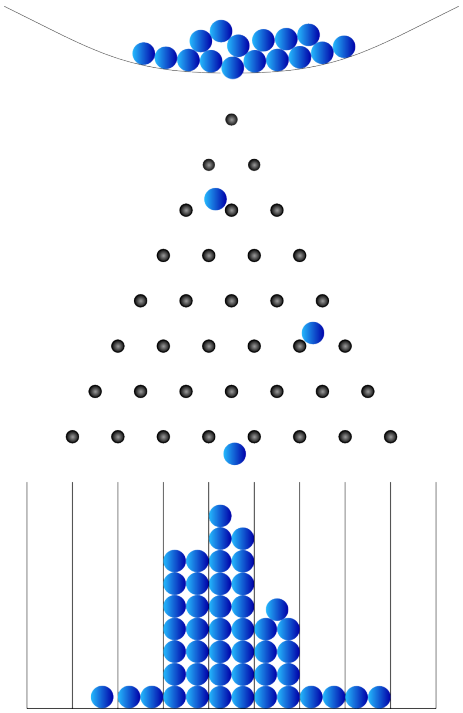
UNIVERSITÉ DE GENÈVE

Probabilités et Statistique

Y. Velenik



— Version du 24 mai 2012 —

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

2011-2012

Table des matières

Chapitre 1

Introduction

Si la théorie des probabilités a été originellement motivée par l'analyse des jeux de hasard, elle a pris 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 a aussi de nombreuses applications directes, par exemple en fiabilité, ou dans les assurances et dans 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'oeuvre, 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 en 1563), ainsi qu'à Kepler [2] et Galilée footnoteGalilée ou Galileo Galilei (1564, Pise - 1642, Arcetri), physicien et astronome italien.. Toutefois, il est généralement admis que la théorie des probabilités débute réellement avec les travaux de Pascal [3] et de Fermat [4]. La théorie fut ensuite développée par de nombreuses personnes, dont Huygens [5], J. Bernoulli [6], de Moivre [7], D. Bernoulli [8], Euler [9], Gauss [10] et Laplace [11]. La théorie moderne des probabilités est fondée sur l'approche axiomatique de Kolmogorov [12], basée sur la théorie de la mesure de Borel [13] et Lebesgue [14]. Grâce à cette approche, la théorie a alors connu un développement très rapide tout au long du XX$^{\text{ème}}$ siècle.

1.1 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.

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

1.1.1 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 1.1.1.

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. *Taille d'une personne :* $\Omega = \mathbb{R}^+$.
5. *Durée de vie d'une ampoule :* $\Omega = \mathbb{R}^+$.
6. *Le 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.
7. *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 quatrième, on pourrait limiter, par exemple, Ω à $[0,3]$ (mètres), voire à $\{1,2,\ldots,3000\}$ (millimètres), sans perte de généralité.

1.1.2 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 1.1.2. *On lance successivement deux dés,* $\Omega = \{(m,n) \in \{1,2,3,4,5,6\}\times\{1,2,3,4,5,6\}\}$.

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

L'ensemble des événements associés à une expérience aléatoire est donc un sous-ensemble $\mathcal{F}$ des parties de Ω, $\mathcal{F} \subseteq \mathscr{P}(\Omega)$. Il pourrait paraître raisonnable de prendre $\mathcal{F} = \mathscr{P}(\Omega)$, mais nous verrons par la suite qu'il est alors en général impossible d'associer à chaque événement une probabilité de façon cohérente. Il est donc nécessaire en général de se restreindre à un sous-ensemble strict de $\mathscr{P}(\Omega)$, contenant les événements « intéressants ». Quelle que soit la notion d'« intéressant » que l'on choisisse, il est naturel d'exiger que $\mathcal{F}$ possède un certain nombre de propriétés : si A est un événement intéressant, alors son

complémentaire A^c est également intéressant, puisque demander si A^c est réalisé est équivalent à demander si A ne l'est pas ; de même, si A et B sont des événements intéressants, leur conjonction $A \cap B$ est également intéressante, puisque demander si $A \cap B$ est réalisé revient à demander si A est réalisé et si B est réalisé.

Définition 1.1.1. *Un ensemble $\mathcal{F}$ de parties d'un ensemble Ω est une* algèbre *sur Ω s'il satisfait aux trois conditions suivantes :*

1. $\Omega \in \mathcal{F}$;
2. $A \in \mathcal{F} \implies A^c \in \mathcal{F}$;
3. $A,B \in \mathcal{F} \implies A \cap B \in \mathcal{F}$.

Exemple 1.1.3.
- $\mathscr{P}(\Omega)$ *est une algèbre sur Ω, l'*algèbre triviale *sur Ω.*
- $\{\varnothing,\Omega\}$ *est une algèbre sur Ω, l'*algèbre grossière *sur Ω.*
- *Si $A \subset \Omega$, $\{\varnothing,A,A^c,\Omega\}$ est une algèbre sur Ω.*
- *L'ensemble formé de $\mathbb{R}$, $\varnothing$, et des unions* finies *d'intervalles de la forme*

$$[a,b], (a,b), (a,b], [a,b), (-\infty,a], (-\infty,a), [a,+\infty), (a,+\infty),$$

avec $a \leq b \in \mathbb{R}$, forme une algèbre sur $\mathbb{R}$.

Définition 1.1.2. *Introduisons un peu de terminologie. 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 $A \in \mathcal{F}$, on appelle A^c l'*événement contraire *de A. Si $A,B \in \mathcal{F}$, 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.

Évidemment il suit de la définition que si $\mathcal{F}$ est une algèbre sur Ω, alors $\varnothing \in \mathcal{F}$ (combiner les conditions *1.* et *2.*), et que si $A,B \in \mathcal{F}$, alors $A \cup B \in \mathcal{F}$ (combiner les trois conditions).

En itérant la propriété *3.*, il suit que l'intersection de toute famille finie $A_1,\ldots,A_n \in \mathcal{F}$ est également dans $\mathcal{F}$,

$$A_1,\ldots,A_n \in \mathcal{F} \implies A_1 \cap \cdots \cap A_n \in \mathcal{F},$$

et donc également

$$A_1,\ldots,A_n \in \mathcal{F} \implies A_1 \cup \cdots \cup A_n \in \mathcal{F}.$$

Par contre, le fait que $\mathcal{F}$ soit une algèbre n'implique pas que l'union ou l'intersection d'une collection *infinie* $A_1,A_2,\ldots$ d'événements soient également dans $\mathcal{F}$. De nombreux événements importants s'expriment toutefois comme union ou intersection d'un nombre infini d'événements.

Exemple 1.1.4. *On considère une expérience consistant à jeter une infinité de fois une pièce de monnaie. On a donc comme univers* $\Omega = \{a_1a_2a_3\ldots : a_i \in \{0,1\}\}$*, l'ensemble des suites infinies de* 0 *et de* 1*, où l'on a décidé de représenter par* 0*, resp.* 1*, un pile, resp. face. On considère l'ensemble* $\mathcal{A}$ *composé des sous-ensembles de* Ω *de la forme*

$$\{\omega \in \Omega : (a_1,\ldots,a_n) \in A\},$$

avec $n \geq 1$ *un entier arbitraire et* $A \subseteq \{0,1\}^n$*. On vérifie facilement que* $\mathcal{A}$ *contient* $\varnothing$ *(en prenant* $n = 1$ *et* $A = \varnothing$*) et* Ω *(en prenant* $n = 1$ *et* $A = \{0,1\}$*), et que* $\mathcal{A}$ *est une algèbre.*

Un événement intéressant[15] *est «* $\frac{1}{n}\sum_{i=1}^n a_i$ *converge vers* $\frac{1}{2}$ *», qui affirme que si l'on lance un grand nombre de fois une pièce de monnaie, pile est sorti en moyenne une fois sur deux. Or cet événement ne fait pas partie de* $\mathcal{A}$ *: on voit en effet immédiatement qu'il ne dépend pas des premiers termes* $a_1,\ldots,a_n$*,* quel que soit n fixé*, alors qu'un événement de* $\mathcal{A}$ *doit toujours pouvoir, par définition, s'exprimer en fonction du début de la suite infinie de lancers.*

Pour cette raison on remplace habituellement la contrainte que $\mathcal{F}$ est une algèbre par la contrainte plus forte que $\mathcal{F}$ est une σ-algèbre, ou tribu, sur Ω.

Définition 1.1.3. *Une algèbre sur* Ω *est une* σ*-algèbre, ou tribu, sur* Ω *si*

3'. $A_1,A_2,\ldots \in \mathcal{F} \implies \bigcap_{i=1}^\infty A_i \in \mathcal{F}$.

Comme précédemment, si $\mathcal{F}$ est une tribu sur Ω, il suit que

$$A_1,A_2,\ldots \in \mathcal{F} \implies \bigcup_{i=1}^\infty A_i \in \mathcal{F}.$$

Une tribu est toujours une algèbre, mais la réciproque n'est pas vraie.

Exemple 1.1.5.

1. *Les trois premiers exemples de l'Exemple 1.1.3 sont des tribus (mais pas le quatrième).*
2. *Revenons à l'Exemple 1.1.4. Soit* $\mathcal{F}$ *une tribu contenant* $\mathcal{A}$*, nous allons vérifier que l'événement* $A =$ *«* $\frac{1}{n}\sum_{i=1}^n a_i$ *converge vers* $\frac{1}{2}$ *» appartient bien à* $\mathcal{F}$*. Soit* $N \in \mathbb{N}^*$ *et* $\epsilon > 0$*; l'événement*

$$A_{N,\epsilon} = \text{« } \Big|\frac{1}{n}\sum_{i=1}^n a_i - \tfrac{1}{2}\Big| \leq \epsilon \text{ pour tout } n \geq N \text{ »}$$

15. Pour un mathématicien du moins. D'un point de vue pratique, cela est moins clair. Toutefois, le fait d'autoriser ce type d'événements enrichit substantiellement la théorie mathématique. De plus, il y a une raison importante de s'intéresser à des événements « asymptotiques » : ce n'est que pour ceux-ci que la théorie des probabilités est falsifiable ! En effet, l'affirmation « la probabilité que, lors du prochain lancer, cette pièce tombe sur « pile » est égale à 1/2 » n'est pas falsifiable. Les seules affirmations falsifiables sont celles correspondant à des événements dont la probabilité est 0 ou 1 (ou éventuellement très proche de 0 ou 1). Par exemple, affirmer que si on lance une pièce 1000000 fois, le nombre de « pile » sera compris entre 497500 et 502500 peut être considéré comme falsifiable, car la théorie prédit que la probabilité que cet événement n'ait pas lieu est négligeable en pratique (de l'ordre de 6 $cdot10^{-7}$).

peut s'écrire

$$A_{N,\epsilon} = \bigcap_{n\geq N} \Big\{\omega \in \Omega \,:\, |\frac{1}{n}\sum_{i=1}^{n} a_i - \tfrac{1}{2}| \leq \epsilon\Big\},$$

et par conséquent $A_{N,\epsilon} \in \mathcal{F}$, pour tout $N \in \mathbb{N}^*$ et $\epsilon > 0$, puisqu'il s'écrit comme une intersection d'événements dans $\mathcal{A}$. Ceci implique que l'événement

$$A_\epsilon = \text{« } |\frac{1}{n}\sum_{i=1}^{n} a_i - \tfrac{1}{2}| \leq \epsilon \text{ pour tout } n \text{ suffisamment grand »},$$

qui peut s'écrire

$$A_\epsilon = \bigcup_{N\geq 1} A_{N,\epsilon}$$

appartient aussi à $\mathcal{F}$, pour tout $\epsilon > 0$ (c'est une union dénombrable d'éléments de $\mathcal{F}$). Or l'événement A qui nous intéresse peut s'écrire quant à lui

$$A = \bigcap_{M\geq 1} A_{1/M},$$

et appartient donc bien à $\mathcal{F}$.

La construction décrite dans ce dernier exemple, dans laquelle on part d'une algèbre facile à décrire, que l'on complète ensuite en une tribu, est très courant. L'observation essentielle (simple) est la suivante.

Lemme 1.1.1. *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 1.1.4. *Soit $\mathcal{C} \subseteq \mathscr{P}(\Omega)$. On appelle* tribu engendrée par $\mathcal{C}$ *, notée $\sigma(\mathcal{C})$, la plus petite tribu contenant $\mathcal{C}$,*

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

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

Définition 1.1.5. *Soit $\Omega = \mathbb{R}$. La* tribu borélienne *est la tribu $\mathcal{B}$ sur Ω engendrée par la classe des ouverts. Une partie de $\mathbb{R}$ appartenant à $\mathcal{B}$ est appelée un* borélien.

On peut vérifier assez facilement que $\mathcal{B}$ coïncide avec la tribu engendrée par les intervalles de la forme $(-\infty, a]$, avec $a \in \mathbb{Q}$.

1.1.3 Mesure de probabilité

Étant en possession d'une tribu 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. 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.

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). Notre but est d'obtenir un cadre mathématique permettant de décrire des phénomènes aléatoires, mais déterminer les paramètres permettant d'optimiser l'adéquation entre notre modèle et des expériences réelles 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, 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 15% de chance qu'il y ait un tremblement de terre d'au moins 7 sur l'échelle de Richter en Californie dans les 20 années à venir ». 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 $f_N(A)$ 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 $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Ces conditions sont tout à fait naturelles, et suffisent presque à construire la théorie des probabilités : pour la même raison qu'il est utile de passer de la structure d'algèbre à celle de tribu, il est utile de remplacer la condition d'additivité de $\mathbb{P}$ (3. ci-dessus) par la propriété plus forte de σ-additivité,

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

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

Exemple 1.1.6. *On jette deux dés non pipés. Il est alors naturel de prendre* $\Omega = \{(n,m) \in \{1,2,3,4,5,6\}^2\}$ *et* $\mathcal{F} = \mathscr{P}(\Omega)$. *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. 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* $\mathbb{P}(\{(6,4),(5,5),(4,6)\}) = 3/36 = 1/12$.

1.2 Résumé du chapitre

L'objet de base de la théorie des probabilités, l'espace probabilisé, est un triplet $(\Omega,\mathcal{F},\mathbb{P})$ composé d'un univers Ω arbitraire, d'une tribu $\mathcal{F}$ sur Ω, et d'une application $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ satisfaisant les conditions 1., 2. et 3'. ci-dessus.

Chapitre 2

Probabilité, probabilité conditionnelle et indépendance

2.1 Axiomatique de la théorie des probabilités

Comme discuté dans l'introduction, la structure mathématique de base de la théorie des probabilités est un espace probabilisé [1], c'est-à-dire un triplet $(\Omega,\mathcal{F},\mathbb{P})$, où l'univers Ω est un ensemble quelconque, l'ensemble des événements $\mathcal{F}$ est une tribu sur Ω, et $\mathbb{P}$ est une probabilité sur $\mathcal{F}$, comme définie ci-dessous.

Définition 2.1.1. *Une mesure de probabilité, ou plus simplement une probabilité, sur $\mathcal{F}$ est une application $\mathbb{P} : \mathcal{F} \to [0,1]$ possédant les deux propriétés suivantes :*

1. $\mathbb{P}(\Omega) = 1$.
2. *(σ-additivité) Pour toute famille $A_1,A_2,\ldots \in \mathcal{F}$ d'événements deux-à-deux disjoints,*

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

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

Lemme 2.1.1.

1. $\mathbb{P}(\varnothing) = 0$.
2. *Pour tout $A \in \mathcal{F}$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.*
3. *(Additivité) Pour tout $A,B \in \mathcal{F}$ tels que $A \cap B = \varnothing$,*

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

4. *Pour tout $A \subseteq B \in \mathcal{F}$,*

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A).$$

1. La paire $(\Omega,\mathcal{F})$ seule forme un espace probabilisable.

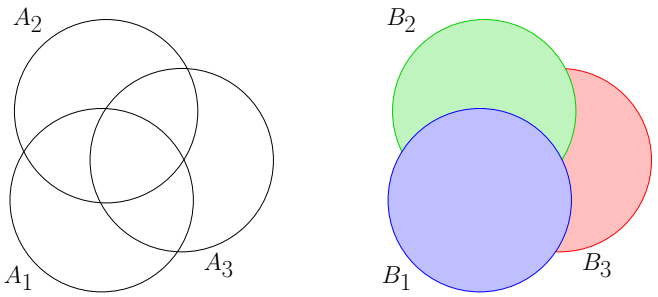


FIGURE 2.1: Trois ensembles A_1,A_2,A_3 (délimités par des cercles) à gauche, et les ensembles B_1,B_2,B_3 correspondant à droite (représentés par les régions coloriées en bleu, vert et rouge, respectivement.)

5. *Pour tout* A,$B \in \mathcal{F}$,

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

6. *Plus généralement,* $\forall A_1, A_2, \ldots, A_n \in \mathcal{F}$,

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

7. *(Sous-*σ*-additivité) Pour toute collection* $A_1, A_2, \ldots \in \mathcal{F}$,

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

Démonstration. 1. Suit de la propriété de σ-additivité avec $A_k = \varnothing$, pour tout $k \ge 1$.

