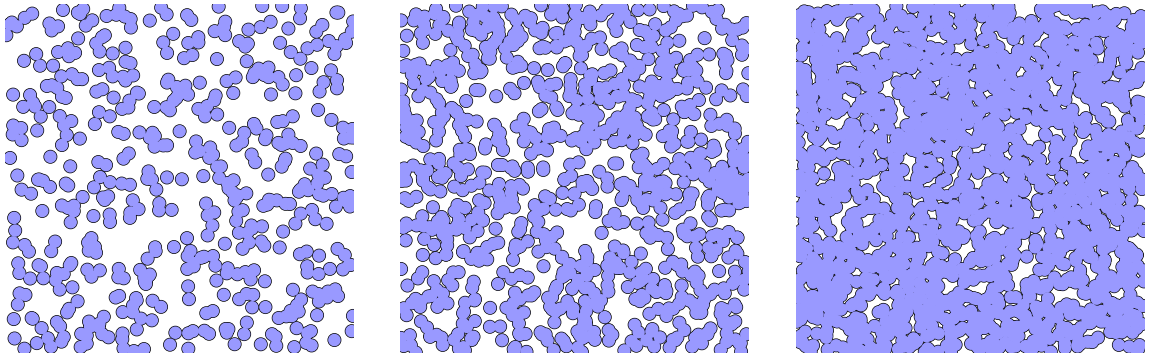


FIGURE 11.4: Trois réalisations du processus booléen de l'Exemple 11.3 pour des intensités croissantes du processus de Poisson sous-jacent.

Exemple 11.3 (Modèle booléen). Nous allons à présent décrire un cas particulier du modèle booléen. Dans ce modèle, on associe à chaque réalisation Π d'un processus de Poisson d'intensité λ dans $\mathbb{R}^2$ le sous-ensemble $\overline{\Pi}$ de $\mathbb{R}^2$ donné par l'union des disques de rayon $r > 0$ centrés sur les points de Π,

$$\overline{\Pi} = \bigcup_{\mathbf{x} \in \Pi} D_r(\mathbf{x}),$$

où $D_r(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^2 : \|\mathbf{y} - \mathbf{x}\|_2 \leq r\}$; c.f. Fig. 11.4. (Dans une version plus générale du modèle booléen, on remplace les disques par des compacts eux-mêmes aléatoires.) On peut voir ce modèle comme une version continue du modèle de percolation du chapitre 10.

Soit A un borélien de $\mathbb{R}^2$ tel que $0 < |A| < \infty$. On désire déterminer la fraction moyenne de A couverte par les disques. On a

$$\begin{aligned}\mathbb{E}(|A \cap \overline{\Pi}|) &= \mathbb{E}\left(\int_A \mathbf{1}_{\overline{\Pi}}(\mathbf{x})\mathrm{d}\mathbf{x}\right) \\ &= \int_A \mathbb{P}(\mathbf{x} \in \overline{\Pi})\mathrm{d}\mathbf{x}.\end{aligned}$$

Or, par définition du processus,

$$\begin{aligned}\mathbb{P}(\mathbf{x} \notin \overline{\Pi}) &= \mathbb{P}(\text{Aucun point de } \Pi \text{ ne se trouve à distance au plus } r \text{ de } \mathbf{x}) \\ &= \mathbb{P}\left(\Pi \cap D_r(\mathbf{x}) = \varnothing\right) \\ &= \mathbb{P}\left(N(D_r(\mathbf{x})) = 0\right) \\ &= \exp(-\lambda\pi r^2).\end{aligned}$$

Par conséquent, la fraction de A couverte par les disques est donnée par

$$\frac{\mathbb{E}(|A \cap \overline{\Pi}|)}{|A|} = 1 - e^{-\lambda\pi r^2}.$$

11.2.7 Processus de renouvellement

Fonction de renouvellement, équation de renouvellement

Avant de clore ce chapitre, nous allons brièvement discuter des processus de renouvellement généraux. Il s'agit d'un sujet de grande importance, que nous ne ferons qu'effleurer.

Soit $(N_t)_{t \geq 0}$ un processus de renouvellement, c'est-à-dire un processus de comptage pour lequel les temps d'attente sont i.i.d., et supposons pour simplifier [3] que la loi commune des temps d'attente possède la densité f. On notera F la fonction de répartition correspondante.

3. Mais tout ce que nous dirons ici s'étend à des lois quelconques.

Il est aisé d'exprimer la loi des temps de renouvellement T_k à partir de celle des temps d'attente.

Lemme 11.3. *$f_{T_1}(t) = f(t)$, et $f_{T_{k+1}}(t) = \int f_{T_k}(t-s)f(s)\mathrm{d}s$, pour $k \geq 1$.*

Démonstration. Cela suit immédiatement de la relation $T_{k+1} = T_k + X_{k+1}$ et de l'indépendance des variables aléatoires T_k et X_{k+1}. □

Lemme 11.4. *$\mathbb{P}(N_t = k) = F_{T_k}(t) - F_{T_{k+1}}(t)$.*

Démonstration. Il suffit d'observer que $\{N_t = k\} = \{N_t \geq k\} \setminus \{N_t \geq k+1\}$, et d'utiliser le fait que $N_t \geq n \Leftrightarrow T_n \leq t$. □

Il est en général impossible de déterminer explicitement la loi de N_t, et il faudra souvent se satisfaire d'informations sur $\mathbb{E}(N_t)$.

Définition 11.11. *La **fonction de renouvellement** est définie par $m(t) = \mathbb{E}(N_t)$.*

Lemme 11.5. *$m(t) = \sum_{k=1}^{\infty} F_{T_k}(t)$.*

Démonstration. Manifestement, $N_t = \sum_{k\geq 1} \mathbf{1}_{\{T_k \leq t\}}$. Par conséquent,

$$m(t) = \mathbb{E}(\sum_{k\geq 1} \mathbf{1}_{\{T_k \leq t\}}) = \sum_{k\geq 1} \mathbb{P}(T_k \leq t).$$

□

Le résultat précédent n'est que de peu d'utilité en général. Une approche alternative pour déterminer m est la suivante.

Lemme 11.6. *La fonction de renouvellement satisfait l'**équation de renouvellement**,*

$$m(t) = F(t) + \int_0^t m(t-s)f(s)\,\mathrm{d}s, \qquad t \geq 0.$$

Démonstration. En conditionnant sur X_1, on a

$$m(t) = \mathbb{E}(\mathbb{E}(N_t \mid X_1)).$$

À présent, $\mathbb{E}(N_t \mid X_1 = x) = 0$ si $t < x$. D'un autre côté,

$$\mathbb{E}(N_t \mid X_1 = x) = 1 + \mathbb{E}(N_{t-x}), \qquad \text{si } t \geq x.$$

On en déduit que

$$m(t) = \int_0^{\infty} \mathbb{E}(N_t \mid X_1 = x) f(x)\,\mathrm{d}x = \int_0^t (1 + m(t-x)) f(x)\,\mathrm{d}x.$$

□

Remarque 11.4. *Évidemment, $m(t) = \sum_{k=1}^{\infty} \mathbb{P}(T_k \leq t)$ est une solution de l'équation de renouvellement. En fait, on peut montrer qu'il s'agit de l'unique solution bornée sur tout intervalle fini.*

Remarque 11.5. *On peut montrer qu'il y a bijection entre les lois des temps d'attente et la fonction de renouvellement. En particulier, le processus de Poisson est le seul processus de renouvellement dont la fonction de renouvellement est linéaire.*

Théorèmes limites

Nous allons à présent nous intéresser au comportement asymptotique de N_t et $m(t)$, lorsque t est grand.