2. Puisque $A \cup A^{\mathrm{c}} = \Omega$ et $A \cap A^{\mathrm{c}} = \varnothing$, cela suit du point suivant.
3. Suit de la propriété de σ-additivité avec $A_1 = A$, $A_2 = B$, et $A_k = \varnothing$, $k \ge 3$.
4. Suit de l'additivité, puisque $B = A \cup (B \setminus A)$ et $A \cap (B \setminus A) = \varnothing$.
5. Puisque $A \cup B = A \cup (B \setminus A)$, et $A \cap (B \setminus A) = \varnothing$, on a, par additivité,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(A) + \mathbb{P}(B \setminus (A \cap B)) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),$$

puisque $\mathbb{P}(B) = \mathbb{P}(B \setminus (A \cap B)) + \mathbb{P}(A \cap B)$.
6. La démonstration, par récurrence, est laissée en exercice.
7. Il suffit d'observer que les événements $B_1 = A_1$ et, pour $k \ge 1$, $B_{k+1} = A_{k+1} \setminus \bigcup_{i=1}^{k} A_i$ sont deux-à-deux disjoints et satisfont $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$ (*cf.* Fig. 2.1).

□

Nous allons à présent énoncer une propriété plus abstraite, qui nous sera utile à plusieurs reprises dans le cours.

Lemme 2.1.2. *Soit $(A_i)_{i\geq 1}$ une suite croissante d'événements, c'est-à-dire telle que $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$, et soit A leur limite,*

$$A = \bigcup_{i=1}^{\infty} A_i \equiv \lim_{i\to\infty} A_i.$$

Alors

$$\mathbb{P}(A) = \lim_{i\to\infty} \mathbb{P}(A_i).$$

Soit $(B_i)_{i\geq 1}$ une suite décroissante d'événements, c'est-à-dire telle que $B_1 \supseteq B_2 \supseteq B_3 \supseteq \cdots$, et soit B leur limite,

$$B = \bigcap_{i=1}^{\infty} B_i \equiv \lim_{i\to\infty} B_i.$$

Alors

$$\mathbb{P}(B) = \lim_{i\to\infty} \mathbb{P}(B_i).$$

Démonstration. $A = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \cdots$ est l'union d'une famille d'événements deux-à-deux disjoints. Par conséquent,

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

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

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

□

Notation. Nous emploierons très fréquemment dans la suite la notation suivante : si A,B sont deux événements, alors on pose

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

2.2 Construction d'espaces probabilisés

Il convient à présent de montrer qu'il est possible de construire de tels espaces probabilisés assez riches pour pouvoir décrire les phénomènes aléatoires. Nous le ferons pour des univers de plus en plus généraux.

2.2.1 Univers fini

Commençons par la situation la plus simple, dans laquelle l'univers Ω est fini. Dans ce cas, la construction d'un espace probabilisé est particulièrement élémentaire. La tribu des événements est simplement $\mathcal{F} = \mathscr{P}(\Omega)$. On se donne une fonction $f : \Omega \to [0,1]$ telle que

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

On associe tout d'abord à chaque événement élémentaire $\omega \in \Omega$ la probabilité $\mathbb{P}(\{\omega\}) = f(\omega)$. On étend ensuite $\mathbb{P}$ à $\mathcal{F}$ par additivité :

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

Lemme 2.2.1. *L'application $\mathbb{P} : \mathcal{F} \to [0,1]$ construite ci-dessus est une mesure de probabilité sur $\mathcal{P}(\Omega)$.*

Démonstration. Il est clair que $\mathbb{P}(\Omega) = \sum_{\omega\in\Omega} f(\omega) = 1$. La seule chose à vérifier est donc la condition d'additivité. Soient $A,B \in \mathcal{F}$, avec $A \cap B = \varnothing$. On a

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

□

Remarque 2.2.1. *Observez également que toute mesure de probabilité sur $\mathcal{P}(\Omega)$ avec Ω fini est de cette forme : étant donné $\mathbb{P}$, il suffit de poser $f(\omega) = \mathbb{P}(\{\omega\})$. L'additivité de $\mathbb{P}$ implique bien que la fonction f satisfait $\sum_{\omega\in\Omega} f(\omega) = 1$, et $\mathbb{P}(A) = \sum_{\omega\in A} f(\omega)$.*

On voit donc qu'une mesure de probabilité sur un univers fini est entièrement caractérisée par les probabilités associées aux événements élémentaires.

Exemple 2.2.1.
- *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 $\omega \in \Omega = \{\mathsf{P},\mathsf{F}\}^5$.*

Un cas particulièrement important est celui où la même probabilité est associée à chaque événement élémentaire, comme dans le premier et le troisième exemples ci-dessus.

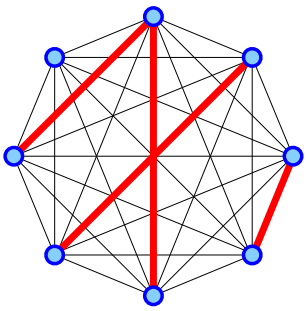


FIGURE 2.2: Une réalisation du graphe aléatoire $\mathcal{G}(8,4)$ (les arêtes présentes sont indiquées en rouge).

Définition 2.2.1. *On appelle* distribution de probabilité uniforme *sur un univers Ω fini, la mesure de probabilité définie par $f(\omega) = 1/|\Omega|$, 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 $\mathbb{P}(A) = |A|/|\Omega|$.

Exemple 2.2.2. *Nous allons à présent introduire un exemple non-trivial d'espace probabilisé fini : le graphe aléatoire d'Erdős–Rényi*[2] [3]*. Soient $m \geq 0$ et $n \geq 1$ deux entiers. Le* graphe aléatoire $\mathcal{G}(n,m)$ *est l'espace probabilisé sur l'ensemble des graphes $G = (S,A)$ à n sommets et m arêtes : $S = \{1,\ldots,n\}$, et $A \subset \big\{\{i,j\} : 1 \leq i < j \leq n\big\}$ avec $|A| = m$. La mesure de probabilité sur cet ensemble est la mesure uniforme.*

Quelques résultats combinatoires

Nous allons à présent rappeler certains résultats de combinatoire élémentaires qui sont régulièrement utilisés. 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 $(n)_0 = 1$.

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. Deux procédures sont possibles : le tirage avec remise, durant lequel chaque élément de l'ensemble peut être choisi à plusieurs reprises, et 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 $r \leq n$).

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.

Lemme 2.2.2. *On considère un ensemble à $n \geq 1$ éléments, et $r \in \mathbb{N}$.*

1. *Le nombre d'échantillons de taille r différents avec remise est égal à n^r.*
2. *Pour $r \leq n$, le nombre d'échantillons de taille r différents sans remise 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 pour le premier élément ne peut plus 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. □

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 distribution de probabilité. Sauf mention explicite du contraire, on considérera la distribution 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 théorème 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 2.2.3. *Supposons que dans une ville donnée il y a 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 2.2.4. *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 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*[5]*, par exemple), etc. Pour le modèle 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.

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

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 donné une population de taille n, nous cherchons à déterminer le nombre de sous-populations de taille r.

Lemme 2.2.3. *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 sans remise de taille r 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 2.2.5. *Au poker, chaque joueur reçoit 5 cartes parmi 52. Le nombre de mains possibles est donc de $\binom{52}{5} = 2598960$. 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 2.2.6. *Considérons la distribution aléatoire de r balles dans n urnes. Quelle est la probabilité qu'une urne donnée contienne exactement k balles ? On peut choisir les k balles de $\binom{r}{k}$ façons. Les autres $r-k$ balles 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*

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

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

Exemple 2.2.7. *Retournons au graphe aléatoire de l'Exemple 2.2.2. On a clairement*

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

Par conséquent, le nombre total de graphes dans $\mathcal{G}(n,m)$ est donné par $\binom{N}{m}$, et donc la probabilité de chaque graphe est donnée par

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

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

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

Lemme 2.2.4. *Soit $r_1, \ldots, r_k$ 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 $n - r_1$, 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 2.2.8. *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*

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

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 du Théorème du binôme de Newton.

Lemme 2.2.5. *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 assurent la convergence de la série. □

Formule de Stirling L'équivalence asymptotique suivante pour $n!$, dûe à Stirling[6], est très utile dans de nombreux problèmes de nature combinatoire.

Lemme 2.2.6. *Lorsque $n \to \infty$, on a*

$$n! = n^n e^{-n} \sqrt{2\pi n}\,(1 + o(1)).$$

Démonstration. Sera faite en exercice. □

2.2.2 Univers dénombrable

On peut procéder à la construction d'espaces probabilisés avec un univers Ω dénombrable exactement de la même façon que dans le cas fini : on prend $\mathcal{F} = \mathscr{P}(\Omega)$, et on associe à chaque événement élémentaire $\omega \in \Omega$ sa probabilité, $\mathbb{P}(\{\omega\}) \equiv f(\omega) \in [0,1]$, avec

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

Remarque 2.2.2. *La somme ci-dessus est définie de la manière suivante. Ω étant dénombrable, il est possible de numéroter ses éléments, disons $\Omega = \{\omega_1,\omega_2,\ldots\}$. On pose alors, pour tout $A \subseteq \Omega$,*

$$\sum_{\omega\in A} f(\omega) = \sum_{i=1}^{\infty} f(\omega_i)\,\mathbf{1}_A(\omega_i).$$

Il est important d'observer que cette définition ne dépend pas *de l'ordre choisi pour les éléments de Ω : toutes les séries intervenant sont à termes positifs, et ceux-ci peuvent donc être réorganisés à notre guise.*

On pose ensuite, pour $A \in \mathcal{F}$, $\mathbb{P}(A) = \sum_{\omega\in A} f(\omega)$. On vérifie alors de la même façon que dans le cas fini que $\mathbb{P}$ est bien une mesure de probabilité et que toute mesure de probabilité sur un univers dénombrable est nécessairement de cette forme.

Exemple 2.2.9. *On jette une pièce de monnaie jusqu'à l'obtention du premier pile. On peut choisir $\Omega = \mathbb{N}^* \cup \{\infty\}$ où le dernier événement représente la possibilité que pile ne sorte jamais. Si la pièce est équilibrée, on aura*

$$f(k) = 2^{-k}, \qquad k = 1,2,\ldots$$

En particulier, la probabilité que pile ne sorte jamais est donnée par

$$f(\infty) = 1 - \sum_{k=1}^{\infty} 2^{-k} = 0,$$

comme le veut l'intuition.

En particulier, la probabilité que le premier pile sorte après un nombre pair de lancers est de

$$\mathbb{P}(\{2,4,6,\ldots\}) = \sum_{k=1}^{\infty} f(2k) = \sum_{k=1}^{\infty} 2^{-2k} = 1/3.$$

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

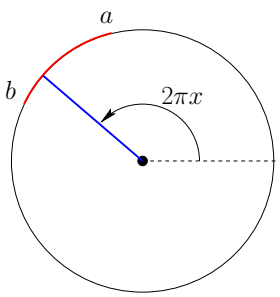


FIGURE 2.3: L'aiguille de l'Exemple 2.2.10. La position de l'aiguille (en bleu) est représentée par le nombre $x \in [0,1)$. La direction de l'aiguille tombe dans un intervalle $[a,b)$ quelconque avec probabilité $b-a$.

2.2.3 Univers non-dénombrable

Nous allons à présent brièvement discuter le cas d'espaces probabilisés construits à partir d'un univers Ω infini non dénombrable. Cette situation est substantiellement plus subtile que les cas étudiés précédemment. Commençons par considérer un exemple.

Exemple 2.2.10. *On désire modéliser l'expérience suivante : on considère une aiguille dont une extrémité est fixée à un axe autour duquel elle peut tourner (cf. Fig. 2.3). On peut encoder la position de l'aiguille par l'angle qu'elle fait avec une direction fixée. On peut donc prendre, dans des unités appropriées,* $\Omega = [0,1)$. *On suppose qu'une fois lancée, l'aiguille peut s'arrêter en pointant dans n'importe quelle direction, avec la même probabilité. Plus précisément, on va demander à ce que la probabilité que l'aiguille s'arrête dans un intervalle* $[a,b)$ *(*$a \neq b \in [0,1]$*) le long du cercle, ne dépende que de sa longueur,* $\mathbb{P}([a,b)) = b-a$.

Manifestement, on ne peut plus construire une telle probabilité comme précédemment, en spécifiant les probabilités des événements élémentaires, puis en définissant les probabilités d'événements généraux à partir de celles-ci. En effet, la probabilité de n'importe quel événement élémentaire doit être nulle : si $x \in [0,1)$, $\mathbb{P}(\{x\}) \leq \mathbb{P}([x,x+\epsilon)) = \epsilon$, *pour tout* $\epsilon > 0$. *Les seuls événements dont il est possible d'évaluer la probabilité à partir de celles des événements élémentaires sont les unions dénombrables de points (et leurs compléments), et les probabilités de celles-ci sont toutes nulles (ou égales à* 1*).*

La question est de savoir s'il est possible de construire une tribu sur $[0,1)$, *contenant tous les intervalles, sur laquelle on puisse définir une mesure de probabilité* $\mathbb{P}$ *associant à chaque intervalle sa longueur. La réponse est positive, mais la construction n'est pas triviale. Elle sera faite en détail dans le cours de théorie de la mesure (Analyse III). La tribu correspondante est celle des boréliens de* $[0,1)$ *(la tribu engendrée par les ouverts de* $[0,1)$*). Elle ne contient* pas *toutes les parties de* $[0,1)$ *: il n'est pas possible d'attribuer une « longueur » (on dit* mesurer*) à tous les sous ensembles de* $[0,1)$ *de façon cohérente.*

Pour être un peu plus précis, l'affirmation « toute partie de $[0,1)$ *est mesurable » est*

indépendante du système d'axiomes de Zermelo-Fraenkel[7][8] *: il n'est pas possible de la prouver, ni de prouver son contraire. En fait, si l'on accepte l'axiome du choix (non dénombrable), alors il est possible de montrer l'existence de sous-ensembles de* $[0,1)$ *qui ne sont pas mesurables*[9]*. Ceci dit, même dans ces conditions, l'existence de tels ensembles auxquels on ne peut associer de probabilité ne limite en rien l'applicabilité de la théorie des probabilités, puisque ces ensembles sont pathologiques (il est impossible de les décrire explicitement, puisque leur existence repose de façon essentielle sur l'axiome du choix), et ne correspondent donc pas à des événements intéressants dans la pratique.*

Comme expliqué dans l'exemple précédent, il est nécessaire en général de restreindre la classe des événements, afin de pouvoir construire un espace probabilisé. La procédure est la suivante :

1. On commence par déterminer une algèbre d'événements intéressants, sur laquelle on définit une probabilité. *Dans l'exemple, on part des intervalles, dont on connaît la probabilité. On considère ensuite l'algèbre engendrée par les intervalles. On définit sur cette algèbre une mesure de probabilité finiment additive, la probabilité de chaque élément étant déterminée à partir de celle des intervalles et des règles d'additivité. On montre ensuite que cette mesure est en fait* σ*-additive.*
2. On fait appel à un résultat fondamental de théorie de la mesure, le Théorème d'extension de Carathéodory, qui affirme qu'une mesure de probabilité sur une algèbre s'étend de façon unique en une mesure de probabilité sur la tribu engendrée par l'algèbre.

Exemple 2.2.11. *Revenons à un problème déjà discuté dans les Exemples 1.1.4 et 1.1.5 : une infinité de jets d'une pièce de monnaie. On a vu que les ensembles*

$$\{\omega \in \Omega \,:\, (a_1, \ldots, a_n) \in A\},$$

avec $n \geq 1$ *un entier arbitraire et* $A \subseteq \{0,1\}^n$*, forment une algèbre sur* Ω*. Or, chaque élément de cette algèbre ne fait intervenir qu'un nombre fini de lancers, et par conséquent, on peut aisément leur associer une probabilité (nous reviendrons sur la façon de le faire une*

7. Ernst Friedrich Ferdinand Zermelo (1871, Berlin - 1953, Fribourg-en-Brisgau), mathématicien allemand.

8. Abraham Adolf Halevi Fraenkel (1891, Münich - 1965, Jérusalem), mathématicien d'abord allemand puis israélien.

9. Esquissons brièvement une construction due à Vitali. On note $\mathbb{S}^1$ le cercle unité. Nous allons montrer, en utilisant l'axiome du choix, qu'il est possible d'écrire $\mathbb{S}^1 = \bigcup_{n\in\mathbb{Z}} A_n$, où les ensembles A_n sont disjoints et peuvent tous être obtenus à partir de A_0 par rotation. Si A_0 possèdait une longueur $\ell(A)$, alors la σ-additivité impliquerait que $2\pi = \infty \times \ell(A)$, ce qui est impossible. Pour construire A_n, on procède comme suit. On identifie $\mathbb{S}^1$ à l'ensemble $\{e^{\mathrm{i}\theta} \,:\, \theta \in \mathbb{R}\}$ dans $\mathbb{C}$. On introduit une relation d'équivalence sur $\mathbb{S}^1$ en posant $x \sim y$ s'il existe $\alpha,\beta \in \mathbb{R}$ tels que $x = e^{\mathrm{i}\alpha}, y = e^{\mathrm{i}\beta}$, avec $\alpha - \beta \in \mathbb{Z}$. On utilise l'axiome du choix pour construire l'ensemble A_0 composé d'exactement un représentant de chaque classe d'équivalence. On pose alors, pour $n \in \mathbb{Z}^*$, $A_n = e^{\mathrm{i}n} A_0 = \{e^{\mathrm{i}n} x \,:\, x \in A_0\}$. La famille ainsi construite possède les propriétés désirées. En effet, si $y \in A_n$ alors il existe $x \in A_0$ tel que $y = e^{\mathrm{i}n} x$, et donc $y \sim x$; comme A_0 ne contient qu'un seul représentant de chaque classe d'équivalence, on en déduit que $y \notin A_0$. Ceci montre que les ensembles A_n sont disjoints. De plus, si $y \in \mathbb{S}^1$, sa classe d'équivalence est donnée par $\{e^{\mathrm{i}k} y \,:\, k \in \mathbb{Z}\}$, et il existe donc $n \in \mathbb{Z}$ tel que $e^{-\mathrm{i}n} y \in A_0$, puisque A_0 contient un représentant de chaque classe d'équivalence ; on en déduit que $y \in A_n$, et donc que les A_n forment une partition de $\mathbb{S}^1$.

fois le concept d'indépendance introduit), et vérifier que celle-ci est σ-additive. On obtient alors notre espace probabilisé, sur la tribu engendrée par cette algèbre, par une application du Théorème d'extension de Carathéodory.