Soit $\mu = \mathbb{E}(X_1)$. Dans cette sous-section, nous supposerons que $\mu < \infty$.

Théorème 11.9. *$\frac{1}{t} N_t \overset{\text{p.s.}}{\to} \frac{1}{\mu}$, lorsque $t \to \infty$.*

Démonstration. Puisque $T_{N_t} \leq t < T_{N_t+1}$, on a, lorsque $N_t > 0$,

$$\frac{T_{N_t}}{N_t} \leq \frac{t}{N_t} \leq \frac{T_{N_t+1}}{N_t+1}(1 + \frac{1}{N_t}).$$

D'une part, $N_t \overset{\text{p.s.}}{\to} \infty$ lorsque $t \to \infty$. D'autre part, par la loi forte des grands nombres, $\frac{1}{N}\sum_{i=1}^{N} X_i \overset{\text{p.s.}}{\to} \mu$, lorsque $N \to \infty$. Par conséquent,

$$\frac{T_{N_t}}{N_t} = \frac{1}{N_t}\sum_{i=1}^{N_t} X_i \overset{\text{p.s.}}{\to} \mu,$$

et donc

$$\mu \leq \lim_{t\to\infty} \frac{t}{N_t} \leq \mu,$$

presque sûrement. □

Théorème 11.10. *Supposons que $0 < \sigma^2 = \text{Var}(X_1) < \infty$. Alors la variable aléatoire*

$$\frac{N_t - (t/\mu)}{\sqrt{t\sigma^2/\mu^3}}$$

converge en loi vers une variable aléatoire $\mathcal{N}(0,1)$, lorsque $t \to \infty$.

Démonstration. Fixons $x \in \mathbb{R}$. Alors

$$\mathbb{P}(\frac{N_t - (t/\mu)}{\sqrt{t\sigma^2/\mu^3}} \geq x) = \mathbb{P}(N_t \geq (t/\mu) + x\sqrt{t\sigma^2/\mu^3}) = \mathbb{P}(T_{a(t)} \leq t),$$

où $a(t) = \lceil (t/\mu) + x\sqrt{t\sigma^2/\mu^3} \rceil$. À présent,

$$\mathbb{P}(T_{a(t)} \leq t) = \mathbb{P}(\frac{T_{a(t)} - \mu a(t)}{\sigma\sqrt{a(t)}} \leq \frac{t - \mu a(t)}{\sigma\sqrt{a(t)}}).$$

D'une part,

$$\lim_{t\to\infty} \frac{t - \mu a(t)}{\sigma\sqrt{a(t)}} = -x.$$

D'autre part, on vérifie aisément que le Théorème central limite implique la convergence en loi de $(T_{a(t)} - \mu a(t))/(\sigma\sqrt{a(t)})$ vers une variable aléatoire $\mathcal{N}(0,1)$, lorsque $t \to \infty$. Par conséquent,

$$\lim_{t\to\infty} \mathbb{P}(\frac{N_t - (t/\mu)}{\sqrt{t\sigma^2/\mu^3}} \geq x) = \Phi(-x).$$

□

Remarque 11.6. *On peut établir des résultats analogues sur le comportement asymptotique de la fonction de renouvellement $m(t)$. Nous ne le ferons pas ici, car les preuves sont plus délicates. On peut montrer en particulier que*

$$\lim_{t\to\infty} \frac{m(t)}{t} = \frac{1}{\mu},$$

et, pour tout $h > 0$,

$$\lim_{t\to\infty} \big(m(t+h) - m(t)\big) = \frac{h}{\mu}.$$

Chapitre 12

Introduction à la statistique

Dans ce chapitre, nous présentons une brève introduction aux méthodes statistiques. Il est important d'observer que le point de vue de ce chapitre est très différent de celui des autres chapitres, dont la nature est plus probabiliste. Plutôt que de se donner à priori un espace de probabilité (ou une collection de variables aléatoires de lois données) 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 indépendantes, et on cherche à déterminer la loi des variables aléatoires correspondantes.

12.1 Estimateurs

12.1.1 Définition, consistance, biais

Soit $\mathbb{P}$ une mesure de probabilité sur $\mathbb{R}^d$.

Définition 12.1. *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 12.1.
- ▷ 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 $p \in [0, 1]$.
- ▷ Durée de vie de composants électroniques. Dans ce cas, on modélise les durées de vie par une famille de variables aléatoires i.i.d. de loi exponentielle de paramètre $\lambda > 0$.
- ▷ Répartition de la taille des individus dans une population homogène. On peut 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 $\mathbb{P}$ connue, dépendant d'un ou plusieurs paramètres, en général inconnus ; on notera θ la collection de paramètres, Θ l'ensemble des valeur que θ peut prendre, et $\mathbb{P}_\theta$ la loi correspondante. Pour les exemples précédents :
- ▷ $\theta = p \in \Theta = [0, 1]$.
- ▷ $\theta = \lambda \in \Theta = \mathbb{R}_+^*$.
- ▷ $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R}_+^* \times \mathbb{R}_+^*$.

Le problème fondamental est de prédire (une valeur approchée de) θ à partir des données (c'est-à-dire du n-échantillon). On parle alors d'**estimation paramétrique**.

Définition 12.2. *Soit $X_1, \dots, X_n$ un n-échantillon.*
On appelle ***statistique*** *toute fonction mesurable $F(X_1, \dots, X_n)$.*
On appelle ***estimateur de*** *$f(\theta)$ toute statistique à valeurs dans $f(\Theta)$, utilisée pour estimer $f(\theta)$.*

Insistons sur le fait qu'un estimateur est une fonction de l'échantillon, et ne dépend pas de θ.

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 que l'on cherche à estimer.

Définition 12.3. *Un estimateur T_n de $f(\theta)$ est* ***consistant*** *(ou* ***convergent****) s'il converge en probabilité vers $f(\theta)$,*

$$\lim_{n\to\infty} \mathbb{P}_\theta(|T_n - f(\theta)| \geq \epsilon) = 0, \quad \forall \epsilon > 0, \forall \theta \in \Theta.$$

Exemple 12.2. La moyenne empirique

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

est un estimateur de $f(\theta) = \mathbb{E}_\theta(X)$. La loi des grands nombres implique que cet estimateur est consistant.

Une caractéristique classique d'un estimateur est son biais.

Définition 12.4. *Le* ***biais*** *d'un estimateur T de $f(\theta)$ est défini par $\mathbb{E}_\theta(T - f(\theta)) = \mathbb{E}_\theta(T) - f(\theta)$. On dit que T est un estimateur* ***sans biais*** *de $f(\theta)$ si $\mathbb{E}_\theta(T) = f(\theta)$, $\forall \theta \in \Theta$, sinon on dit qu'il est* ***biaisé****.*

Insister sur l'absence de biais est utile lorsqu'on veut démontrer l'optimalité de certains estimateurs dans une certaine classe ; dans la pratique, ce n'est pas une condition toujours désirable : il est tout à fait possible qu'un estimateur biaisé soit meilleur qu'un estimateur sans biais. Nous reviendrons sur ce point plus tard.