Le cas de $\mathbb{R}$

Le cas de $\mathbb{R}$ est particulièrement important. Donnons donc brièvement quelques définitions et résultats dans ce contexte. Ceux-ci seront étudiés de façon détaillée dans le cours de théorie de la mesure (Analyse III).

Définition 2.2.2. *La tribu borélienne sur $\Omega \subseteq \mathbb{R}$, $\mathcal{B}(\Omega)$, est la tribu sur Ω engendrée par les ouverts de Ω. Ses éléments sont appelés les boréliens.*

Dans la suite, lorsque nous considérerons $\mathbb{R}$ comme espace probabilisé, nous le supposerons toujours muni de sa tribu borélienne, sauf mention du contraire.

Lemme 2.2.7. *La tribu borélienne est engendrée par les intervalles $(-\infty,a]$, $a \in \mathbb{Q}$.*

Une mesure de probabilité $\mathbb{P}$ sur $\Omega \subseteq \mathbb{R}$ peut être caractérisée par les valeurs qu'elle attribue aux intervalles de cette forme. Ceci motive l'introduction d'une fonction $F_{\mathbb{P}} : \mathbb{R} \to [0,1]$, $F_{\mathbb{P}}(x) = \mathbb{P}((-\infty,x])$.

Définition 2.2.3. *Une fonction de répartition est une fonction $F : \mathbb{R} \to [0,1]$ possédant les propriétés suivantes :*

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

Lemme 2.2.8. *$F_{\mathbb{P}}$ est une fonction de répartition.*

Démonstration. Laissée en exercice. Pour la continuité à droite, utiliser le Lemme 2.1.2. □

On peut donc associer à chaque mesure de probabilité une fonction de répartition. Le résultat suivant montre que la réciproque est également vraie.

Théorème 2.2.1. *Soit $F : \mathbb{R} \to \mathbb{R}$. Alors il existe une mesure de probabilité $\mathbb{P}$ sur $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ telle que $F = F_{\mathbb{P}}$ si et seulement si F est une fonction de répartition.*

Ce résultat montre que les mesures de probabilité sur $\mathbb{R}$ sont en bijection avec les fonctions de répartition sur $\mathbb{R}$.

2.3 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 pleut 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. La 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

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

Lorsque N devient grand, on s'attend à ce que le terme de gauche converge vers la probabilité de A conditionnellement à la réalisation de l'événement B, alors que le terme de droite devrait converger vers $\mathbb{P}(A\cap B)/\mathbb{P}(B)$. Ceci motive la définition suivante.

Définition 2.3.1. *Soit $B \in \mathcal{F}$ un événement tel que $\mathbb{P}(B) > 0$. Pour tout $A \in \mathcal{F}$, la probabilité conditionnelle de A sachant B est la quantité*

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

Lemme 2.3.1. *Soit $B \in \mathcal{F}$ un événement tel que $\mathbb{P}(B) > 0$. Alors la probabilité conditionnelle $\mathbb{P}(\,\cdot\,|\,B) : \mathcal{F} \to \mathbb{R}$ est une mesure de probabilité, et $(\Omega,\mathcal{F},\mathbb{P}(\,\cdot\,|\,B))$ est un espace probabilisé. De plus, $\mathcal{F}_B = \{A\cap B \,:\, A\in\mathcal{F}\}$ est une tribu et $(B,\mathcal{F}_B,\mathbb{P}(\,\cdot\,|\,B))$ est également un espace probabilisé.*

Démonstration. On a manifestement $\mathbb{P}(A\cap B)/\mathbb{P}(B) \in [0,1]$, pour tout $A\in\mathcal{F}$. Comme $\Omega\cap B = B$, on a également $\mathbb{P}(\Omega\,|\,B) = 1$. Finalement, si $A_1,A_2,\ldots$ sont des événements deux-à-deux disjoints, la σ-additivité de $\mathbb{P}$ implique que

$$\mathbb{P}\Big((\bigcup_{i=1}^{\infty} A_i)\cap B\Big) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty}(A_i\cap B)\Big) = \sum_{i=1}^{\infty}\mathbb{P}(A_i\cap B),$$

et donc que

$$\mathbb{P}\Big((\bigcup_{i=1}^{\infty} A_i)\;\Big|\; B\Big) = \sum_{i=1}^{\infty}\frac{\mathbb{P}(A_i\cap B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty}\mathbb{P}(A_i\,|\,B).$$

La preuve de la seconde affirmation est laissée en exercice. □

Exemple 2.3.1. *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 2.3.2. *On choisit une famille au hasard parmi toutes les familles ayant deux enfants et dont au moins un est un garçon. Quelle est la probabilité que les deux enfants soient des garçons ? Introduisant les événements* $B = \{(\mathrm{G},\mathrm{G}),(\mathrm{F},\mathrm{G}),(\mathrm{G},\mathrm{F})\}$ *et* $A = A \cap B = \{(\mathrm{G},\mathrm{G})\}$*, on voit que*

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

On choisit une famille au hasard parmi toutes les familles ayant deux enfants et dont l'aîné est un garçon. Quelle est la probabilité que les deux enfants soient des garçons ? À présent, $B = \{(\mathrm{G},\mathrm{G}),(\mathrm{G},\mathrm{F})\}$*,* $A = A \cap B = \{(\mathrm{G},\mathrm{G})\}$*. Donc*

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

Définition 2.3.2. *Une famille* $(B_i)_{i\in I}$*,* I *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 2.3.1. *Soit* $(B_i)_{i\in I}$ *une partition de* Ω *telle que* $\mathbb{P}(B_i) > 0$*, pour tout* $i \in I$*, et soit* $A \in \mathcal{F}$*.*

1. *(Loi de la probabilité totale)*

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

2. *(Formule de Bayes)*

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

Démonstration. Par σ-additivité,

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

La seconde relation suit de l'observation que

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

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

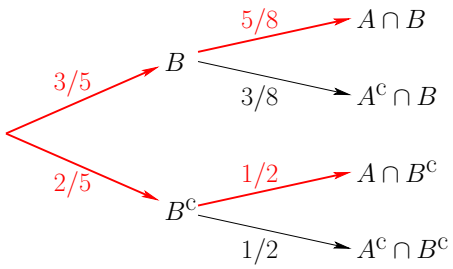


FIGURE 2.4: L'arbre représentant le processus décrit dans l'Exemple 2.3.3

Remarque 2.3.1. *Dans la terminologie statistique, on appelle $\mathbb{P}(B_i)$ la probabilité à priori de B_i et $\mathbb{P}(B_i \,|\, A)$ la probabilité à posteriori de B_i (sachant A). La formule de Bayes donne donc un moyen de transformer les probabilités à priori en probabilités à posteriori.*

Exemple 2.3.3. *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

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

À présent,

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

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 de ce type comme sur la Fig. 2.4.

Exemple 2.3.4 (Problème du ballot)**.** *Lors d'une élection opposant deux candidats A et B, le premier reçoit n voix et le second $m < n$ voix. En supposant équiprobables les différents ordres d'apparition des bulletins (et en ignorant les bulletins blancs ou non-valides), montrer que la probabilité $P(n,m)$ que le candidat A soit toujours en tête lors du dépouillement est égale à $(n-m)/(n+m)$.*

En conditionnant sur le résultat du dernier bulletin, il suit de la loi de la probabilité totale et de l'hypothèse d'équiprobabilité que

$$P(n,m) = \mathbb{P}(A \text{ toujours en tête} \,|\, \text{dernier vote en faveur de } A)\, \frac{n}{n+m} + \mathbb{P}(A \text{ toujours en tête} \,|\, \text{dernier vote en faveur de } B)\, \frac{m}{m+n}.$$

Un instant de réflexion montre que $\mathbb{P}(A$ *toujours en tête*$\,|\,$*dernier vote en faveur de* $A) = P(n{-}1,m)$ *et* $\mathbb{P}(A$ *toujours en tête*$\,|\,$*dernier vote en faveur de* $B) = P(n,m{-}1)$*. Par conséquent, le problème se réduit à vérifier que* $P(n,m) = (n-m)/(n+m)$ *est bien la solution du système*

$$P(n,m) = \frac{n}{n+m}P(n-1,m) + \frac{m}{m+n}P(n,m-1), \qquad n > m \geq 1,$$

avec les conditions au bord $P(n,n) = 0$ *(*A *ne peut avoir été toujours en tête s'il est à égalité avec* B *à la fin) et* $P(n,0) = 1$ *(*A *a forcément toujours été en tête si personne n'a voté pour* B*). Les conditions au bord sont clairement vérifiées. Pour démontrer le résultat, on procède par récurrence sur* $n+m$*. Supposons le résultat valide pour* $n+m \leq k$ *(*$n \geq m$*,* $k \geq 1$*), ainsi que pour* $n = m$ *arbitraires. Considérons à présent* $n+m = k+1$*,* $n > m$*. On a alors, par hypothèse de récurrence,*

$$P(n,m) = \frac{n}{n+m}\frac{n-1-m}{n-1+m} + \frac{m}{m+n}\frac{n-(m-1)}{n+(m-1)} = \frac{n-m}{n+m},$$

et le résultat est établi.

Exemple 2.3.5. *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 :

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

On a donc les informations suivantes :

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

et on veut calculer

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

La formule de Bayes nous dit que

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

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

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

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

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

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 $2/100000$.

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 2.3.6. *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* 1000000 *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 2.3.7. *(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 donc qu'entre lui-même et* C*, l'un va être libéré, et l'autre exécuté. Quelle est la probabilité que* A *soit exécuté ?*

Remarque 2.3.2. *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 2.3.8. *(Paradoxe de Simpson*[10]*) 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 avoir annoncé ce résultat, un de ses assistants vient vers lui. Il 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 : 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. 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

10. 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.

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 candidater dans des départements avec forts taux d'admission.

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

Définition 2.4.1. *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.

Proposition 2.4.1. *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 $B_i \in \{A_i, A_i^{\mathrm{c}}\}$, sont aussi indépendants.

Démonstration. Laissée en exercice. □

Remarque 2.4.1. *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, *ou* indépendante par paires. *L'indépendance par paires* n'implique pas l'indépendance. *Un exemple : considérez $\Omega = \{1,2,3,4\}$, avec la distribution uniforme, et les événements $A = \{1,2\}$, $B = \{2,3\}$ et $C = \{1,3\}$; on vérifie aisément que A,B,C sont indépendants par paires, et pourtant $\mathbb{P}(A \cap B \cap C) = 0 \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$.*

Exemple 2.4.1. *Retournons au graphe aléatoire $\mathcal{G}(n,m)$. La probabilité que deux sommets distincts i et j donnés soient reliés par une arête (noté $i \sim j$) 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 façon 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 montre les deux exemples suivants.

Exemple 2.4.2. *Un événement peut être indépendant de lui-même. En effet, ceci a lieu si et seulement s'il a probabilité 0 ou 1, car dans ce cas, on a bien*

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

Exemple 2.4.3. *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 2.4.2. *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*

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

pour tous les sous-ensembles finis J de I.

2.5 Expériences répétées, espace produit

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

L'espace probabilisé correspondant à une instance de l'expérience est noté $(\Omega_1,\mathcal{F}_1,\mathbb{P}_1)$.

Exemple 2.5.1. *Dans le cas d'un jet d'une pièce de monnaie,* $\Omega_1 = \{\mathsf{P},\mathsf{F}\}$*, et la tribu correspondante est* $\mathcal{F} = \mathscr{P}(\Omega_1)$*.* Ω_1 *étant fini, il est suffisant, pour déterminer la mesure de probabilité, de donner ses valeurs sur les événements élémentaires : on posera donc* $\mathbb{P}_1(\{\mathsf{P}\}) = p$*,* $\mathbb{P}_1(\{\mathsf{F}\}) = 1-p \equiv q$*, où* $p \in [0,1]$ *est la probabilité que la pièce tombe sur pile.* $p = \frac{1}{2}$ *dans le cas d'une pièce équilibrée.*

Nous allons à présent construire l'espace probabilisé correspondant à 2 répétitions de l'expérience. L'univers correspondant est donné par le produit cartésien de 2 copies de Ω_1 : $\Omega_2 = \Omega_1 \times \Omega_1 = \{(a_1,a_2) : a_i \in \Omega_1\}$.

En ce qui concerne la construction de la tribu sur Ω_2, nous voulons garantir que celle-ci contienne tous les événements du type « l'événement A s'est produit lors de la première expérience, et l'événement B s'est produit lors de la seconde ». Ceci conduit à la définition suivante.

Définition 2.5.1. *Si* $\mathcal{F}$ *et* $\mathcal{F}'$ *sont deux tribus sur des univers* Ω *et* Ω'*, la* tribu produit *$\mathcal{F} \times \mathcal{F}'$ sur $\Omega \times \Omega'$ est la tribu engendrée par les* rectangles*, c'est-à-dire les ensembles de la forme* $A \times B$ *avec* $A \in \mathcal{F}$ *et* $B \in \mathcal{F}'$*.*

Exemple 2.5.2. *La* tribu borélienne *sur* $\mathbb{R}^n$ *est la tribu produit* $\mathcal{B}(\mathbb{R}) \times \cdots \times \mathcal{B}(\mathbb{R})$ *(n fois). On la notera* $\mathcal{B}(\mathbb{R}^n)$*. On peut montrer qu'elle coïncide avec la tribu engendrée par les ouverts de* $\mathbb{R}^n$*, et qu'elle est en fait également engendrée par les ensembles de la forme* $(-\infty,x_1] \times \cdots \times (-\infty,x_n]$*, avec* $x_1,\ldots,x_n \in \mathbb{Q}$*.*

Nous désirons à présent définir la mesure de probabilité $\mathbb{P}_2$ sur $(\Omega_2,\mathcal{F}_2)$. Nous voulons modéliser l'indépendance des expériences successives, par conséquent deux événements A et B portant l'un sur la première expérience, et l'autre sur la seconde doivent être indépendants. Cela implique que

$$\mathbb{P}_2(A \cap B) = \mathbb{P}_2(A)\mathbb{P}_2(B),$$

pour tout A de la forme $\tilde{A} \times \Omega_1$, et B de la forme $\Omega_1 \times \tilde{B}$, avec $\tilde{A},\tilde{B} \in \mathcal{F}_1$. De plus la réalisation de l'événement A ne dépendant que de la réalisation de $\tilde{A}$ lors de la première expérience, on doit avoir $\mathbb{P}_2(A) = \mathbb{P}_1(\tilde{A})$; similairement $\mathbb{P}_2(B) = \mathbb{P}_1(\tilde{B})$. Observant que $A \cap B = \tilde{A} \times \tilde{B}$, ceci conduit à chercher à définir $\mathbb{P}_2$ par

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

L'existence d'une telle mesure de probabilité est un résultat classique de théorie de la mesure (*cf.* Analyse III).

Théorème 2.5.1. *Soient* $(\Omega,\mathcal{F},\mathbb{P})$ *et* $(\Omega',\mathcal{F}',\mathbb{P}')$ *deux espaces probabilisés. Il existe une unique mesure de probabilité* $\mathbb{P}\times\mathbb{P}'$ *sur l'espace probabilisable* $(\Omega\times\Omega',\mathcal{F}\times\mathcal{F}')$ *telle que*

$$\mathbb{P}\times\mathbb{P}'(A\times B)=\mathbb{P}(A)\mathbb{P}'(B),\qquad \forall A\in\mathcal{F},B\in\mathcal{F}'.$$

$\mathbb{P}\times\mathbb{P}'$ *est appelé* mesure produit *de* $\mathbb{P}$ *et* $\mathbb{P}'$.

Définition 2.5.2. *Soient* $(\Omega,\mathcal{F},\mathbb{P})$ *et* $(\Omega',\mathcal{F}',\mathbb{P}')$ *deux espaces probabilisés. L'espace probabilisé* $(\Omega\times\Omega',\mathcal{F}\times\mathcal{F}',\mathbb{P}\times\mathbb{P}')$ *est leur* espace probabilisé produit.