Définition 12.5. *Une famille d'estimateurs $(T_n)_{n\geq 1}$ est appelée* ***estimateur asymptotiquement sans biais de*** *$f(\theta)$ si*

$$\lim_{n\to\infty} (\mathbb{E}_\theta(T_n) - f(\theta)) = 0, \ \forall \theta \in \Theta.$$

Proposition 12.1. *Si T_n est un estimateur de $f(\theta)$ asymptotiquement sans biais, et tel que sa variance tende vers 0 lorsque $n \to \infty$, alors T_n est un estimateur consistant de $f(\theta)$.*

Démonstration. Soit $\epsilon > 0$. Par le Théorème 7.3,

$$\mathbb{P}_\theta(|T_n - f(\theta)| \geq \epsilon) = \mathbb{P}_\theta((T_n - f(\theta))^2 \geq \epsilon^2) \leq \epsilon^{-2}\mathbb{E}_\theta((T_n - f(\theta))^2),$$

pour tout $\theta \in \Theta$. Puisque $\mathbb{E}_\theta((T_n - f(\theta))^2) = \mathrm{Var}_\theta(T_n) + (\mathbb{E}_\theta(T_n - f(\theta)))^2$, et que chacun de ces deux termes tend vers 0 par hypothèse, la conclusion suit. □

12.1.2 Quelques exemples

Moyenne empirique

Soit $X_1, \ldots, X_n$ un n-échantillon de loi $\mathbb{P}_\theta$. On cherche à estimer $f(\theta) = \mathbb{E}_\theta(X_1)$. Un estimateur naturel est la moyenne de l'échantillon :

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

Comme mentionné plus haut, sa consistance suit de la loi des grands nombres. D'autre part,

$$\mathbb{E}_\theta(\bar{X}_n) = \frac{1}{n}(\mathbb{E}_\theta(X_1) + \cdots + \mathbb{E}_\theta(X_n)) = \mathbb{E}_\theta(X_1) = f(\theta),$$

et il s'agit donc d'un estimateur sans biais de $f(\theta)$.

Variance empirique

On désire à présent estimer la variance σ^2 de X_1. Un estimateur naturel est

$$\tilde{\sigma}_n^2 = \frac{1}{n}(X_1^2 + \cdots + X_n^2) - (\frac{1}{n}(X_1 + \cdots + X_n))^2.$$

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

$$\mathbb{E}_\theta\left(\frac{1}{n}(X_1^2 + \cdots + X_n^2)\right) = \mathbb{E}_\theta(X_1^2),$$
$$\mathbb{E}_\theta\left((\frac{1}{n}(X_1 + \cdots + X_n))^2\right) = \frac{1}{n}\mathbb{E}_\theta(X_1^2) + \frac{n-1}{n}\mathbb{E}_\theta(X_1)^2,$$

et donc

$$\mathbb{E}_\theta(\tilde{\sigma}_n^2) = \frac{n-1}{n}(\mathbb{E}_\theta(X_1^2) - \mathbb{E}_\theta(X_1)^2) = \frac{n-1}{n}\sigma^2.$$

Cet estimateur est donc biaisé. On voit qu'un estimateur non biaisé de la variance est donné par

$$S_n^2 = \frac{n}{n-1}\tilde{\sigma}_n^2.$$

Covariance empirique

On considère un n-échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$, et on cherche à estimer la covariance de X et Y. Des considérations tout à fait similaires à celles faites ci-dessus pour la variance montrent que l'estimateur naturel

$$\tilde{\tau}_n = \frac{1}{n}(X_1Y_1 + \cdots + X_nY_n) - (\frac{1}{n}(X_1 + \cdots + X_n))(\frac{1}{n}(Y_1 + \cdots + Y_n))$$

est consistant et biaisé, mais que l'estimateur

$$\hat{\tau}_n = \frac{n}{n-1}\tilde{\tau}_n$$

est consistant et sans biais.

Méthode de Monte-Carlo.

On cherche à estimer numériquement

$$I = \int_a^b h(x)\,\mathrm{d}x,$$

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 suit une loi uniforme sur $[a, b]$. On va estimer I à l'aide de l'estimateur

$$\hat{I} = (b-a)\frac{1}{n}(h(U_1) + \cdots + h(U_n)),$$

où $U_1, \ldots, U_n$ est un n-échantillon de loi uniforme sur $[a, b]$. $\hat{I}$ est un estimateur sans biais et consistant de I.

12.1.3 Construction d'estimateurs

Un problème important est de trouver une façon de construire des estimateurs de $f(\theta)$. Nous verrons deux méthodes : la méthode des moments, et le maximum de vraisemblance.

Méthode des moments

Soit $X_1, \ldots, X_n$ un n-échantillon de loi $\mathbb{P}_\theta$. Supposons que $\theta = \mathbb{E}_\theta(g(X_1))$. Alors, on peut estimer θ à l'aide de l'estimateur naturel

$$\hat{\theta} = \frac{1}{n}(g(X_1) + \cdots + g(X_n)),$$

et on vérifie immédiatement que ce dernier est consistant et sans biais. Par exemple, si $X_1, \ldots, X_n$ 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 $\hat{\theta} = 2\bar{X}_n$ pour estimer, sans biais, θ.

Un choix classique, qui donne son nom à la méthode, correspond à considérer $g(x) = x^r$, ce qui permet d'estimer θ lorsque ce dernier peut s'exprimer en termes des moments $\mathbb{E}_\theta(X^r)$, $\theta = h(\mathbb{E}_\theta(X^r))$: on considère alors l'estimateur, en général biaisé,

$$\tilde{\theta} = h(\frac{1}{n}(X_1^r + \cdots + X_n^r)).$$

Exemple 12.3. Si $X_1, \ldots, X_n$ est un n-échantillon de loi exponentielle de paramètre θ, alors puisque

$$\mathbb{E}_\theta(X_1) = 1/\theta,$$

on peut utiliser $\hat{\theta} = 1/\bar{X}_n$ 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 $x_1, \dots, x_n$ d'un n-échantillon, une approche naturelle au problème de l'estimation est la suivante : on cherche, 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 12.6. *La* **vraisemblance** *(ou* **fonction de vraisemblance***), notée $L(\theta; x_1, \dots, x_n)$, d'un modèle en $x_1, \dots, x_n$ est la probabilité d'observer $\{X_1 = x_1, \dots, X_n = x_n\}$ lorsque le paramètre est θ.*

Remarque 12.1. *Insistons sur le fait que la variable est θ ; $x_1, \dots, x_n$ sont des paramètres.*

Par indépendance des observations, on peut écrire

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_i = x_i).$$

La définition ci-dessus n'a de sens que pour des variables aléatoires discrètes. Dans le cas continu, on travaille avec les densités :

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^{n} f_\theta(x_i),$$

où f_θ est la densité associée à la loi $\mathbb{P}_\theta$.

Définition 12.7. *On appelle* **estimateur du maximum de vraisemblance** *de θ la variable aléatoire correspondant à la valeur $\hat{\theta}(X_1, \dots, X_n)$ en laquelle la fonction de vraisemblance atteint son maximum.*

Proposition 12.2. *Si $\hat{\theta}$ est l'estimateur du maximum de vraisemblance de θ et f est injective, alors $f(\hat{\theta})$ est l'estimateur du maximum de vraisemblance de $f(\theta)$.*

Démonstration. Évident. □

Exemples

Loi exponentielle de paramètre λ. La fonction de vraisemblance est ($x_i > 0$, $i = 1, \dots, n$)

$$L(\lambda; x_1, \dots, x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda(x_1 + \cdots + x_n)}.$$

Pour trouver le maximum, on considère la **log-vraisemblance**,

$$\log L(\lambda; x_1, \dots, x_n) = n \log \lambda - \lambda(x_1 + \cdots + x_n).$$

La dérivée de cette dernière s'annule en $\lambda = n/(x_1 + \cdots + x_n)$, et on vérifie qu'il s'agit d'un maximum. L'estimateur du maximum de vraisemblance de λ est donc

$$\hat{\lambda} = \frac{n}{X_1 + \cdots + X_n}.$$

Loi normale $\mathcal{N}(\mu, 1)$, $\mu \in \mathbb{R}$. Un calcul similaire au précédent (exercice) montre que l'estimateur du maximum de vraisemblance est donné par

$$\hat{\mu} = \frac{X_1 + \cdots + X_n}{n}.$$

Loi normale $\mathcal{N}(0, \sigma^2)$. Le même type de calcul (exercice) montre que l'estimateur du maximum de vraisemblance est donné par

$$\hat{\sigma}^2 = \frac{X_1^2 + \cdots + X_n^2}{n}.$$

Loi normale $\mathcal{N}(\mu, \sigma^2)$. 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 2 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

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$

Loi uniforme sur $[0, \theta]$, $\theta > 0$. La fonction de vraisemblance prend la forme

$$L(\theta; x_1, \ldots, x_n) = \frac{1}{\theta^n}\prod_{i=1}^{n} \mathbf{1}_{\{x_i \leq \theta\}} = \frac{1}{\theta^n}\mathbf{1}_{\{\max_i x_i \leq \theta\}}.$$

La fonction de vraisemblance est nulle si $\theta < \max_i x_i$. Supposons donc que $\theta \geq \max_i x_i$. Dans ce cas, $L(\theta; x_1, \ldots, x_n) = \theta^{-n}$, 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

$$\hat{\theta} = \max\{X_1, \ldots, X_n\}.$$

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 le faire est de considérer la dispersion de la loi de l'estimateur, puisque celle-ci représente l'erreur typique que l'on fait lors d'une application.

Définition 12.8. *Le **risque quadratique** de l'estimateur $\hat{\theta}$ de θ est défini par*

$$\mathcal{R}_{\hat{\theta}}(\theta) = \mathbb{E}_\theta((\hat{\theta} - \theta)^2).$$

Définition 12.9. *Si $\hat{\theta}$ et $\tilde{\theta}$ sont deux estimateurs de θ, on dira que $\hat{\theta}$ est meilleur que $\tilde{\theta}$ si $\mathcal{R}_{\hat{\theta}}(\theta) < \mathcal{R}_{\tilde{\theta}}(\theta)$, $\forall \theta \in \Theta$.*

Similairement, si on veut estimer $f(\theta)$ avec un estimateur T, alors le risque quadratique de T est défini par

$$\mathcal{R}_T(\theta) = \mathbb{E}_\theta((T - f(\theta))^2).$$

Lemme 12.1. *Soit $\hat{\theta}$ un estimateur de θ. Alors*

$$\mathcal{R}_{\hat{\theta}}(\theta) = \mathrm{Var}_\theta(\hat{\theta}) + \left(\mathbb{E}_\theta(\hat{\theta} - \theta)\right)^2.$$

En particulier, si $\hat{\theta}$ est sans biais, alors

$$\mathcal{R}_{\hat{\theta}}(\theta) = \mathrm{Var}_\theta(\hat{\theta}).$$

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 décroitre la variance.

Exemple 12.4. On considère un n-échantillon distribué uniformément sur $[0, \theta]$, $\theta > 0$. Le risque associé à l'estimateur

$$\bar{\theta} = \frac{2}{n}(X_1 + \cdots + X_n)$$

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,

$$\tilde{\theta} = \max\{X_1, \ldots, X_n\}.$$

Manifestement, cet estimateur est biaisé, puisqu'on a toujours $\mathbb{E}(\tilde{\theta}) < \theta$. Commençons par déterminer la loi de $\tilde{\theta}$:

$$\mathbb{P}_\theta(\tilde{\theta} \leq x) = \mathbb{P}_\theta(X_1 \leq x, \ldots, X_n \leq x) = \left(\mathbb{P}_\theta(X_1 \leq x)\right)^n = \left(\frac{x}{\theta}\right)^n,$$

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,

$$\mathcal{R}_{\tilde{\theta}}(\theta) = \frac{2\theta^2}{(n+1)(n+2)}.$$

On peut à présent comparer les 2 estimateurs ci-dessus : on voit que $\mathcal{R}_{\tilde{\theta}}(\theta) \geq \mathcal{R}_{\bar{\theta}}(\theta)$, pour tout $\theta > 0$, et tout $n \geq 1$, l'inégalité étant stricte dès que $n \geq 3$. 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 $a \leq \theta \leq b$, avec une estimation de la confiance que l'on peut avoir en cette affirmation, plutôt qu'une valeur précise. 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 12.10. *Soit $\alpha \in (0, 1)$. Un intervalle $I = I(X_1, \ldots, X_n)$ (aléatoire, ne dépendant pas de θ) est appelé* ***intervalle de confiance pour θ au niveau*** $1 - \alpha$ *si*

$$\mathbb{P}_\theta(I \ni \theta) = 1 - \alpha, \quad \forall \theta \in \Theta.$$

$1 - \alpha$ *est appelé* ***niveau de confiance*** *de l'estimation.*

Exemple 12.5. On considère un n-échantillon avec loi $\mathcal{N}(\mu, 1)$. On a vu que la moyenne empirique $\bar{X}_n$ est un estimateur sans biais de μ. On veut construire un intervalle $[T_1, T_2]$, avec $T_1 = \bar{X}_n - a$ et $T_2 = \bar{X}_n + a$ (intervalle symétrique autour de la moyenne empirique). Puisque $\bar{X}_n$ est une combinaison linéaire de variables aléatoires normales indépendantes, on trouve qu'il suit une loi $\mathcal{N}(\mu, \frac{1}{n})$. Par conséquent $Z = \sqrt{n}(\bar{X}_n - \mu)$ suit une loi $\mathcal{N}(0, 1)$. On a donc

$$\mathbb{P}_\mu(I \ni \mu) = 1 - \alpha \quad \Leftrightarrow \quad \mathbb{P}_\mu(|\bar{X}_n - \mu| \leq a) = \mathbb{P}(|Z| \leq a\sqrt{n}) = 1 - \alpha.$$

Pour $\alpha = 10\%$, on trouve que cette dernière identité est satisfaite si $a\sqrt{n} \simeq 1{,}64$. Par conséquent, l'intervalle

$$I = [\bar{X}_n - \frac{1{,}64}{\sqrt{n}}, \bar{X}_n + \frac{1{,}64}{\sqrt{n}}]$$

est un intervalle de confiance à 90% pour μ.