Exemple 2.5.3. *Pour deux jets de pièces de monnaie, on obtient*

$$\Omega_2=\{\mathsf{PP},\mathsf{PF},\mathsf{FP},\mathsf{FF}\},\mathcal{F}_2=\mathscr{P}(\Omega_2),$$

et $\mathbb{P}_2$ *est déterminée par* $\mathbb{P}_2(\{\mathsf{PP}\})=p^2$, $\mathbb{P}_2(\{\mathsf{PF}\})=\mathbb{P}_2(\{\mathsf{FP}\})=pq$ *et* $\mathbb{P}_2(\{\mathsf{FF}\})=q^2$.

En itérant la construction ci-dessus, on construit l'espace probabilisé $(\Omega_n,\mathcal{F}_n,\mathbb{P}_n)$ correspondant à la répétition d'un nombre fini quelconque d'expériences indépendantes : $\Omega_n=\Omega_1\times\cdots\times\Omega_1$ (n fois), $\mathcal{F}_n=\mathcal{F}_1\times\cdots\times\mathcal{F}_1$ (n fois) et $\mathbb{P}_n=\mathbb{P}_1\times\cdots\times\mathbb{P}_1$ (n fois).

Pour diverses raisons, en particulier la discussion de la loi forte des grands nombres, il est important de pouvoir discuter de la répétition d'un nombre *infini* d'expériences indépendantes. La façon de procéder est la suivante (déjà esquissée dans l'Exemple 1.1.4). On définit évidemment l'univers comme le produit cartésien d'une infinité de copies de Ω_1, $\Omega_\infty=\Omega_1\times\Omega_1\times\cdots$. La tribu correspondante $\mathcal{F}_\infty$ est la tribu engendrée par les événements ne dépendant que des résultats d'un nombre fini d'expériences, c'est-à-dire les événements de la forme

$$\left\{(a_1,a_2,\ldots)\in\Omega_\infty\,:\,(a_1,\ldots,a_n)\in\tilde{A}\right\},$$

avec $n\geq 1$ un entier arbitraire et $\tilde{A}\in\mathcal{F}_n$. Ces ensembles formant une algèbre (*cf.* l'Exemple 1.1.4), il suffit de construire la mesure de probabilité $\mathbb{P}_\infty$ pour ces ensembles, le Théorème d'extension de Carathéodory permettant de l'étendre automatiquement à la tribu $\mathcal{F}_\infty$. Mais, si A est un tel événement, $A=\tilde{A}\times\Omega_1\times\Omega_1\times\cdots$, $\tilde{A}\in\mathcal{F}_n$, $n\geq 1$, on doit avoir $\mathbb{P}_\infty(A)=\mathbb{P}_n(\tilde{A})$.

En particulier, pour déterminer la probabilité de l'événement

$$A=\{(a_1,a_2,\ldots)\in\Omega_\infty\,:\,a_1\in B_1,\ldots,a_n\in B_n\},$$

où $B_i\in\mathcal{F}_1$ ($i=1,\ldots,n$), il suffit de ne considérer que les n premières expériences, et on doit donc avoir

$$\mathbb{P}_\infty(A)=\mathbb{P}_n(\tilde{A})=\mathbb{P}_1(B_1)\cdots\mathbb{P}_1(B_n),$$

où $\tilde{A}=B_1\times\cdots\times B_n$.

2.6 Résumé du chapitre

Continuité des mesures de probabilité. Si $(A_i)_{i\geq 1}$ est une suite croissante d'événements, $A_1 \subseteq A_2 \subseteq \cdots$, alors leur limite $\lim_{i\to\infty} A_i = \bigcup_{i\geq 1} A_i$ satisfait

$$\mathbb{P}(\lim_{i\to\infty} A_i) = \lim_{i\to\infty} \mathbb{P}(A_i).$$

Un résultat analogue est également vérifié pour une suite décroissante d'événements.

Construction d'espaces probabilisés : cas fini et dénombrable. Dans ce cas il est possible de choisir $\mathcal{F} = \mathscr{P}(\Omega)$, et une mesure de probabilité $\mathbb{P}$ est caractérisée par les valeurs qu'elle associe aux événements élémentaires $\omega \in \Omega$, $\mathbb{P}(\{\omega\}) = f(\omega)$. La probabilité d'un événement A quelconque est alors donnée par $\mathbb{P}(A) = \sum_{\omega\in A} f(\omega)$.

Construction d'espaces probabilisés : cas non dénombrable. Dans ce cas il n'est en général pas possible de prendre $\mathcal{F} = \mathscr{P}(\Omega)$. La construction se fait alors par étapes : choix d'une algèbre naturelle d'événements, dont la probabilité peut être aisément définie ; extension de cette mesure de probabilité sur l'algèbre en une mesure de probabilité sur la tribu qu'elle engendre, à l'aide du Théorème d'extension de Carathéodory.

Probabilité conditionnelle. Étant donné un événement B tel que $\mathbb{P}(B) > 0$, la probabilité conditionnelle sachant B est la mesure de probabilité définie par $\mathbb{P}(A \,|\, B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$.

- Loi de la probabilité totale : $\mathbb{P}(A) = \sum_{i\in I} \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i)$, pour toute partition $(B_i)_{i\in I}$ de Ω ;
- Formule de Bayes : $\mathbb{P}(B_i \,|\, A) = \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i) / \sum_{j\in I} \mathbb{P}(A \,|\, B_j)\mathbb{P}(B_j)$.

Indépendance. Une famille $(A_i)_{i\in I}$ d'événements est indépendante (sous $\mathbb{P}$) si, pour tout $J \subseteq I$ fini, $\mathbb{P}(\bigcap_{i\in J} A_i) = \prod_{i\in J} \mathbb{P}(A_i)$. En particulier, si A et B sont indépendants et $\mathbb{P}(B) > 0$, alors $\mathbb{P}(A \,|\, B) = \mathbb{P}(A)$.

Expériences répétées. Si $(\Omega,\mathcal{F},\mathbb{P})$ est l'espace probabilisé associé à une expérience aléatoire, l'espace probabilisé associé à n répétitions indépendantes de l'expérience est donné par l'espace produit, $(\Omega \times \cdots \times \Omega, \mathcal{F} \times \cdots \times \mathcal{F}, \mathbb{P} \times \cdots \times \mathbb{P})$ (tous les produits étant pris n fois), où $\Omega \times \Omega$ est le produit cartésien des ensembles, $\mathcal{F} \times \mathcal{F}$ est la tribu engendrée par les ensembles de la forme $A \times B$, $A,B \in \mathcal{F}$, et $\mathbb{P} \times \mathbb{P}$ est l'unique de mesure de probabilité sur $\mathcal{F} \times \mathcal{F}$ telle que $\mathbb{P} \times \mathbb{P}(A \times B) = \mathbb{P}(A)\mathbb{P}(B)$.

L'espace probabilisé correspondant à une infinité de répétitions indépendantes de l'expérience est $(\Omega_\infty,\mathcal{F}_\infty,\mathbb{P}_\infty)$, où Ω_∞ est le produit cartésien d'une infinité de copies de Ω, $\mathcal{F}_\infty$ est la tribu engendrée par les événements ne dépendant que des n premières expériences, n arbitraire, et $\mathbb{P}_\infty$ est l'unique mesure de probabilité sur $\mathcal{F}_\infty$ telle que $\mathbb{P}(A_1 \times \cdots \times A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n)$, pour tout n.

Chapitre 3

Variables aléatoires

3.1 Définitions

3.1.1 Variables aléatoires 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 2.2.2 est un objet trop complexe pour être directement intéressant (voir la Fig. 3.1). Il sera alors 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.

Exemple 3.1.1. *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 la figure 2.2, on a $N_0 = 1$, $N_1 = 6$, $N_2 = 1$, et $N_k = 0$ pour les autres valeurs de k.*

Soit $(\Omega,\mathcal{F},\mathbb{P})$ un espace probabilisé. Les questions que l'on va se poser concernant une variable aléatoire $X : \Omega \to \mathbb{R}$ 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 $A \subseteq \mathbb{R}$. Or, $\mathbb{P}(X^{-1}(A))$ n'est bien définie que si $X^{-1}(A) \in \mathcal{F}$. De plus, la distribution 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 $A \subseteq \mathbb{R}$,

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

On a vu que ceci ne peut pas être fait de manière cohérente pour toutes les parties de $\mathbb{R}$, et qu'il faudra donc se restreindre aux ensembles $A \in \mathcal{B}$. On est donc conduit à la définition suivante.

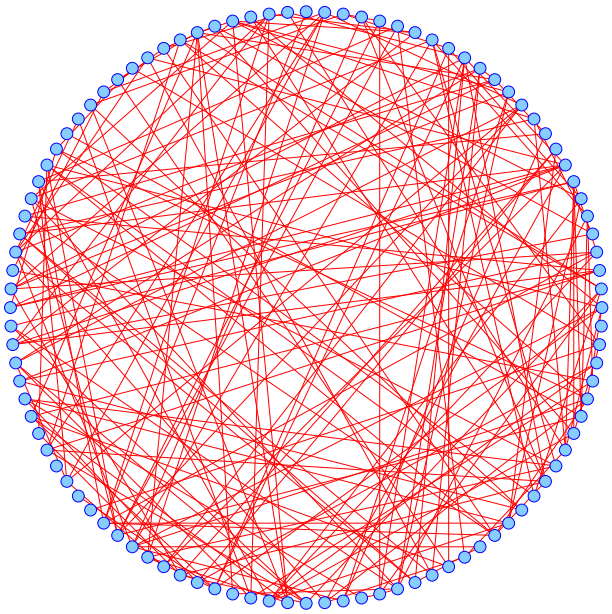


FIGURE 3.1: Une réalisation du graphe aléatoire $\mathcal{G}(100,200)$.

Définition 3.1.1. *Une application $X : \Omega \to \mathbb{R}$ entre les deux espaces probabilisables $(\Omega,\mathcal{F})$ et $(\mathbb{R},\mathcal{B})$ est une* variable aléatoire *si et seulement si*

$$X^{-1}(A) \in \mathcal{F}, \quad \forall A \in \mathcal{B}.$$

La mesure de probabilité $\mathbb{P}_X$ sur $\mathbb{R}$ définie par

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

est appelée la loi *de X.*

Remarque 3.1.1. *On peut montrer qu'il suffit de vérifier que $X^{-1}\big((-\infty,x]\big) \in \mathcal{F}$, pour tout $x \in \mathbb{R}$.*

Exemple 3.1.2. *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 vaille 3 est donnée par*

$$\mathbb{P}_X(\{3\}) = \mathbb{P}(X = 3) = \mathbb{P}(\{(1,2),(2,1)\}) = \frac{2}{36} = \frac{1}{18}.$$

Remarque 3.1.2. *Une fonction $\varphi : \mathbb{R} \to \mathbb{R}$ est dite* mesurable *si $\varphi^{-1}(A) \in \mathcal{B}$, pour tout $A \in \mathcal{B}$. Dans ce cas, on vérifie immédiatement que si $X : \Omega \to \mathbb{R}$ est une variable aléatoire, alors $\varphi(X)$ est également une variable aléatoire. Dans ce cours, à chaque fois que l'on écrit $\varphi(X)$, X une variable aléatoire, la fonction φ sera supposée mesurable. Similairement, on dira qu'une fonction $\varphi : \mathbb{R}^n \to \mathbb{R}$ est mesurable si $\varphi^{-1}(A) \in \mathcal{B}(\mathbb{R}^n)$, pour tout $A \in \mathcal{B}(\mathbb{R})$.*

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 probabilisé de départ $(\Omega,\mathcal{F},\mathbb{P})$ peut être complétement ignoré, et souvent n'est même pas spécifié, l'espace probabilisé pertinent étant $(\mathbb{R},\mathcal{B},\mathbb{P}_X)$.

3.1.2 Variables aléatoires défectives

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.

Définition 3.1.2. *Une variable aléatoire X telle que $\mathbb{P}(X=\infty)>0$ est dite* défective.

Exemple 3.1.3. *On jette une pièce de monnaie jusqu'à ce que le nombre de « pile » et de « face » obtenus soient égaux. On suppose que « face » sort avec probabilité p, indépendamment à chaque lancer. On note τ le nombre de lancers effectués. τ est à priori une variable aléatoire à valeurs dans $\mathbb{R}\cup\{+\infty\}$, $\tau=+\infty$ correspondant à une suite de lancers où l'égalité des « pile » et des « face » n'a jamais lieu.*

La loi de τ peut facilement être déduite du problème du ballot de l'Exemple 2.3.4. Bien entendu, on ne peut avoir égalité entre le nombre de « face » et de « pile » qu'aux temps pairs. Évaluons donc la probabilité de l'événement $\tau=2n$. Une façon de procéder est de conditionner sur le nombre de « face » obtenus lors des premiers $2n$ essais :

$$\mathbb{P}(\tau=2n)=\mathbb{P}(\tau=2n\,|\,n \text{ « face » lors des } 2n \text{ premiers lancers})\binom{2n}{n}p^n(1-p)^n.$$

On vérifie immédiatement que, conditionnellement au fait d'avoir n « face » lors des $2n$ premiers lancers, toutes les séries de $2n$ lancers compatibles sont équiprobables. La probabilité conditionnelle est donc égale à la probabilité qu'au cours du dépouillement des bulletins d'une élection lors de laquelle chacun des deux candidats reçoit n votes, un des deux candidats ait toujours été en avance avant que le dernier bulletin ne soit lu (et mette les deux candidats à égalité). En conditionnant sur le résultat du dernier bulletin, on voit facilement que la probabilité conditionnelle recherchée est égale à $P(n,n-1)$ (dans les notations de l'Exemple 2.3.4). Par conséquent, la loi de τ est donnée par

$$\mathbb{P}(\tau=2n)=\binom{2n}{n}p^n(1-p)^n\,P(n,n-1)=\binom{2n}{n}\frac{p^n(1-p)^n}{2n-1}.$$

Évidemment $\mathbb{P}(\tau<\infty)=\sum_{n\geq 1}\mathbb{P}(\tau=2n)\leq 1$. On vérifie facilement à partir de la formule ci-dessus que le maximum de cette probabilité est atteinte si et seulement si $p=\frac{1}{2}$, ce qui implique que τ est défective pour tout $p\neq\frac{1}{2}$.

Il n'est pas immédiat de calculer $\mathbb{P}(\tau<\infty)$ à l'aide de la formule ci-dessus lorsque $p=1/2$. On verra cependant au Chapitre 7 que $\mathbb{P}(\tau<\infty)=1$ lorsque $p=1/2$, et que τ n'est donc pas défective.

Sauf mention explicite du contraire, nous supposerons toujours les variables aléatoires non défectives.

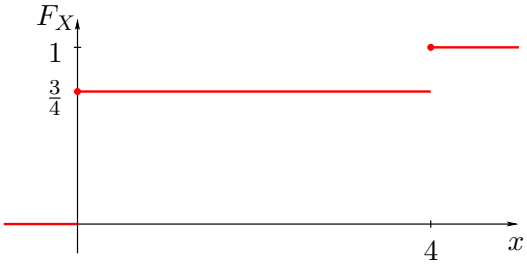


FIGURE 3.2: La fonction de répartition de la variable aléatoire X de l'Exemple 3.1.4.

3.1.3 Fonction de répartition d'une variable aléatoire

La loi $\mathbb{P}_X$ d'une variable aléatoire est une mesure de probabilité sur $\mathbb{R}$, et nous avons vu que ces dernières sont caractérisées par leur fonction de répartition. Il est donc naturel d'associer à toute variable aléatoire sa fonction de répartition.

Définition 3.1.3. *La fonction de répartition d'une variable aléatoire X est la fonction de répartition associée à sa loi, c'est-à-dire la fonction $F_X : \mathbb{R} \to [0,1]$ définie par*

$$F_X(x) = \mathbb{P}_X\big((-\infty,x]\big) = \mathbb{P}(X \leq x).$$

Exemple 3.1.4. *On jette successivement 2 pièces ; $\Omega = \{\mathsf{PP},\mathsf{PF},\mathsf{FP},\mathsf{FF}\}$. Supposons qu'un joueur mise sa fortune de 1 CHF au jeu suivant basé sur cette expérience aléatoire : à chaque fois que face sort, sa fortune double, mais si pile sort, il perd tout. La variable aléatoire X donnant sa fortune à la fin du jeu est donnée par*

$$X(\mathsf{PP}) = X(\mathsf{PF}) = X(\mathsf{FP}) = 0, \quad X(\mathsf{FF}) = 4.$$

La fonction de répartition de cette variable aléatoire est donnée par (cf. Fig. 3.2)

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0, \\ \frac{3}{4} & \text{si } 0 \leq x < 4, \\ 1 & \text{si } x \geq 4. \end{cases}$$

Lemme 3.1.1. *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 2.1.2 que

$$\mathbb{P}(X = x) = \lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \big(F_X(x) - F_X(x - \frac{1}{n})\big),$$

par le point 2. (la limite existe par monotonicité de F_X). □

Développer la théorie pour des variables aléatoires générales requiert des outils de théorie de la mesure, que l'on ne développera pas dans ce cours ; nous nous contenterons d'en donner un aperçu dans la Section 3.8. Dans la suite, nous allons principalement nous concentrer sur deux types particulièrement importants de variables aléatoires : les variables aléatoires discrètes, et les variables aléatoires à densité.