Exemple 12.6. On considère un n-échantillon distribué uniformément sur $[0, \theta]$, $\theta > 0$. Manifestement, l'estimateur du maximum de vraisemblance $\hat{\theta} = \max\{X_1, \ldots, X_n\}$ satisfait toujours $\hat{\theta} \leq \theta$. On peut donc prendre $T_1 = \hat{\theta}$. On cherche T_2 de la forme $C\hat{\theta}$ avec $\mathbb{P}_\theta(C\hat{\theta} \geq \theta) = 1 - \alpha$. Dans ce cas,

$$I = [\hat{\theta}, C\hat{\theta}]$$

sera un intervalle de confiance au niveau $1 - \alpha$. On a déjà vu que

$$\mathbb{P}_\theta(\hat{\theta} \leq x) = (\frac{x}{\theta})^n.$$

On a donc

$$\mathbb{P}_\theta(C\hat{\theta} \geq \theta) = 1 - \mathbb{P}_\theta(C\hat{\theta} < \theta) = 1 - (\frac{1}{C})^n.$$

L'intervalle recherché est donc

$$I = [\hat{\theta}, \alpha^{-1/n}\hat{\theta}].$$

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 12.11. *Un intervalle $I = I(X_1, \ldots, X_n)$ (indépendant de θ) est un intervalle de confiance pour θ au niveau $1 - \alpha$* ***par excès*** *si*

$$\mathbb{P}_\theta(I \ni \theta) \geq 1 - \alpha, \quad \forall \theta \in \Theta.$$

Exemple 12.7. Soit $X_1, \ldots, X_n$ 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 $\bar{X}_n$ la moyenne empirique, on a par le Théorème 7.3 que

$$\mathbb{P}_\theta(|\bar{X}_n - f(\theta)| < \delta) \geq 1 - \frac{\sigma^2}{n\delta^2}.$$

On en conclut que

$$I = [\bar{X}_n - \frac{\sigma}{\sqrt{n\alpha}}, \bar{X}_n + \frac{\sigma}{\sqrt{n\alpha}}]$$

est un intervalle de confiance par excès au niveau $1 - \alpha$.

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.

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 12.12. *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*

$$\lim_{n\to\infty} \mathbb{P}_\theta(I_n \ni \theta) = 1-\alpha, \quad \forall \theta \in \Theta.$$

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*

$$\lim_{n\to\infty} \mathbb{P}_\theta(I_n \ni \theta) \geq 1-\alpha, \quad \forall \theta \in \Theta.$$

Exemple 12.8. 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,

$$\mathbb{P}_\theta(\bar{X}_n \in [\mu - \frac{a\sigma}{\sqrt{n}}, \mu + \frac{a\sigma}{\sqrt{n}}]) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(Z \in [-a, a]),$$

où Z suit une loi $\mathcal{N}(0,1)$. Si l'on choisit a tel que $\mathbb{P}(Z \in [-a,a]) = 1-\alpha$, l'intervalle

$$I_n = [\bar{X}_n - \frac{a\sigma}{\sqrt{n}}, \bar{X}_n + \frac{a\sigma}{\sqrt{n}}]$$

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

$$I_n = [\bar{X}_n - \frac{1{,}88}{\sqrt{n}}, \bar{X}_n + \frac{1{,}88}{\sqrt{n}}].$$

On choisit à présent le plus petit n tel que $1{,}88/\sqrt{n} \leq 0{,}1$, c'est-à-dire $n \geq 355$.

Exemple 12.9. Considérons maintenant le cas d'un n-échantillon, dont on désire estimer la moyenne $\mu = \mathbb{E}_\theta(X_1)$, sans connaître la variance. On part de l'intervalle obtenu précédemment,

$$I_n = [\bar{X}_n - \frac{a\sigma}{\sqrt{n}}, \bar{X}_n + \frac{a\sigma}{\sqrt{n}}].$$

Ce n'est pas un intervalle de confiance, puisque σ est inconnu. On considère donc l'intervalle

$$J_n = [\bar{X}_n - \frac{aS_n}{\sqrt{n}}, \bar{X}_n + \frac{aS_n}{\sqrt{n}}],$$

où S_n^2 est l'estimateur sans biais de la variance défini par

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

On a vu que S_n^2 est un estimateur consistant de σ^2. On a donc

$$\lim_{n\to\infty} \mathbb{P}_\theta(I_n \ni \mu) = \mathbb{P}(Z \in [-a,a]), \qquad \forall a > 0,$$

$$S_n^2 \overset{\mathbb{P}_\theta}{\to} \sigma^2.$$

On va voir que cela implique que

$$\lim_{n\to\infty} \mathbb{P}_\theta(J_n \ni \mu) = \mathbb{P}(Z \in [-a,a]), \qquad \forall a > 0,$$

et donc que J_n est un intervalle de confiance asymptotique pour μ au niveau $\mathbb{P}(Z \in [-a,a]) = 1-\alpha$. Pour vérifier cela, il suffit d'observer que

$$\mathbb{P}_\theta(J_n \ni \mu) = \mathbb{P}_\theta(J_n \ni \mu, |S_n - \sigma| \le \epsilon) + \mathbb{P}_\theta(J_n \ni \mu, |S_n - \sigma| > \epsilon).$$

Le second terme du membre de droite tend vers 0, puisqu'il est borné supérieurement par $\mathbb{P}_\theta(|S_n - \sigma| > \epsilon)$, qui tend vers 0 pour tout $\epsilon > 0$. Le premier terme du membre de droite peut, lui, être borné supérieurement par

$$\mathbb{P}_\theta([\bar{X}_n - \frac{a(\sigma+\epsilon)}{\sqrt{n}}, \bar{X}_n + \frac{a(\sigma+\epsilon)}{\sqrt{n}}] \ni \mu)$$

qui converge vers $\mathbb{P}(Z \in [-a(1+\epsilon/\sigma), a(1+\epsilon/\sigma)])$. Comme cette borne est valide pour tout $\epsilon > 0$, on obtient

$$\limsup_{n\to\infty} \mathbb{P}_\theta(J_n \ni \mu) \le \mathbb{P}(Z \in [-a,a]).$$

Pour la borne inférieure, on procède similairement

$$\begin{aligned}
&\mathbb{P}_\theta(J_n \ni \mu, |S_n - \sigma| \le \epsilon) \\
&\qquad \ge \mathbb{P}_\theta([\bar{X}_n - \frac{a(\sigma-\epsilon)}{\sqrt{n}}, \bar{X}_n + \frac{a(\sigma-\epsilon)}{\sqrt{n}}] \ni \mu, |S_n - \sigma| \le \epsilon) \\
&\qquad \ge \mathbb{P}_\theta([\bar{X}_n - \frac{a(\sigma-\epsilon)}{\sqrt{n}}, \bar{X}_n + \frac{a(\sigma-\epsilon)}{\sqrt{n}}] \ni \mu) - \mathbb{P}_\theta(|S_n - \sigma| > \epsilon).
\end{aligned}$$

Le second terme du membre de droite tend vers 0, pour tout $\epsilon > 0$, et le premier terme tend vers $\mathbb{P}(Z \in [-a(1-\epsilon/\sigma), a(1-\epsilon/\sigma)])$. Par conséquent,

$$\liminf_{n\to\infty} \mathbb{P}_\theta(J_n \ni \mu) \ge \mathbb{P}(Z \in [-a,a]),$$

et l'affirmation est démontrée.

12.2.3 Normalité asymptotique

On a vu dans les exemples précédents que la convergence de l'estimateur vers une loi normale est particulièrement pratique pour construire des intervalles de confiance.

Définition 12.13. *Une suite d'estimateurs T_n de $f(\theta)$ est* ***asymptotiquement normale*** *s'il existe $\sigma(\theta) > 0$ tels que $\frac{\sqrt{n}}{\sigma(\theta)}(T_n - f(\theta))$ converge en loi $\mathbb{P}_\theta$ vers $\mathcal{N}(0,1)$, pour tout $\theta \in \Theta$.*

Proposition 12.3. *Un estimateur de θ asymptotiquement normal est nécessairement consistant.*

Démonstration. Soit $\epsilon > 0$. On a

$$\mathbb{P}_\theta(|T_n - \theta| \ge \epsilon) = \mathbb{P}_\theta(\sqrt{n}(T_n - \theta) \notin [-\epsilon\sqrt{n}, \epsilon\sqrt{n}]) \le \mathbb{P}_\theta(\sqrt{n}(T_n - \theta) \notin [-A, A]),$$

pour tout $n \ge A^2\epsilon^{-2}$. Par normalité asymptotique, cette dernière probabilité converge vers

$$\mathbb{P}(Z \notin [-A, A]),$$

où Z suit une loi $\mathcal{N}(0, \sigma^2(\theta))$, $\forall \theta \in \Theta$, ce qui tend vers 0 lorsque $A \to \infty$. □

Il y a une façon naturelle de comparer deux estimateurs asymptotiquement normaux.

Définition 12.14. *Si T_n et T'_n sont deux estimateurs asymptotiquement normaux de $f(\theta)$, c'est-à-dire tels que, pour tout $\theta \in \Theta$, il existe $\sigma(\theta)$ et $\sigma'(\theta)$ tels que $\sqrt{n}(T_n - f(\theta))$ converge en loi $\mathbb{P}_\theta$ vers $\mathcal{N}(0, \sigma^2(\theta))$ et $\sqrt{n}(T'_n - f(\theta))$ converge en loi $\mathbb{P}_\theta$ vers $\mathcal{N}(0, \sigma'^2(\theta))$, alors on dit que T_n est meilleur que T'_n si $\sigma^2(\theta) < \sigma'^2(\theta)$, $\forall \theta \in \Theta$.*

On interprète σ^2/n comme le risque quadratique asymptotique de T_n.

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 $\mathbb{P}_\theta(T \geq 2) \geq 0{,}9$, où T est le temps de vie de ces composants, de loi supposée exponentielle de paramètre $1/\theta$. Ceci revient à s'assurer que $\theta \geq -2/\log(0{,}9) = \theta^\star \simeq 19$. 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 $\theta \geq \theta^\star$, 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 1ère 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

$$\mathbb{P}_\theta(\hat{\theta}_n \geq z_0) \leq 5\%, \qquad \forall \theta \in (0, \theta^\star].$$

Dans ce cas, si l'on observe $\hat{\theta}_n \geq z_0$, il ne sera pas raisonnable de supposer que $\theta \in (0, \theta^\star]$, puisque cela arrive dans seulement 5% des cas. Le fabricant rejettera donc l'hypothèse $\theta < \theta^\star$, et aura raison dans 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 $X_1, \ldots, X_n$ 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 12.15. *Une* ***région de rejet*** *est un événement $D = D(X_1, \dots, X_n)$.*

Définition 12.16. *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 12.17. *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 12.18. *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 l'erreur de seconde espèce β. Ce critère permet de comparer des tests.

Définition 12.19. *Une hypothèse H est dite* ***simple*** *si l'ensemble Θ correspondant est réduit à un seul élément, sinon elle est dite* ***composite****.*

Exemple 12.10. 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

$$\mathbb{P}_{\theta_0}(I \not\ni \theta_0) = \alpha.$$

12.3.3 Cas gaussien

On considère un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$.

Test de moyenne à variance connue

Test de « $\mu = \mu_0$ » contre « $\mu \neq \mu_0$ ». Soit $\bar{X}_n$ la moyenne empirique (de loi $\mathcal{N}(\mu, \sigma^2/n)$) ; on prend pour région de rejet

$$D = \{|\bar{X}_n - \mu_0| \geq C\}.$$

On veut un niveau de risque de 5%, c'est-à-dire

$$\mathbb{P}_{\mu_0}(|\bar{X}_n - \mu_0| \geq C) = 0{,}05,$$

et donc $C \simeq 1{,}96\sigma/\sqrt{n}$.

Test de « $\mu \leq \mu_0$ » contre « $\mu > \mu_0$ ». Cette fois, on prend pour région de rejet

$$D = \{\bar{X}_n > C\}.$$

On veut un niveau de risque de 5%, c'est-à-dire

$$\sup_{\mu \leq \mu_0} \mathbb{P}_\mu(D) = \sup_{\mu \leq \mu_0} \mathbb{P}\left(\frac{\sigma}{\sqrt{n}} Z > C - \mu\right) = 0{,}05,$$

où Z est normale standard. La borne supérieure est atteinte pour $\mu = \mu_0$, et on obtient donc $C \simeq \mu_0 + 1{,}64\sigma/\sqrt{n}$.

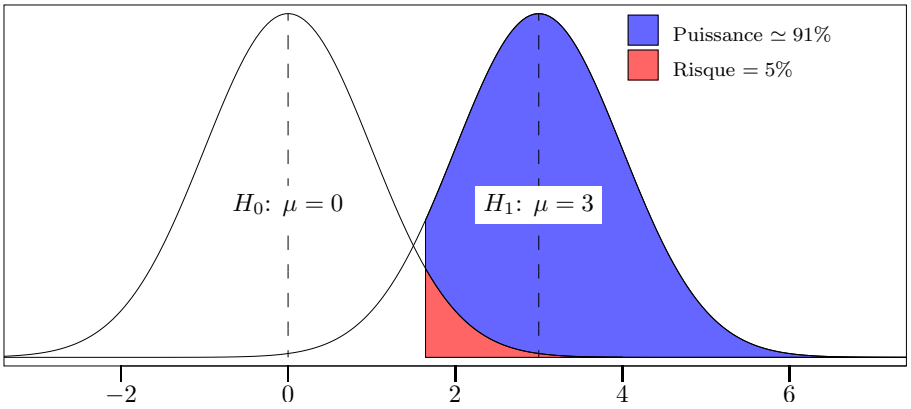


FIGURE 12.1: Test de deux hypothèses simples.

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 précédent : on estime $\mu - \nu$ par $\bar{X}_n - \bar{Y}_m$, qui est de loi $\mathcal{N}(\mu - \nu, \frac{\sigma^2}{n} + \frac{\tau^2}{m})$, et on teste « $\mu - \nu = 0$ » contre « $\mu - \nu \neq 0$ ».

Test de moyenne à variance inconnue

On veut tester « $\mu = \mu_0$ » contre « $\mu \neq \mu_0$ », dans le cas où la variance σ^2 n'est pas connue.

On considère comme estimateurs la moyenne empirique $\bar{X}_n$ et la variance empirique débiaisée S_n^2. Un calcul montre que la variable aléatoire

$$T_{n-1} = \frac{\sqrt{n}}{S_n}(\bar{X}_n - \mu)$$

suit la loi de Student à $n-1$ degrés de liberté.