3.2 Variables aléatoires discrètes

Définition 3.2.1. *Une variable aléatoire discrète est une variable aléatoire prenant une quantité dénombrable*[1] *de valeurs différentes.*

Soit $X : \Omega \to \mathbb{R}$ une variable aléatoire discrète, et notons $X(\Omega)$ l'ensemble dénombrable des valeurs prises par X. Dans ce cas, la loi $\mathbb{P}_X$ est caractérisée par les valeurs $\mathbb{P}_X(\{x\})$, $x \in X(\Omega)$, comme cela a été discuté dans la Section 2.2.2.

Définition 3.2.2. *La fonction de masse d'une variable aléatoire discrète 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 $x \notin X(\Omega)$.

Lemme 3.2.1. *Soit X une variable aléatoire discrète. Alors,*

1. $F_X(x) = \sum_{y \in X(\Omega) : y \leq x} f_X(y)$;
2. *si x et y sont deux points consécutifs de $X(\Omega)$, alors F_X est constante sur $[x,y)$;*
3. *la hauteur du saut en $x \in X(\Omega)$ est donnée par $f_X(x)$.*

Démonstration. *1.* $F_X(x) = \mathbb{P}_X\big((-\infty,x]\big) = \mathbb{P}_X\big(\{y \in X(\Omega) \,:\, y \leq x\}\big)$.

2. Soient $x < y$ deux points consécutifs de $X(\Omega)$. Alors, pour tout $x \leq z < y$, il suit du point *1.* que

$$F_X(z) - F_X(x) = \sum_{u \in X(\Omega) : x < u \leq z} f_X(u) = 0,$$

puisque $X(\Omega) \cap (x,y) = \varnothing$.

3. Soit $x \in X(\Omega)$. Il suit du Lemme 3.1.1 que la hauteur du saut en x est donnée par

$$F_X(x) - \lim_{y \uparrow x} F_X(y) = \mathbb{P}(X = x) = f_X(x).$$

□

Exemple 3.2.1. *On veut modéliser un jeu de fléchettes. La cible est donnée par le disque de rayon 1, noté D_1. On suppose pour simplifier que le joueur est assuré de toucher la cible quelque part. De plus, on suppose que la probabilité que la fléchette se retrouve dans une région $A \subseteq D_1$ est proportionelle à la surface $|A|$ de A (à nouveau, il est impossible*

1. Ici et dans la suite, nous emploierons le qualificatif dénombrable pour tout ensemble dont les éléments peuvent être numérotés (c'est-à-dire tel qu'il existe une injection de l'ensemble dans $\mathbb{N}$), que l'ensemble soit fini ou infini.

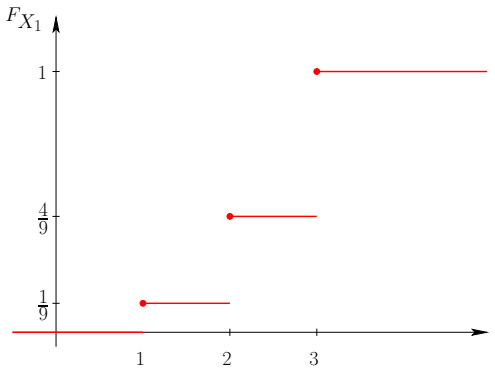


FIGURE 3.3: La fonction de répartition de la variable aléatoire X_1 de l'Exemple 3.2.1

.

de définir une surface pour tous les sous-ensembles de D_1 ; on le fait pour une algèbre de bons sous-ensembles, par exemple celle obtenue à partir des rectangles, puis on l'étend par Carathéodory à la tribu $\mathcal{B}(D_1)$ engendrée par cette algèbre). On a donc

$$\Omega = D_1 = \left\{(x,y) \in \mathbb{R}^2 \,:\, x^2 + y^2 < 1\right\},$$

muni de sa tribu borélienne $\mathcal{B}(D_1)$, et, pour pour $A \in \mathcal{B}(D_1)$, on a

$$\mathbb{P}(A) = \frac{|A|}{|D_1|} = \frac{1}{\pi}|A|.$$

Supposons à présent que la cible soit décomposée en trois anneaux concentriques, A_1,A_2 et A_3, de rayons $\frac{1}{3}$,$\frac{2}{3}$ et 1,

$$A_k = \left\{(x,y) \in D_1 \,:\, \frac{k-1}{3} \le \sqrt{x^2+y^2} < \frac{k}{3}\right\}.$$

Le joueur reçoit k points si la fléchette tombe dans l'anneau A_k, ce qui correspond à la variable aléatoire

$$X_1(\omega) = k, \qquad si\ \omega = (x,y) \in A_k.$$

La probabilité que la fléchette s'arrête dans l'anneau A_k est donnée par $|A_k|/\pi = (2k-1)/9$. La fonction de répartition de X_1 est donc donnée par (cf. Fig. 3.3)

$$F_{X_1}(x) = \mathbb{P}(X_1 \le x) = \begin{cases} 0 & si\ x < 1, \\ \frac{1}{9} & si\ 1 \le x < 2, \\ \frac{4}{9} & si\ 2 \le x < 3, \\ 1 & si\ 3 \le x. \end{cases}$$

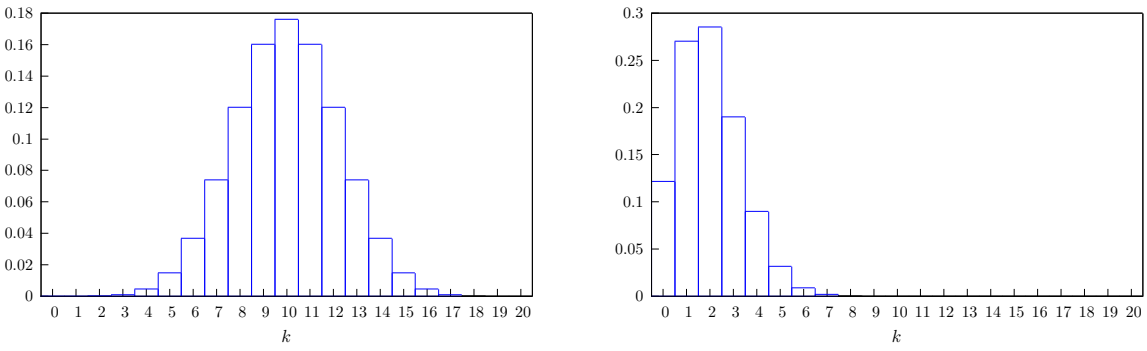


FIGURE 3.4: Loi binomiale pour $n = 20, p = 0{,}5$ (gauche) et $n = 20, p = 0{,}1$ (droite).

3.2.1 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 il est laissé en exercice de vérifier que celles-ci sont proprement normalisées (c'est-à-dire de somme 1).

Variable aléatoire constante

Une variable aléatoire X est dite constante s'il existe c 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 3.2.2. *1. Un lancer à pile ou face est une épreuve de Bernoulli.*

2. Si $A \in \mathcal{F}$, la fonction indicatrice de A, $\mathbf{1}_A : \Omega \to \{0,1\}$, définie par

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

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 ; $X \sim \text{binom}(n, p)$.

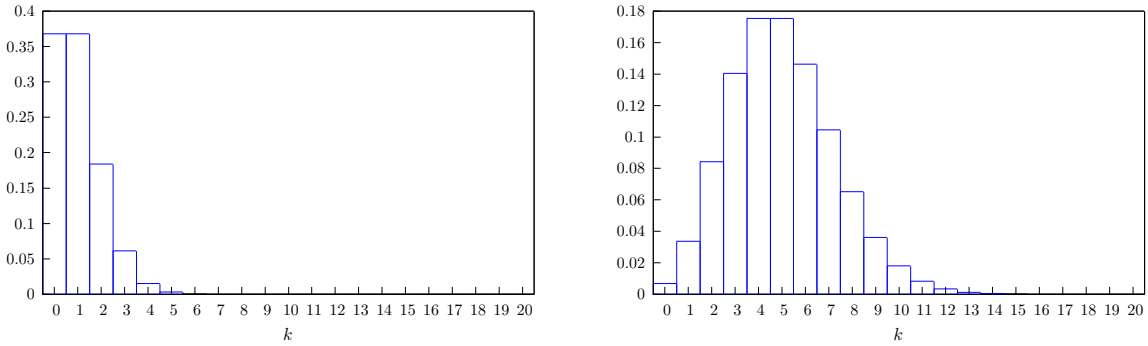


FIGURE 3.5: Loi de Poisson pour $\lambda = 1$ (gauche) et $\lambda = 5$ (droite).

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

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

Loi de Poisson

Une variable aléatoire X suit une loi de Poisson [2] de paramètre $\lambda > 0$, $X \sim \text{poisson}(\lambda)$, si elle prend ses valeurs dans $\mathbb{N}$, et la fonction de masse associée est donnée par

$$f_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0,1,2,\ldots$$

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 $n \to \infty, p \to 0$ de telle sorte que $np \to \lambda$, on voit que chacun des rapports 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 $k = 0,1,2,\ldots$

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

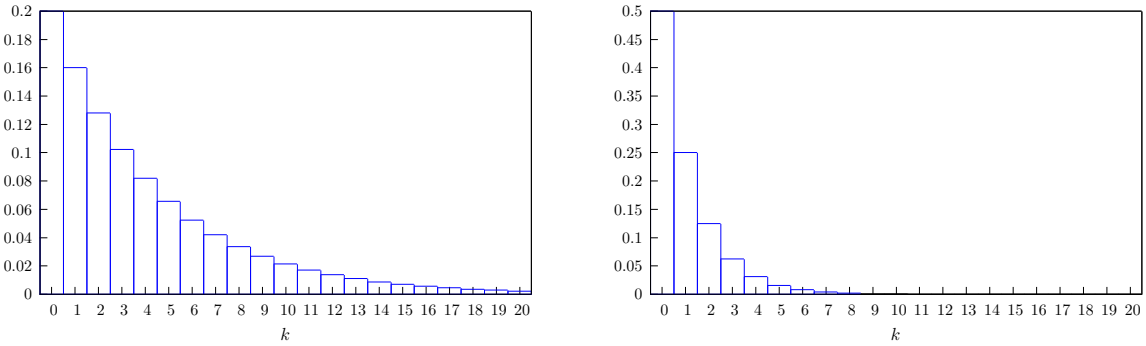


FIGURE 3.6: Loi géométrique pour $p = 0{,}2$ (gauche) et $p = 0{,}5$ (droite).

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 ; $X \sim \mathsf{geom}(p)$. La fonction de masse associée est donc donnée par

$$f_X(k) = p(1-p)^{k-1}, \qquad k = 1,2,\ldots$$

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

Lemme 3.2.2. *Soit X une variable aléatoire suivant une loi géométrique. Alors, pour tout $k \geq 1$,*

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

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 6 n'est pas sorti pendant 50 semaines consécutives à la loterie, cela ne rend pas sa prochaine apparition plus probable.

Loi hypergéométrique

Une urne contient N balles, dont b sont bleues et $r = N - b$ sont rouges. Un échantillon de n balles est tiré de l'urne, sans remise. On vérifie facilement que le nombre B de balles bleues dans l'échantillon suit la loi hypergéométrique de paramètres N, b et n, $B \sim \mathsf{hypergeom}(N, b, n)$, dont la fonction de masse est [3]

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

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

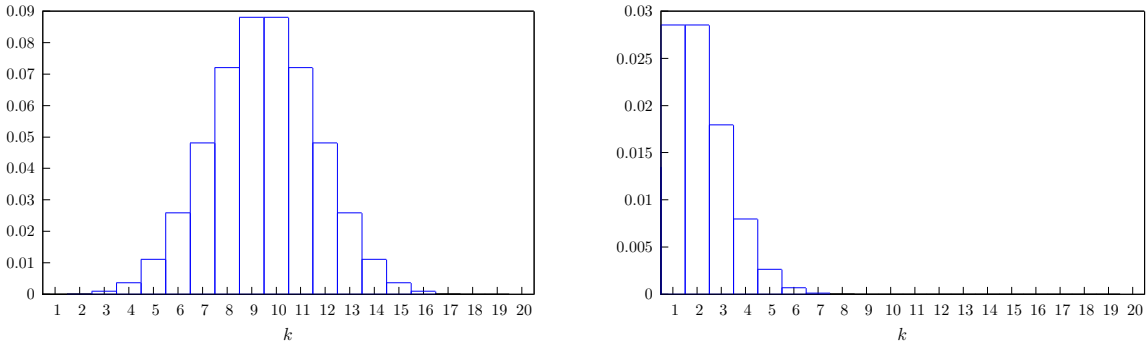


FIGURE 3.7: Loi de Pascal dans le cas $k+r=20$ pour $p=0{,}5$ (gauche) et $p=0{,}1$ (droite).

Lemme 3.2.3. *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.

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, $X \sim \mathsf{pascal}(r,p)$, dont la fonction de masse est

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

Dans certaines applications, il est utile d'autoriser le paramètre r à prendre des valeurs réelles positives pas nécessairement entières. Dans ce cas, on parle de loi binomiale négative de paramètre r et p.

3.3 Variables aléatoires à densité

Dans ce cours, toutes les intégrales

$$\int_A f(x)\,\mathrm{d}x,$$

seront prises au sens de Lebesgue, avec $A \in \mathcal{B}(\mathbb{R})$ et f une fonction Lebesgue-intégrable. Nous décrirons brièvement ce concept dans la Section 3.8. Pour le moment, il suffit d'interpréter les formules au sens de l'intégrale de Riemann [4], et la notion de Lebesgue-intégrabilité comme la condition minimale pour pouvoir définir l'intégrale de Lebesgue. Lorsqu'elles existent toutes deux, les intégrales de Lebesgue et Riemann coïncident, et comme nous le verrons plus loin, la classe des fonctions Lebesgue-intégrables est beaucoup, beaucoup plus grande que celle des fonctions Riemann-intégrables (et contient toutes les fonctions dont la valeur absolue est Riemann intégrable).

Nous nous permettrons également d'interchanger sans discussion l'ordre d'intégration lorsque nous aurons à faire à des intégrales multiples. Le Théorème de Fubini, justifiant nos calculs, est également énoncé dans la Section 3.8 (Théorème 3.8.3).

Finalement, nous emploierons la terminologie suivante, expliquée elle aussi dans la Section 3.8 : une propriété est vérifiée presque partout (p.p.) si l'ensemble des points où elle n'est pas vérifiée est de mesure de Lebesgue nulle ; ceci signifie qu'il est possible de recouvrir ce dernier par une union dénombrable d'intervalles disjoints de longueur totale arbitrairement petite.

Définition 3.3.1. *Une variable aléatoire X est à densité s'il existe une fonction Lebesgue-intégrable positive f_X telle que*

$$\mathbb{P}(X \in A) = \int_A f_X(x)\,\mathrm{d}x,$$

pour tout $A \in \mathcal{B}$. f_X est la densité (de probabilité) de X.

Remarque 3.3.1.

1. *Insistons sur le fait que la valeur $f_X(x)$ n'est **pas** une probabilité (en particulier, $f_X(x)$ peut être plus grande que 1). Par contre, il peut être utile de penser à $f_X(x)\mathrm{d}x$ comme à la probabilité que $X \in [x, x+\mathrm{d}x]$.*
2. *Lorsqu'elle existe, la densité d'une variable aléatoire n'est pas unique : par exemple, changer la valeur d'une densité f sur un ensemble de mesure de Lebesgue nulle ne change pas la valeur de l'intégrale. Toutefois, deux densités différentes f_1, f_2 de X coïncident presque partout.*

Lemme 3.3.1. *Soit X une variable aléatoire de densité f_X et de fonction de répartition F_X. Alors*

1. $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$;
2. $F_X(x) = \int_{(-\infty,x]} f_X(y)\,\mathrm{d}y$;

4. Georg Friedrich Bernhard Riemann (1826, Breselenz - 1866, Selasca), mathématicien allemand.

3. $\mathbb{P}(X = x) = 0$, *pour tout* $x \in \mathbb{R}$.

Démonstration. *1.* On a, par définition,

$$1 = \mathbb{P}(X \in \mathbb{R}) = \int_{\mathbb{R}} f(x)\,\mathrm{d}x.$$

2. $F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}\big(X \in (-\infty,x]\big) = \int_{(-\infty,x]} f_X(y)\,\mathrm{d}y$.

3. Cela suit du Lemme 3.1.1, puisque *2.* implique que F_X est continue[5]. □

Définition 3.3.2. *Une fonction de répartition* F *est dite* absolument continue *s'il existe une fonction positive Lebesgue-intégrable* f *telle que*

$$F(y) = \int_{(-\infty,y]} f(x)\,\mathrm{d}x,$$

pour tout $y \in \mathbb{R}$. f *est la* densité *associée à* F.

Remarque 3.3.2. *La continuité* absolue *d'une fonction de répartition est strictement plus forte que sa continuité. On peut fabriquer des fonctions* F *(assez pathologiques) qui sont continues, mais pas absolument continues. Les variables aléatoires correspondantes sont dites* singulières.