Prenons $n = 20$, $\mu_0 = 0$ et un risque $\alpha = 5\%$. On choisit comme région de rejet

$$D = \{\frac{|\bar{X}_n - \mu_0|}{S_n} \geq C\},$$

avec C déterminée par la relation

$$\mathbb{P}_{\mu_0}(D) = \mathbb{P}(|T_{n-1}| \geq C\sqrt{n}) = 0{,}05.$$

La loi de Student étant tabulée, on trouve, pour 19 degrés de liberté, $C\sqrt{n} \simeq 2{,}093$, et donc

$$D = \{\frac{|\bar{X}_n|}{S_n} \geq \frac{2{,}093}{\sqrt{20}}\}.$$

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 à l'aide des fonctions de vraisemblance, c'est-à-dire en comparant $L(\theta_0; x_1, \dots, x_n)$ et $L(\theta_1; x_1, \dots, x_n)$. C'est ce qu'on appelle le **test de Neyman**[1]**-Pearson**[2].

1. Jerzy Neyman (1894, Bendery - 1981, Berkeley), statisticien polonais ; un des grands fondateurs de la statistique moderne.

2. Egon Sharpe Pearson (1895, Hampstead - 1980, London), statisticien anglais. Fils du célèbre statisticien Karl Pearson.

L'objet central est le **rapport de vraisemblance**,

$$R(\theta_0, \theta_1; x_1, \dots, x_n) = \frac{L(\theta_1; x_1, \dots, x_n)}{L(\theta_0; x_1, \dots, x_n)}.$$

On prend pour région de rejet

$$D = \{R(\theta_0, \theta_1; X_1, \dots, X_n) > C\},$$

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 12.11. 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 on effectue 100 lancers. Elle obtient $F = 60$ « face ». Comment déterminer quelle pièce a été utilisée ?

Le modèle est clair : on a un $n = 100$-échantillon 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

$$p^f(1-p)^{n-f} = (1-p)^n \left(\frac{p}{1-p}\right)^f.$$

Le rapport de vraisemblance est donc donné, dans la situation étudiée ici, par

$$R = \left(\frac{1-\frac{2}{3}}{1-\frac{1}{2}}\right)^n \left(\frac{\frac{2}{3}/(1-\frac{2}{3})}{\frac{1}{2}/(1-\frac{1}{2})}\right)^f = \left(\tfrac{2}{3}\right)^n 2^f.$$

Il s'agit d'une fonction monotone de f, donc prendre une région de rejet de la forme

$$D = \{R > C\}$$

revient à prendre une région

$$D' = \{F > C'\},$$

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 trouve que $C' \simeq 56{,}4$.

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 12.2 (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 f_{θ_0} et f_{θ_1}. Soient $\alpha \in (0,1)$ et*

$$D = \left\{ (x_1, \dots, x_n) \,:\, \prod_{i=1}^n f_{\theta_1}(x_i) > C \prod_{i=1}^n f_{\theta_0}(x_i) \right\},$$

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

$$\mathbb{P}_{\theta_1}(B) \leq \mathbb{P}_{\theta_1}(D),$$

avec l'inégalité stricte si $\mathbb{P}_{\theta_1}(D \setminus B) > 0$.

Démonstration. Notons $\mathbf{x} = (x_1, \dots, x_n)$, $\mathrm{d}\mathbf{x} = \mathrm{d}x_1 \cdots \mathrm{d}x_n$, et $f(\mathbf{x}) = f(x_1) \cdots f(x_n)$. On a

$$\int_{D \setminus B} f_{\theta_0}(\mathbf{x}) \mathrm{d}\mathbf{x} = \alpha - \int_{D \cap B} f_{\theta_0}(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_{B \setminus D} f_{\theta_0}(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

D'autre part, puisque $D \setminus B \subseteq D$ et $B \setminus D \subseteq D^{\mathrm{c}}$, on déduit de l'identité précédente que

$$\int_{D \setminus B} f_{\theta_1}(\mathbf{x}) \mathrm{d}\mathbf{x} \geq C \int_{D \setminus B} f_{\theta_0}(\mathbf{x}) \mathrm{d}\mathbf{x} = C \int_{B \setminus D} f_{\theta_0}(\mathbf{x}) \mathrm{d}\mathbf{x} \geq \int_{B \setminus D} f_{\theta_1}(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

(La première inégalité est stricte si $\mathbb{P}_{\theta_1}(D \setminus B) > 0$.) On a donc bien

$$\mathbb{P}_{\theta_1}(D) = \mathbb{P}_{\theta_1}(D \setminus B) + \mathbb{P}_{\theta_1}(D \cap B) \geq \mathbb{P}_{\theta_1}(B \setminus D) + \mathbb{P}_{\theta_1}(D \cap B) = \mathbb{P}_{\theta_1}(B).$$

□

Remarque 12.2. *Dans le cas de lois discrètes, un résultat similaire est encore vérifié. Il y a toutefois deux choses à observer : 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 fait des sauts ; d'autre part, l'ensemble $\{(x_1, \dots, x_n) \,:\, p_{\theta_1}(x_1) \cdots p_{\theta_1}(x_n) = C p_{\theta_0}(x_1) \cdots p_{\theta_0}(x_n)\}$ n'a plus nécessairement probabilité nulle. Une manière de résoudre simultanément ces deux problèmes est d'utiliser la procédure suivante. Soit $R(\theta_0, \theta_1; x_1, \dots, x_n)$ 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 de façon à ce 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, et le problème se réduisait 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 $N_n(j) = \sum_{i=1}^n \mathbf{1}_{\{X_i = j\}}$ est le nombre

d'occurrence 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 $H_0 : p_i = q_i$, $i = 1, \dots, k$, contre $H_1 : \exists j : q_j \neq p_j$.

q nous donne donc les probabilités de chacune des classes sous l'hypothèse nulle, et on est donc 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.

Proposition 12.4. *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 nous rappelons que la densité est

$$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}\, \mathbf{1}_{[0,\infty)}(x).$$

Remarque 12.3. *La raison pour laquelle on a $k-1$ degrés de liberté et non k est qu'une contrainte lie les N_i : $N_1 + \cdots + N_k = n$.*

Ainsi, 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 diverge.

On peut donc prendre une région de rejet de la forme

$$D = \{Z_n > C\},$$

en choisissant C de sorte que

$$\lim_{n\to\infty} \mathbb{P}_q(Z_n > C) = \mathbb{P}(\chi^2_{k-1} > C) = \alpha.$$

Remarque 12.4. *Il est important de réaliser qu'il s'agit d'une approximation asymptotique. Pour qu'elle soit applicable en pratique, il faut que les effectifs théoriques nq_j soient supérieurs à 5.*

Exemple 12.12. 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 jaune et ronds avec une fréquence $9/16$, jaune et ridé avec une fréquence $3/16$, vert et rond avec une fréquence $3/16$, et vert et ridé 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 de l'usage du test d'adéquation du χ^2. On obtient,

$$\begin{aligned} Z_{556} &= \frac{\left(315 - 556 \cdot \frac{9}{16}\right)^2}{556 \cdot \frac{9}{16}} + \frac{\left(101 - 556 \cdot \frac{3}{16}\right)^2}{556 \cdot \frac{3}{16}} + \frac{\left(108 - 556 \cdot \frac{3}{16}\right)^2}{556 \cdot \frac{3}{16}} + \frac{\left(32 - 556 \cdot \frac{1}{16}\right)^2}{556 \cdot \frac{1}{16}} \\ &\simeq 0{,}47. \end{aligned}$$

Pour un seuil de 5%, on obtient que $\mathbb{P}_{H_0}(\chi^2_3 > C) = 0{,}05$ 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.

	ch. blonds	ch. bruns	ch. roux	ch. noirs	total	fréquence
y. bleus	25	9	7	3	44	44/124
y. gris	13	17	7	10	47	47/124
y. 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 connait pas ces

3. Johann Gregor Mendel (1822, Heinzendorf – 1884, Brünn), moine et botaniste Autrichien. Il est communément reconnu comme le père fondateur de la génétique.