Remarque 3.3.3. *Comme mentionné précédemment, il n'y a pas unicité de la densité associée à une fonction de répartition* F *(ou une variable aléatoire). Cependant, on choisira une densité* f *telle que* $f(x) = F'(x)$ *en tout point où* F *est différentiable. En fait, il est possible de montrer qu'une fonction de répartition absolument continue* F *est différentiable presque partout. Comme la valeur de l'intégrale ne dépend pas de changements effectués sur un ensemble de mesure nulle, on peut choisir un représentant canonique pour la densité, en prenant* $f(x) = F'(x)$ *en tout point où* F *est différentiable, et* $f(x) = 0$ *(par exemple) ailleurs.*

Exemple 3.3.1. *Revenons à l'Exemple 3.2.1 du jeu de fléchettes. On suppose à présent que le score du joueur est donné par la distance entre le centre de la cible et la position de la fléchette. Ceci correspond à la variable aléatoire*

$$X_2(\omega) = \sqrt{x^2 + y^2}, \qquad \textit{si } \omega = (x,y).$$

5. Cela n'est pas complètement évident si f_X n'est pas bornée. Une façon de procéder est la suivante. On fixe $\epsilon > 0$. Pour $n \geq 1$, on introduit $f_n = \min(f_X,n)$. On a alors $f_n \uparrow f$ lorsque $n \to \infty$. Par le Théorème de convergence monotone (Théorème 3.8.2), on a que $\int_{\mathbb{R}} f_n(x)\mathrm{d}x \to \int_{\mathbb{R}} f_X(x)\mathrm{d}x$. On peut donc trouver n assez grand pour que $\int_{\mathbb{R}}(f_X(x) - f_n(x)) < \epsilon$. On a alors, en notant ℓ la mesure de Lebesgue (*cf.* Sous-section 3.8.1),

$$\int_A f_X(x)\mathrm{d}x = \int_A (f_X(x) - f_n(x))\mathrm{d}x + \int_A f_n(x)\mathrm{d}x \leq \int_{\mathbb{R}} (f_X(x) - f_n(x))\mathrm{d}x + n\ell(A) \leq \epsilon + n\ell(A) \leq 2\epsilon,$$

pour tout $A \in \mathcal{B}$ tel que $\ell(A) \leq \delta = \epsilon/n$. La continuité suit, puisque $\ell([x,x+\delta]) \leq \delta$.

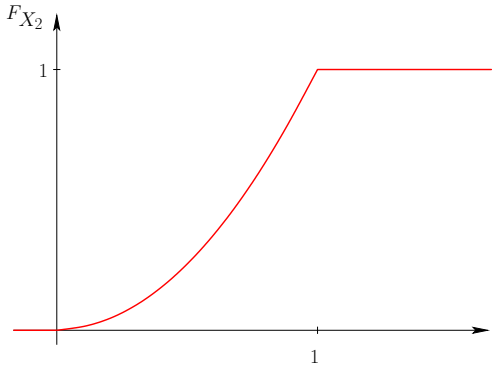


FIGURE 3.8: La fonction de répartition de la variable aléatoire X_2 de l'Exemple 3.3.1

.

Clairement, pour $0 \leq x < 1$, $\mathbb{P}(X_2 \leq x) = |D_x|/|D_1| = x^2$, *où* D_x *est le disque de rayon* x. *Par conséquent, la fonction de répartition de* X_2 *est donnée par (cf. Fig. 3.8)*

$$F_{X_2}(x) = \mathbb{P}(X_2 \leq x) = \begin{cases} 0 & si\ x < 0, \\ x^2 & si\ 0 \leq x < 1, \\ 1 & si\ 1 \leq x. \end{cases}$$

Exemple 3.3.2. *On continue avec le jeu de fléchettes. On va supposer à présent que le joueur touche la cible avec probabilité* $p \in [0,1]$. *Son score est alors calculé comme suit : s'il touche la cible, son score est égal à la distance entre le centre de la cible et la position de la fléchette, c'est-à-dire est donné par la variable aléatoire* X_2 *de l'Exemple 3.3.1. S'il rate la cible, son score est de* 2. *Notons* X_3 *cette variable aléatoire. On a alors, par la loi de la probabilité totale,*

$$\begin{aligned} \mathbb{P}(X_3 \leq x) &= \mathbb{P}(X_3 \leq x \,|\, \text{cible touchée})\mathbb{P}(\text{cible touchée}) \\ &\qquad + \mathbb{P}(X_3 \leq x \,|\, \text{cible ratée})\mathbb{P}(\text{cible ratée}) \\ &= pF_{X_2}(x) + (1-p)\mathbf{1}_{\{x \geq 2\}}. \end{aligned}$$

Par conséquent, la fonction de répartition de X_3 *est (cf. Fig. 3.9)*

$$F_{X_3}(x) = \begin{cases} 0 & si\ x < 0, \\ p\,x^2 & si\ 0 \leq x < 1, \\ p & si\ 1 \leq x < 2, \\ 1 & si\ 2 \leq x. \end{cases}$$

On voit que X_3 *est un mélange de variables aléatoires discrètes et à densité. En général, une variable aléatoire sera le mélange d'une variable aléatoire discrète, d'une variable aléatoire à densité et d'une variable aléatoire singulière.*

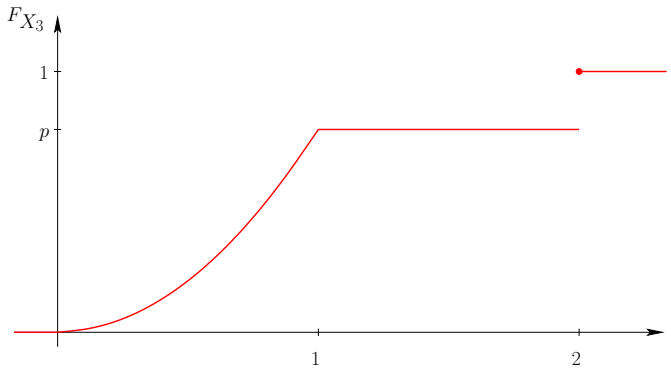


FIGURE 3.9: La fonction de répartition de la variable aléatoire X_3 de l'Exemple 3.3.2

.

3.3.1 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, et il est laissé en exercice de vérifier que ses densités sont proprement normalisées (c'est-à-dire d'intégrale 1).

Loi uniforme

X est uniforme sur $[a,b]$, noté $X \sim \mathsf{U}(a,b)$, si elle a densité

$$f_X(x) = \frac{1}{|b-a|}\,\mathbf{1}_{[a,b]}(x).$$

Ceci correspond grossièrement à dire que X prend n'importe quelle valeur entre a et b avec la même probabilité.

Loi exponentielle

X est exponentielle de paramètre $\lambda > 0$, $X \sim \mathsf{exp}(\lambda)$ si elle admet pour densité

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

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

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

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

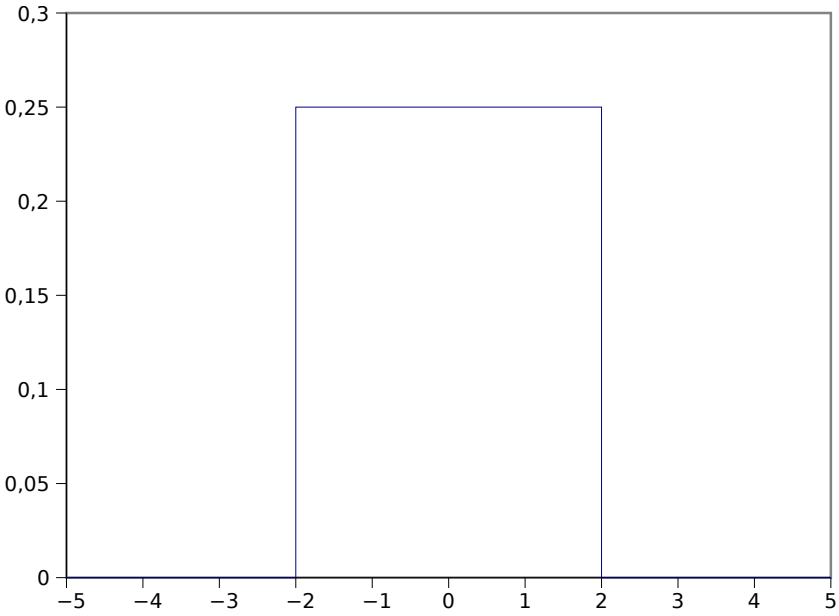


FIGURE 3.10: Loi uniforme sur $[-2,2]$.

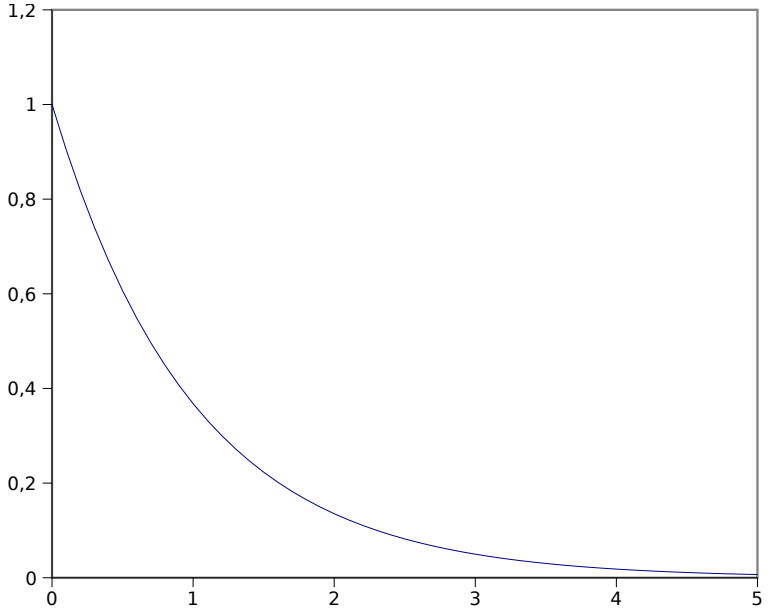


FIGURE 3.11: Loi exponentielle pour $\lambda = 1$.

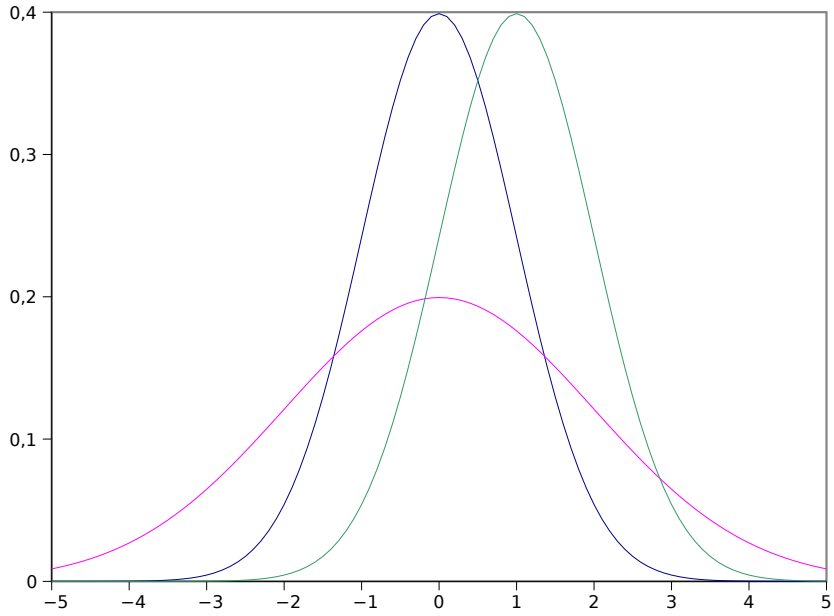


FIGURE 3.12: Loi normale : $\mu = 0$,$\sigma^2 = 1$ (bleu), $\mu = 0$,$\sigma^2 = 2$ (magenta) et $\mu = 1$,$\sigma^2 = 1$ (vert).

également que p tende vers 0 de façon à ce que p/δ tende vers une constante $\lambda > 0$. Dans ce cas,

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

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 3.2.2.

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

$$f_X(x) \equiv \varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\frac{(x-\mu)^2}{2\sigma^2}\big),$$

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.

Loi gamma

X suit la loi gamma de paramètres λ,$t > 0$, $X \sim \mathsf{gamma}(\lambda, t)$, si elle a densité

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

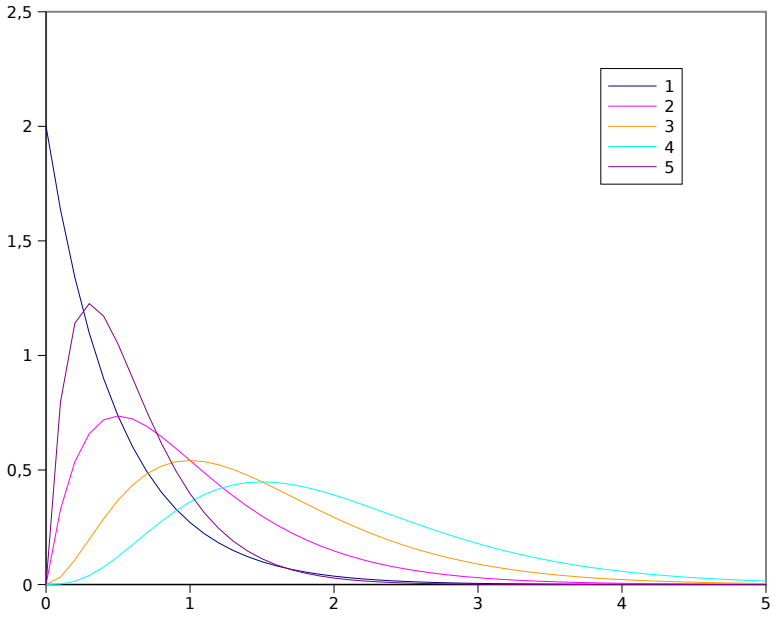


FIGURE 3.13: Loi Gamma pour $\lambda = 0.5$ et diverses valeurs de t.

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.

Loi de Cauchy

X suit la loi de Cauchy[6], $X \sim \mathsf{cauchy}$, si elle a densité

$$f_X(x) = \frac{1}{\pi(1+x^2)},$$

pour tout $x \in \mathbb{R}$.

Cette loi a un certain nombre de propriétés « pathologiques », et apparaît souvent dans des contre-exemples.

Loi bêta

X suit une loi beta de paramètres a,$b > 0$, $X \sim \mathsf{beta}(a,b)$, si elle a densité

$$f_X(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}\, \mathbf{1}_{[0,1]}(x),$$

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

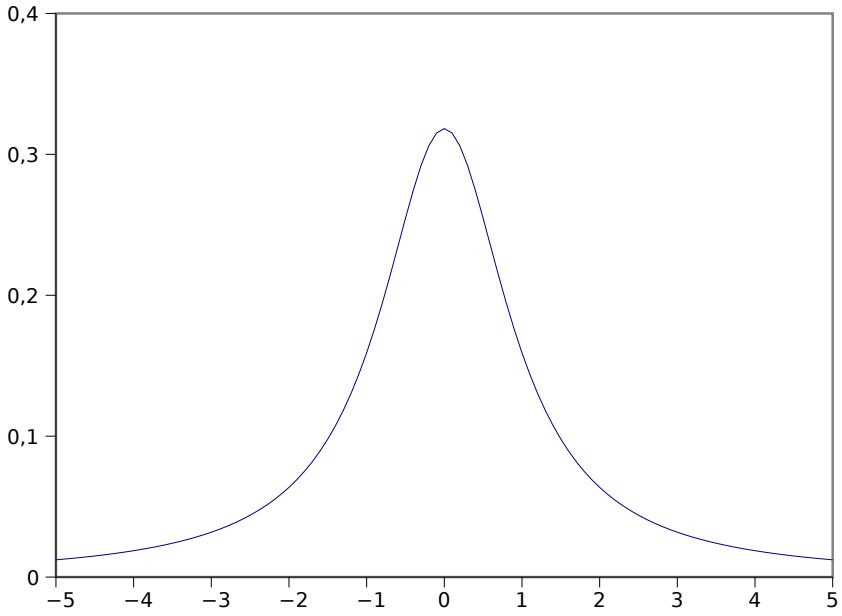


FIGURE 3.14: Loi de Cauchy.

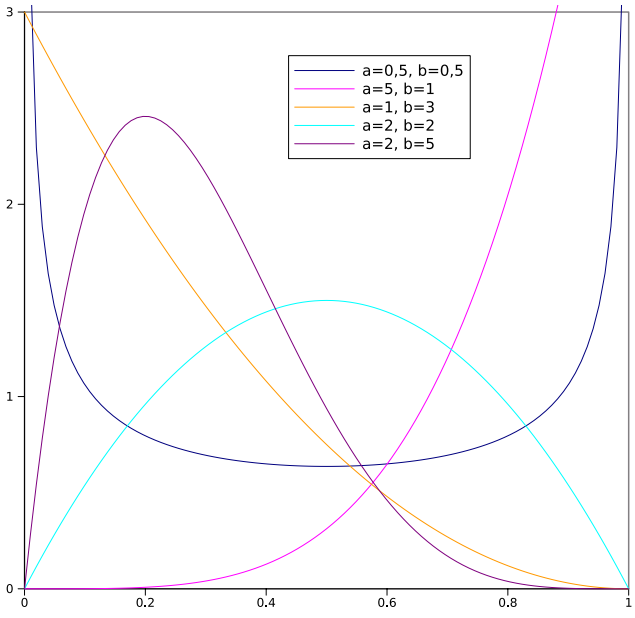


FIGURE 3.15: Loi bêta pour diverses valeurs de a et b.

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.

Loi de Student

X suit une loi de Student [7] ou loi t à ν degrés de liberté, $X \sim \mathsf{student}(\nu)$, si elle a densité

$$f_X(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}(1 + \frac{x^2}{\nu})^{-(\nu+1)/2},$$

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.

Loi de Weibull

X suit une loi de Weibull [8] de paramètres $\alpha,\beta > 0$ si elle a densité

$$f_X(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^\beta} \mathbf{1}_{[0,\infty)}(x).$$

Lorsque $\beta = 1$, 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.

3.4 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.

7. 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.

8. Ernst Hjalmar Waloddi Weibull (1887, ? ? ? – 1979, Annecy), ingénieur et mathématicien suédois.

Définition 3.4.1. *Deux variables aléatoires X et Y sur un espace probabilisé $(\Omega,\mathcal{F},\mathbb{P})$ sont* **indépendantes** *si et seulement si les événements*

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

sont indépendants pour tout $A,B \in \mathcal{B}$. 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 $A_i \in \mathcal{B}$, $i \in J$, et tout $J \subset I$ fini.

Le résultat suivant montre qu'il est suffisant de vérifier l'indépendance pour des ensembles de la forme $(-\infty,x]$, $x \in \mathbb{R}$.

Lemme 3.4.1. *La famille $(X_i)_{i\in I}$ de variables aléatoires est indépendante si et seulement si les événements*

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

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

Démonstration. Le cas discret sera fait en exercice. On admettra le cas général. □

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

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

est également indépendante.

Démonstration. φ_i étant mesurable, $\varphi_i^{-1}(A) \in \mathcal{B}$ pour tout $A \in \mathcal{B}$. Par conséquent, 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 3.4.2. *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.*

3.5 Vecteurs aléatoires

Soient X et Y deux variables aléatoires sur un même espace probabilisé $(\Omega,\mathcal{F},\mathbb{P})$. Leurs fonctions de répartition F_X et F_Y encodent toute l'information nécessaire à une étude statistique de chacune des variables. Par contre, elles ne fournissent aucune information sur les propriétés relativement l'une à l'autre.

Exemple 3.5.1. *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 3.5.2. *Si l'on considère l'évolution d'un grain de pollen dans un liquide, la position au temps t du grain de Pollen est donné par un vecteur aléatoire (X,Y,Z), dont les composantes sont les variables aléatoires correspondant aux trois coordonnées.*

Exemple 3.5.3. *On effectue n lancers à pile ou face. On peut représenter les résultats obtenus à l'aide d'un vecteur aléatoire $(X_1,\ldots,X_n)$, où X_i est la variable aléatoire prenant la valeur 1 ou 0 selon qu'un pile ou un face a été obtenu au $i^{ème}$ jet. Le nombre de pile s'exprime alors comme la variable aléatoire $X_1 + \cdots + X_n$.*

3.5.1 Loi conjointe et fonction de répartition conjointe

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

Définition 3.5.1. *On appelle loi conjointe du vecteur aléatoire $\mathbf{X} = (X_1,\ldots,X_n)$ la mesure de probabilité sur $(\mathbb{R}^n,\mathcal{B}(\mathbb{R}^n))$ définie par*

$$\mathbb{P}_{\mathbf{X}}(A) = \mathbb{P}(\mathbf{X} \in A), \qquad \forall A \in \mathcal{B}(\mathbb{R}^n).$$

Tout comme sa fonction de répartition encode toute l'information sur une variable aléatoire, la fonction de répartition conjointe encode celle d'un vecteur aléatoire. Afin de simplifier les notations, si $\mathbf{x} = (x_1,\ldots,x_n)$ et $\mathbf{y} = (y_1,\ldots,y_n)$, on écrira $\mathbf{x} \leq \mathbf{y}$ lorsque $x_i \leq y_i$, $i = 1,\ldots,n$.

Définition 3.5.2. *Soient $\mathbf{X} = (X_1,\ldots,X_n)$ un vecteur aléatoire sur un espace probabilisé $(\Omega,\mathcal{F},\mathbb{P})$. On appelle fonction de répartition conjointe de $\mathbf{X}$ la fonction $F_{\mathbf{X}} : \mathbb{R}^n \to [0,1]$ définie par*

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

Remarque 3.5.1. *La fonction de répartition conjointe $F_{\mathbf{X}}$ caractérise la loi de $\mathbf{X}$. Considérons pour simplifier $\mathbf{X} = (X_1,X_2)$. Alors la probabilité $\mathbb{P}_X([a,b]\times[c,d]) = \mathbb{P}(X_1 \in [a,b], X_2 \in [c,d]) = \mathbb{P}(X_1 \leq b, X_2 \leq d) - \mathbb{P}(X_1 \leq b, X_2 \leq c) - \mathbb{P}(X_1 \leq a, X_2 \leq d) + \mathbb{P}(X_1 \leq a, X_2 \leq c)$, et par conséquent $F_{(X_1,X_2)}$ permet de déterminer la probabilité des produits d'intervalles ; un résultat de théorie de la mesure montre que cela caractérise de façon unique la mesure $\mathbb{P}_{(X_1,X_2)}$.*

Les fonctions de répartition conjointes possèdent des propriétés tout à fait analogues à celles des fonctions de répartition.

Lemme 3.5.1. *La fonction de répartition conjointe $F_{\mathbf{X}}$ d'un vecteur aléatoire $\mathbf{X}$ satisfait*

1. $\lim_{x_i\to-\infty,i=1,\ldots,n} F_{\mathbf{X}}(\mathbf{x}) = 0$, $\lim_{x_i\to+\infty,i=1,\ldots,n} F_{\mathbf{X}}(\mathbf{x}) = 1$;
2. $\lim_{x_k\to+\infty} F_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n) = F_{(X_1,\ldots,X_{k-1},X_{k+1},\ldots,X_n)}(x_1,\ldots,x_{k-1},x_{k+1},\ldots,x_n)$;
3. *si* $\mathbf{x}_1 \leq \mathbf{x}_2$, *alors* $F_{\mathbf{X}}(\mathbf{x}_1) \leq F_{\mathbf{X}}(\mathbf{x}_2)$;
4. $F_{\mathbf{X}}$ *est continue par au-dessus, dans le sens que*

$$\lim_{\mathbf{u}\downarrow\mathbf{0}} F_{\mathbf{X}}(\mathbf{x}+\mathbf{u}) = F_{\mathbf{X}}(\mathbf{x}).$$

Démonstration. Laissée en exercice. □

La seconde affirmation du lemme montre qu'il est possible de récupérer la fonction de répartition de n'importe quelle composante d'un vecteur aléatoire : on a par exemple

$$F_{X_1}(x_1) = \lim_{x_2,\ldots,x_n\to+\infty} F_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n).$$

Définition 3.5.3. *Soit $\mathbf{X} = (X_1,\ldots,X_n)$ un vecteur aléatoire. Alors, les fonctions de répartition F_{X_i}, $i = 1,\ldots,n$, sont appelées* fonctions de répartition marginales *de $F_{\mathbf{X}}$.*

Il est possible de caractériser simplement l'indépendance de variables aléatoires en termes de leur fonction de répartition conjointe.

Lemme 3.5.2. *La famille $X_1,\ldots,X_n$ de variables aléatoires est indépendante si et seulement si*

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

Démonstration. L'affirmation suit du Lemme 3.4.1, puisque $\{\mathbf{X} \leq \mathbf{x}\} = \{X_1 \leq x_1\} \cap \cdots \cap \{X_n \leq x_n\}$. □

Comme pour les variables aléatoires, deux classes de vecteurs aléatoires sont particulièrement intéressantes : les vecteurs aléatoires discrets, et les vecteurs aléatoires à densité.

3.5.2 Vecteurs aléatoires discrets

Définition 3.5.4. *Un vecteur aléatoire $(X_1,\ldots,X_n)$ est* discret *s'il prend ses valeurs dans un sous-ensemble dénombrable de $\mathbb{R}^n$.*

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

Définition 3.5.5. *La* fonction de masse conjointe *d'un vecteur aléatoire discret $\mathbf{X} = (X_1,\ldots,X_n)$ est la fonction $f_{\mathbf{X}} : \mathbb{R}^n \to [0,1]$ définie par*

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

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

Lemme 3.5.3. *La famille $X_1,\ldots,X_n$ de variables aléatoires discrètes est indépendante si et seulement si*

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

Démonstration. Supposons $X_1,\ldots,X_n$ indépendantes. Alors, on a

$$\begin{aligned} f_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n) &= \mathbb{P}(X_1 = x_1,\ldots,X_n = x_n) \\ &= \mathbb{P}(X_1 = x_1)\cdots\mathbb{P}(X_n = x_n) \\ &= f_{X_1}(x_1)\cdots f_{X_n}(x_n). \end{aligned}$$

Réciproquement, si la fonction de masse se factorise,

$$\begin{aligned} \mathbb{P}(X_1 = x_1,\ldots,X_n = x_n) &= f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_{X_1}(x_1)\cdots f_{X_n}(x_n) \\ &= \mathbb{P}(X_1 = x_1)\cdots\mathbb{P}(X_n = x_n). \end{aligned}$$

□

Définition 3.5.6. *Étant donné une fonction de masse conjointe $f_{X_1,\ldots,X_n}$, 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 3.5.4.

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

Démonstration. Laissée en exercice. □

3.5.3 Vecteurs aléatoires à densité

Définition 3.5.7. *Un vecteur aléatoire* $\mathbf{X} = (X_1, \ldots, X_n)$ *est à densité s'il existe une fonction positive Lebesgue-intégrable* $f_{\mathbf{X}} : \mathbb{R}^n \to \mathbb{R}$ *telle que*

$$\mathbb{P}(\mathbf{X} \in A) = \int_A f_{\mathbf{X}}(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n, \qquad \forall A \in \mathcal{B}(\mathbb{R}^n).$$

$f_{\mathbf{X}}$ *est la densité conjointe du vecteur aléatoire* $\mathbf{X}$.

Remarque 3.5.2.
1. *On peut montrer qu'il suffit de vérifier la condition pour des ensembles* A *de la forme* $(-\infty, x_1] \times \cdots \times (-\infty, x_n]$, $x_1, \ldots, x_n \in \mathbb{R}$. *En d'autres termes, il suffit de vérifier que*
$$F_{\mathbf{X}}(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\mathbf{X}}(y_1, \ldots, y_n)\, \mathrm{d}y_1 \cdots \mathrm{d}y_n, \qquad \forall x_1, \ldots, x_n \in \mathbb{R}.$$
2. *À nouveau, il n'y a pas unicité de la densité conjointe, et on choisira toujours une version de* $f_{\mathbf{X}}$ *satisfaisant* $f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F_{\mathbf{X}}(x_1, \ldots, x_n)$, *en chaque point où la fonction de répartition conjointe est suffisamment différentiable.*
3. *Il peut à nouveau être utile d'interpréter* $f_{\mathbf{X}}(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n$ *comme la probabilité* $\mathbb{P}(X_1 \in [x_1, x_1 + \mathrm{d}x_1], \ldots, X_n \in [x_n, x_n + \mathrm{d}x_n])$.

Les densités des composantes d'un vecteur aléatoire $\mathbf{X}$ peuvent aisément être extraites de la densité conjointe.

Lemme 3.5.5. *Soit* $\mathbf{X} = (X_1, \ldots, X_n)$ *un vecteur aléatoire à densité. Alors, pour tout* $1 \leq k \leq n$,

$$f_{X_k}(x_k) = \int_{-\infty}^{\infty} \mathrm{d}x_1 \cdots \int_{-\infty}^{\infty} \mathrm{d}x_{k-1} \int_{-\infty}^{\infty} \mathrm{d}x_{k+1} \cdots \int_{-\infty}^{\infty} \mathrm{d}x_n\, f_{\mathbf{X}}(x_1, \ldots, x_n).$$

Définition 3.5.8. *Les densités* f_{X_k}, $1 \leq k \leq n$, *d'un vecteur aléatoire* $\mathbf{X} = (X_1, \ldots, X_n)$ *sont appelées ses densités marginales.*

L'indépendance de variables aléatoires peut se caractériser simplement en termes de leur densité conjointe.

Lemme 3.5.6. *Soit* $\mathbf{X} = (X_1, \ldots, X_n)$ *un vecteur aléatoire à densité. Les variables aléatoires* $X_1, \ldots, X_n$ *sont indépendantes si et seulement si*

$$f_{\mathbf{X}}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n),$$

pour presque tout $(x_1, \ldots, x_n)$.

Démonstration. Pour tout $x_1, \ldots, x_n \in \mathbb{R}$, il suit du Lemme 3.5.2 que

$$\begin{aligned} F_{\mathbf{X}}(x_1, \ldots, x_n) &= \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \cdots \mathbb{P}(X_n \leq x_n) \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1}(y_1) \cdots f_{X_n}(y_n)\, \mathrm{d}y_1 \cdots \mathrm{d}y_n, \end{aligned}$$

et par conséquent $f_{X_1}(x_1) \cdots f_{X_n}(x_n)$ est une densité de $\mathbb{P}_{\mathbf{X}}$. □

Exemple 3.5.4. *Retournons une fois de plus à l'exemple du jeu de fléchettes ;* $\Omega = D_1 = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$. *On considère les quatre variables aléatoires suivantes :* $X(\omega) = x$, $Y(\omega) = y$, $R(\omega) = \sqrt{x^2+y^2}$ *et* $\Theta(\omega) = \operatorname{atan}(y/x)$. *Ainsi les vecteurs aléatoires* (X,Y) *et* (R,Θ) *correspondent à la position de la fléchette en coordonnées cartésiennes et polaires, respectivement. Déterminons leurs lois conjointes, ainsi que les lois de ces quatre variables aléatoires.*

Pour le couple (X,Y), *on a*

$$\mathbb{P}((X,Y) \in A) = |A \cap D_1|/\pi = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{\pi}\mathbf{1}_{\{x^2+y^2<1\}}\mathbf{1}_A \,\mathrm{d}x\mathrm{d}y,$$

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,* $f_Y(y) = \frac{2}{\pi}\sqrt{1-y^2}$. *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.*

Pour le couple (R,Θ), *on a*

$$\mathbb{P}((R,\Theta) \in A) = |A \cap D_1|/\pi = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{\pi}\mathbf{1}_{\{0\le r<1, 0\le\theta<2\pi\}}\mathbf{1}_A \, r\mathrm{d}r\mathrm{d}\theta,$$

d'où l'on tire la densité 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 $\mathbf{X} = (X_1,\ldots,X_n)$ 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 $\Psi : U \to \mathbb{R}^n$, $\Psi(\mathbf{x}) = (\psi_1(\mathbf{x}),\ldots,\psi_n(\mathbf{x}))$. 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 3.5.1. *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 Lebesgue-intégrable* $f : V \to \mathbb{R}$, *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 est une fonction Lebesgue-intégrable quelconque repose sur la construction de l'intégrale de Lebesgue, et nous ne la ferons pas ici. □

Corollaire 3.5.1. *On considère un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ 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 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 3.5.7. *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 $X + Y$ est donnée par la convolution des densités 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 conjointe se factorise et la seconde affirmation suit. □

Une autre conséquence utile (et immédiate) du Corollaire précédent est le résultat suivant.

Lemme 3.5.8. *Soit X une variable aléatoire à densité et $a,b \in \mathbb{R}$, $a \neq 0$. La densité de la variable aléatoire $aX+b$ est donnée par*

$$f_{aX+b}(y) = \frac{1}{|a|} f_X\big((y-b)/a\big).$$

Démonstration. Laissée en exercice. □

On déduit immédiatement des deux lemmes précédents l'important résultat suivant.

Lemme 3.5.9. *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 3.5.8, ces variables suivent respectivement les lois $\mathcal{N}(0,\sigma_1^2)$ et $\mathcal{N}(0,\sigma_2^2)$. Une application du Lemme 3.5.7 montre que la densité 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

$$\sigma_2^2x^2 + \sigma_1^2(z-x)^2 = (\sqrt{\sigma_1^2+\sigma_2^2}\,x - \frac{\sigma_1^2 z}{\sqrt{\sigma_1^2+\sigma_2^2}})^2 + \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}z^2,$$

l'intégration sur x montre que cette densité 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 3.5.9. *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 3.5.10. *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 3.5.9. Pour la seconde affirmation, il suffit d'observer que, pour tout $\mathbf{a} \in \mathbb{R}^n$,

$$\langle \mathbf{a}, A\mathbf{X}\rangle = \langle A^{\mathrm{t}}\mathbf{a}, \mathbf{X}\rangle$$

est bien gaussien. □

Remarque 3.5.3. *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 3.6.8).*

Exemple 3.5.5. *Un exemple de vecteur aléatoire gaussien est le vecteur* $(X_1,\dots,X_n)$ *composé de* n *variables aléatoires* indépendantes *suivant des lois normales. En effet,* $a_1X_1+\cdots+a_nX_n$ *est une somme de variables aléatoires normales, et donc, par le Lemme 3.5.9, suit également une loi normale.*

Il suit de l'exemple précédent et du Lemme 3.5.10 que l'image d'un vecteur $(X_1,\dots,X_n)$ 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 3.5.11. *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 conjointe*

$$f_{\mathbf{X}}(\mathbf{x})=\frac{1}{\sqrt{(2\pi)^n\det\mathbf{C}}}\exp\left(-\tfrac{1}{2}\langle\mathbf{x}-\boldsymbol{\mu},\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\rangle\right), \tag{3.1}$$

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 $\mathbf{C}^{-1}=\mathbf{O}^{\mathrm{t}}\mathbf{D}\mathbf{O}$. 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 3.5.10 que $\mathbf{X}$ est gaussien. □

3.6 Espérance, variance, covariance et moments

3.6.1 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 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, $\sum_{x\in E}xf(x)$ devrait fournir une approximation asymptotiquement correcte de m ; on appelle la quantité correspondante espérance.