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 qu'il faut observer que sur les $4 \cdot 3 = 12$ fréquences empiriques, seules $3 \cdot 2 = 6$ sont indépendantes. On doit donc considérer une variable de loi χ^2_6.

En procédant comme précédemment, on arrive à la conclusion qu'avec un seuil de 5%, l'hypothèse nulle (d'indépendance) doit être rejetée.

Index

accroissement, 169
    indépendant, 169
    stationnaire, 169
amas, 159
Avogadro, Lorenzo Romano Amedeo Carlo, 134

Bernoulli
    Daniel, 6
    Jacques, 6
Berry, Andrew C., 126
Berry–Esséen (inégalité de), 126
biais, 184
Bienaymé, Irénée-Jules, 46
Bonferroni, Carlo Emilio, 15
Borel, Félix Édouard Justin Émile, 6
Borel-Cantelli (lemmes de), 118
Brown, Robert, 134

Cardano, Girolamo, 5
Cauchy, Augustin Louis, 97
chaîne de Markov, 141
    absorbante, 143
    apériodique, 154
    ergodique, 154
    irréductible, 143
    récurrente, 150
    récurrente-positive, 151
    renversée, 156
    réversible, 156
coefficient de corrélation, 48
condition d'équilibre local, 156
conditions de consistance de Kolmogorov, 117
confiance, 193, 194
convergence
    en loi, 121
    en moyenne, 121
    en probabilité, 121
    presque sûre, 121
convexité, 44
    stricte, 44
couplage, 163
covariance, 47

densité
    conditionnelle, 104
    conjointe, 100
    marginale, 100
distribution stationnaire, 152
distribution uniforme, 16

écart-type, 46
échantillon, 183
échantillon aléatoire, 17
Ehrenfest
    Paul, 145
    Tatiana Alexeyevna Afanaseva, 145
Einstein, Albert, 134
épreuve de Bernoulli, 33
équation de renouvellement, 180
équiprobabilité, 16
Erdős, Pál, 16
erreur
    première espèce, 193
    seconde espèce, 193
espace des états, 141
espace des observables, 6
espace échantillon, 6
espérance
    variables aléatoires discrètes, 39
    vecteur aléatoire, 50
espérance conditionnelle, 53, 105
Esséen, Carl-Gustav, 126
estimateur
    maximum de vraisemblance, 187
    normalité asymptotique, 192

estimation paramétrique, 183
état
    absorbant, 143
    apériodique, 154
    atteignable, 143
    période, 154
    périodique, 154
    récurrent, 150
Euler, Leonhard, 6
événement
    asymptotique, 128
    composite, 7
    disjoints, 7
    élémentaire, 7
    incompatibles, 7

Fermat, Pierre de, 6
fonction caractéristique, 109
    conjointe, 111
fonction de densité, 176
fonction de masse, 13, 32
    conditionnelle, 53
    conjointe, 38
    marginale, 38
fonction de renouvellement, 180
fonction gamma, 97
fonction génératrice, 65
    fonction génératrice conjointe, 75
    fonction génératrice des moments, 67
fonction harmonique, 148
fonction indicatrice, 33
formule de Bayes, 23

Galilée, 6
Gauss, Johann Carl Friedrich, 6
Gosset, William Sealy, 99
grande déviation, 124
graphe aléatoire, 16

Huygens, Christiaan, 6
hypothèse
    alternative, 193
    composite, 194
    nulle, 193
    simple, 194

indépendance
    év. deux-à-deux indép., 27
    événements indépendants, 27
    indépendance conditionnelle, 29
    variables aléatoires, 36
inégalités de Bonferroni, 15
inégalité de Cauchy-Schwarz, 48
inégalité de Jensen, 45
intervalle de confiance, 189
    asymptotique, 191
    asymptotique par excès, 191
    par excès, 190
Ising, Ernst, 161

Kepler, Johannes, 6
Kolmogorov, Andreï Nikolaïevich, 6

Laplace, Pierre-Simon, 6
Lebesgue, Henri Léon, 6
loi, 32
    χ^2, 97
    beta, 97
    binomiale, 33
    binomiale négative, 36
    de Bernoulli, 33
    de Cauchy, 97
    de Pascal, 36
    de Poisson, 33
    de Student, 99
    de Weibull, 99
    gamma, 96
    gaussienne, 96
    géométrique, 35
    hypergéométrique, 35
    multinomiale, 198
    normale, 96
    normale standard, 96
    t, 99
    uniforme, 94
Loi 0-1 de Kolmogorov, 128
loi conjointe, 38
loi de la probabilité totale, 23
loi de Pólya, 36
loi des petits nombres, 34
loi faible des grands nombres, 51, 122
loi forte des grands nombres, 124
lois fini-dimensionnelle, 117

marche aléatoire, 134
    simple
        symétrique, 56
matrice de covariance, 50
matrice de transition, 141
matrice fondamentale, 147
matrice stochastique, 142
modèle booléen, 179

de Moivre, Abraham, 6
moment, 45
mouvement brownien, 139
moyenne empirique, 51

Newton, Sir Isaac, 21
Neyman, Jerzy, 195

paradoxe de Simpson, 26
partition, 23
Pascal, Blaise, 6
Pearson, Egon Sharpe, 195
Peierls, Rudolf Ernst, 161
percolation, 159
Perrin, Jean Baptiste, 134
perte de mémoire, 35
Poisson, Siméon Denis, 33
Pólya, George, 136
principe d'indifférence, 9
principe d'invariance, 139
principe de réflexion, 59
probabilité conditionnelle, 22
probabilités de transition, 141
processus de branchement, 69
processus de comptage, 165
processus de Poisson, 166
    amincissement, 175
    fonction de valeur moyenne, 176
    intensité, 166
    non homogène, 176
    processus de Poisson composé, 177
    spatial, 178
    superposition, 174
processus de renouvellement, 166
processus de Wiener, 139
propriété de Markov, 141
puissance, 194

réalisation, 7, 183
récurrence, 74
    nulle, 74
    positive, 74
région de rejet, 194
Rényi, Alfréd, 16
risque, 194
risque quadratique, 188

seuil, 193
statistique, 184
statistiques d'ordre, 173
Stirling, James, 21
symbole de Pochhammer, 16

Tchebychev, Pafnouti Lvovitch, 46
temps de récurrence, 151
test, 194
    d'adéquation, 197
    d'ajustement, 197
    de Neyman-Pearson, 195
    non paramétrique, 197
    paramétrique, 197
théorème central limite, 126
tirage
    tirage avec remise, 17
    tirage sans remise, 17
transience, 74
tribu
    asymptotique, 128
    engendrée par des v.a., 128
    triviale, 128

univers, 6

Varadhan, S. R. S., 124
variable aléatoire
    asymptotique, 129
    défective, 32
    i.i.d., 37
    v.a. non-corrélées, 47
variable aléatoire discrète, 31
variance, 46
vecteur aléatoire, 38
    à densité, 100
    gaussien, 103
vraisemblance, 187

Weibull, Ernst Hjalmar Waloddi, 99
Wiener, Norbert, 139