Définition 3.6.1. *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*

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

Soit X une variable aléatoire avec densité f_X. On dit que X admet une espérance si

$$\int_{\mathbb{R}} |x| f_X(x)\,\mathrm{d}x < \infty.$$

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

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x.$$

Remarque 3.6.1. *Les conditions d'absolue sommabilité sont importantes : dans le cas discret, elle assure que l'espérance ne dépend pas de l'ordre dans lequel les termes sont sommés. Dans le cas à densité, elle est nécessaire à la définition même de l'intégrale au sens de Lebesgue (cf. Section 3.8).*

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.

Exemple 3.6.1. *Évidemment, les variables aléatoires défectives ont toujours une espérance infinie, ou indéfinie. Cependant, des variables aléatoires finies très naturelles possèdent une espérance infinie. C'est le cas, par exemple, de la variable aléatoire τ de l'Exemple 3.1.3 dans le cas d'une pièce équilibrée, $p = \frac{1}{2}$. On a vu que, pour ce choix de p, τ est presque sûrement finie.*

Par la formule de Stirling et le calcul de la loi de τ effectué précédemment, on voit que $f_\tau(2n) \geq cn^{-3/2}$, pour une constante $c > 0$ et tout $n \geq 1$. Par conséquent, $\mathbb{E}(\tau) = \sum_{n\geq 1} 2n f_\tau(2n) \geq 2c \sum_{n\geq 1} n^{-1/2} = \infty$.

Ainsi, lors d'une série de lancers d'une pièce équilibrée, le nombre moyen de lancers nécessaires avant que le nombre de « face » et de « pile » obtenus ne coïncident est infini !

Nous n'avons défini ici l'espérance que pour des variable aléatoire discrètes et à densité. La définition générale sera donnée dans la Section 3.8. Les propriétés et définitions données ci-dessous restent vraies dans le contexte général.

On voit que ces deux définitions sont formellement les mêmes si on interprète $f_X(x)\mathrm{d}x$ comme $\mathbb{P}(X \in [x, x + \mathrm{d}x])$.

Le résultat élémentaire suivant est extrêmement utile.

Lemme 3.6.1. *Soit A,$B \in \mathcal{F}$. 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. □

Remarque 3.6.2. *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. En particulier, on pourrait être tenté de dire 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. Supposons que je cache une pièce de 2 francs dans une de mes mains et vous invite à payer un droit de participation au jeu suivant : vous choisissez une de mes mains, et si celle-ci contient la pièce, elle est à vous, sinon je garde votre mise. Quelle devrait être la mise pour que le jeu soit équitable ? Il semble qu'une mise de 1 franc soit un bon compromis : après tout, cela correspond au gain moyen. Mais considérez à présent la situation suivante : si au lieu de cacher 2 francs, je cache 1000 francs, quelle serait une mise équitable ? Il semble peu probable que des personnes aux revenus modestes soient prêtes à risquer 500 francs pour pouvoir jouer !*

Une autre façon de se convaincre de cela est de considérer 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 ? Vous pouvez vérifier que l'espérance est infinie ! C'est le célèbre paradoxe de Saint-Pétersbourg.

Démontrons à présent quelques propriétés élémentaires de l'espérance.

Lemme 3.6.2. *1. (Linéarité) $\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}(X) + \beta \mathbb{E}(Y)$.*

2. $X \geq 0 \implies \mathbb{E}(X) \geq 0$.

3. Si $\mathbb{P}(X = c) = 1$ pour un $c \in \mathbb{R}$, alors $\mathbb{E}(X) = c$.

4. $\mathbb{E}(|X|) \geq |\mathbb{E}(X)|$.

Démonstration. 1. On commence par le cas où X et Y sont des variables aléatoires discrètes. On écrit, avec $E = X(\Omega)$, $F = Y(\Omega)$, et $U = \{u = \alpha x + \beta y \,:\, x \in E, y \in F\}$,

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

Dans le cas de variables aléatoires à densité, on a

$$\begin{aligned}
\mathbb{E}(\alpha X+\beta Y) &= \int_{-\infty}^{\infty} u f_{\alpha X+\beta Y}(u)\,\mathrm{d}u \\
&= \int_{-\infty}^{\infty} \mathrm{d}u\, u \int_{-\infty}^{\infty} \mathrm{d}x\, f_{\alpha X,\beta Y}(x,u-x) && \text{(Lemme 3.5.7)} \\
&= \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}u\, u f_{\alpha X,\beta Y}(x,u-x) \\
&= \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y\, (y+x) f_{\alpha X,\beta Y}(x,y) && (y=u-x) \\
&= \int_{-\infty}^{\infty} \mathrm{d}x\, x f_{\alpha X}(x) + \int_{-\infty}^{\infty} \mathrm{d}y\, y f_{\beta Y}(y) && \text{(Lemme 3.5.5)} \\
&= \frac{1}{\alpha}\int_{-\infty}^{\infty} \mathrm{d}x\, x f_X(x/\alpha) + \frac{1}{\beta}\int_{-\infty}^{\infty} \mathrm{d}y\, y f_Y(y/\beta) && \text{(Lemme 3.5.8)} \\
&= \alpha\mathbb{E}(X)+\beta\mathbb{E}(Y).
\end{aligned}$$

2. et *3.* sont immédiats. *4.* suit de l'inégalité triangulaire : Par exemple, dans le cas à densité,

$$\begin{aligned}
\Big|\int_{-\infty}^{\infty} x\, f_X(x)\,\mathrm{d}x\Big| &= \Big|\int_{-\infty}^{0} x\, f_X(x)\,\mathrm{d}x + \int_{0}^{\infty} x\, f_X(x)\,\mathrm{d}x\Big| \\
&\le -\int_{-\infty}^{0} x\, f_X(x)\,\mathrm{d}x + \int_{0}^{\infty} x\, f_X(x)\,\mathrm{d}x \\
&= \int_{0}^{\infty} x\,\big(f_X(x)+f_X(-x)\big)\,\mathrm{d}x \\
&= \int_{0}^{\infty} x\, f_{|X|}(x)\,\mathrm{d}x,
\end{aligned}$$

puisque $f_{|X|}(x) = f_X(x)+f_X(-x)$ pour $x>0$ (en effet, $\mathbb{P}(|X|\in A) = \mathbb{P}(X\in A)+\mathbb{P}(X\in -A)$, pour tout $A\in\mathcal{B}(\mathbb{R}^+)$) et 0 sinon. □

Exemple 3.6.2. *Soit X une variable aléatoire. On désire trouver le nombre $a\in\mathbb{R}$ qui approxime le mieux 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 3.6.3. *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 3.6.1 que*

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

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)	$r(1-p)/p.$	$r(1-p)/p^2$
Uniforme (a,b)	$(a+b)/2$	$(b-a)^2/12$
Exponentielle (λ)	$1/\lambda$	$1/\lambda^2$
Normale (μ,σ^2)	μ	σ^2
Gamma (λ,t)	t/λ	t/λ^2
Cauchy	Pas définie	Pas définie
Beta (a,b)	$a/(a+b)$	$ab/[(a+b)^2(a+b+1))]$

TABLE 3.1: L'espérance et la variance de quelques lois importantes, en fonction de leurs paramètres.

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.

Lemme 3.6.3. *La table 3.1 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 3.6.2, 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 4.1.1.
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 (la vérification est laissée en exercice). On a alors
$$\mathbb{E}(X) = -r + \sum_{i=1}^{r} \mathbb{E}(X_i) = \frac{r}{p} - r.$$
7. *Loi uniforme.* Si X suit une loi $\mathsf{U}(a,b)$, alors $\mathbb{E}(X) = \frac{1}{b-a}\int_a^b x\,\mathrm{d}x = \frac{a+b}{2}$.
8. *Loi exponentielle.* Dans ce cas, $\mathbb{E}(X) = \lambda\int_0^\infty xe^{-\lambda x}\mathrm{d}x = \int_0^\infty e^{-\lambda x}\,\mathrm{d}x = \lambda^{-1}$.
9. *Loi normale.* Soit X de loi $\mathcal{N}(\mu,\sigma^2)$. La variable aléatoire $X-\mu$ suit une loi $\mathcal{N}(0,\sigma^2)$, et, par linéarité et symétrie, $\mathbb{E}(X) = \mathbb{E}(X-\mu) + \mu = \mu$.
10. *Loi gamma.* Il suffit d'observer que
$$\begin{aligned}\mathbb{E}(X) &= \frac{\lambda^t}{\Gamma(t)}\int_0^\infty xx^{t-1}e^{-\lambda x}\,\mathrm{d}x = \frac{\Gamma(t+1)}{\Gamma(t)\lambda}\cdot\frac{\lambda^{t+1}}{\Gamma(t+1)}\int_0^\infty x^{t+1-1}e^{-\lambda x}\,\mathrm{d}x \\ &= \frac{\Gamma(t+1)}{\Gamma(t)\lambda} = \frac{t}{\lambda}.\end{aligned}$$
11. *Loi de Cauchy.* Pour x grand, $xf_X(x) = O(1/x)$. $|x|f_X$ n'est donc pas intégrable, et l'espérance n'existe pas.

12. *Loi bêta.* À nouveau,

$$\begin{aligned}\mathbb{E}(X) &= \frac{1}{B(a,b)} \int_0^1 x x^{a-1}(1-x)^{b-1}\,\mathrm{d}x \\ &= \frac{B(a+1,b)}{B(a,b)} \cdot \frac{1}{B(a+1,b)} \int_0^1 x^{a+1-1}(1-x)^{b-1}\,\mathrm{d}x \\ &= \frac{B(a+1,b)}{B(a,b)} = \frac{a}{a+b}.\end{aligned}$$

□

Exemple 3.6.4. *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. Je place dans chacune de deux enveloppes un papier sur lequel est écrit un nombre entier (positif ou négatif) arbitraire, mais différent dans chaque enveloppe. 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 les mêmes deux nombres, alors la fraction de bonnes réponses serait strictement supérieure à $1/2$).

Lorsque $\mathbf{X} = (X_1, \ldots, X_n)$ est un vecteur aléatoire, et $\varphi : \mathbb{R}^n \to \mathbb{R}$ est une fonction mesurable, nous avons vu que $\varphi(\mathbf{X})$ définit une variable aléatoire. Son espérance est aisément calculée.

Lemme 3.6.4. *Soit $\mathbf{X} = (X_1, \ldots, X_n)$ un vecteur aléatoire et $\varphi : \mathbb{R}^n \to \mathbb{R}$ une fonction mesurable. Alors on a*

1. pour un vecteur aléatoire discret,

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

dès que cette somme est absolument convergente ;

2. pour un vecteur aléatoire à densité,

$$\mathbb{E}(\varphi(\mathbf{X})) = \int_{\mathbb{R}^n} \varphi(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\,\mathrm{d}x_1 \ldots \mathrm{d}x_n,$$

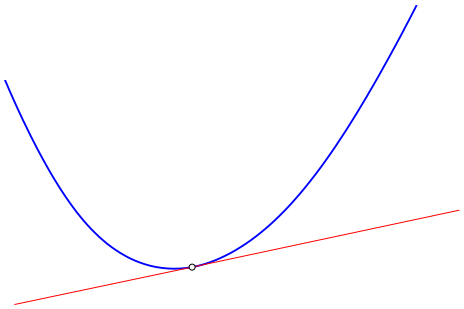


FIGURE 3.16: En chaque point du graphe d'une fonction convexe, il passe au moins une droite restant toujours sous le graphe de la fonction.

dès que cette intégrale est absolument convergente.

Démonstration. 1. 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.

2. La preuve dans le cas d'une variable aléatoire à densité est plus complexe, l'argument le plus naturel reposant sur la définition générale de l'espérance donnée dans la Section 3.8. Nous nous contenterons d'accepter ce résultat. □

Définition 3.6.2. *Une fonction $\varphi : \mathbb{R} \to \mathbb{R}$ est convexe si et seulement si : $\forall x \in \mathbb{R}, \exists a \in \mathbb{R} : \forall y \in \mathbb{R}, \varphi(y) \geq \varphi(x) + a(y-x)$ (cf. Fig. 3.16). Si l'inégalité est toujours stricte lorsque $y \neq x$, alors on dit que φ est strictement convexe .*

Théorème 3.6.1 (Inégalité de Jensen[9])**.** *Soient X une variable aléatoire admettant une espérance et $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction convexe. Alors*

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

avec égalité si et seulement si $\mathbb{P}(X = \mathbb{E}(X)) = 1$, lorsque φ est strictement convexe.

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

Démonstration. On utilise la définition de la convexité de φ, avec $y = X$ et $x = \mathbb{E}(X)$. On a alors, pour un certain $a \in \mathbb{R}$,

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

En prenant l'espérance de chacun des membres, on obtient bien

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

□

3.6.2 Variance, moments d'ordre supérieurs

Définition 3.6.3. *On appelle $\mathbb{E}(X^n)$ le moment d'ordre n de la variable aléatoire X, pourvu que cette espérance soit bien définie.*

Remarque 3.6.3. *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*

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

puisque $n/k > 1$.

Remarque 3.6.4. *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,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.

Définition 3.6.4. *Soit X une variable aléatoire dont l'espérance existe. On appelle variance de X la quantité*

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

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

Lemme 3.6.5. *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. *1.* à *4.* sont évidents. Pour *5.*, on peut utiliser l'observation triviale que $(X+Y)^2 \leq 2X^2 + 2Y^2$. □

Lemme 3.6.6. *La table 3.1 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 :
$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 1 \cdot p + 0 \cdot (1-p) - p^2 = p(1-p).$$
2. *Loi binomiale.* Voir l'Exemple 3.6.6.
3. *Loi de Poisson.* Une façon de calculer la variance d'une variable aléatoire X suivant une loi de Poisson est la suivante.
$$\mathbb{E}\big(X(X-1)\big) = \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda}\lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = \lambda^2.$$
Par conséquent, $\mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}\big(X(X-1)\big) - \mathbb{E}(X)^2 + \mathbb{E}(X) = \lambda$.
4. *Loi géométrique.* 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 $G\check{}(x) = \frac{2}{(1-x)^3} = \sum_{k=1}^{\infty} k(k-1)\,x^{k-2}$. Par conséquent,
$$\mathrm{Var}(X) = p(1-p)G\check{}(1-p) + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$
5. *Loi hypergéométrique.* Voir l'Exemple 4.1.1.
6. *Loi de Pascal.* Voir l'Exemple 3.6.6.
7. *Loi uniforme.* $\mathbb{E}(X^2) = \frac{1}{b-a}\int_a^b x^2 \mathrm{d}x = \frac{1}{3}(a^2 + ab + b^2)$; le résultat suit puisque $\mathbb{E}(X) = (a+b)/2$.
8. *Loi exponentielle.* $\mathbb{E}(X^2) = \lambda \int_0^{\infty} x^2 e^{-\lambda x}\,\mathrm{d}x = 2/\lambda^2$; le résultat suit puisque $\mathbb{E}(X) = 1/\lambda$.
9. *Loi normale.* On peut supposer que X suit une loi $\mathcal{N}(0,1)$ (car $Y = \mu + \sigma X$ suit alors une loi $\mathcal{N}(\mu,\sigma^2)$, et $\mathrm{Var}(Y) = \sigma^2\mathrm{Var}(X)$). On a $\mathrm{Var}(X) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-x^2/2}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x \cdot xe^{-x^2/2}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x = 1$, par intégration par partie.

10. *Loi gamma.* En procédant comme pour l'espérance,

$$\mathbb{E}(X^2) = \frac{\Gamma(t+2)}{\Gamma(t)\lambda^2} = \frac{t(t+1)}{\lambda^2}.$$

Par conséquent, $\mathrm{Var}(X) = t(t+1)/\lambda^2 - (t/\lambda)^2 = t/\lambda^2$.

11. *Loi de Cauchy.* L'espérance n'existe pas, et donc on ne peut pas définir la variance (observez, cependant, que le second moment existe, $\mathbb{E}(X^2) = \infty$).

12. *Loi bêta.* En procédant comme pour l'espérance,

$$\mathbb{E}(X^2) = \frac{B(a+2,b)}{B(a,b)} = \frac{a(a+1)}{(a+b)(a+b+1)},$$

et donc $\mathrm{Var}(X) = a(a+1)/(a+b)(a+b+1) - a^2/(a+b)^2 = ab/(a+b)^2(a+b+1)$. □

3.6.3 Covariance et corrélation

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

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

Ceci motive la définition suivante.

Définition 3.6.5. *On appelle* covariance *de deux variables aléatoires X et Y la quantité*

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

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

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

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

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

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

Démonstration. Laissée en exercice. □

En statistiques, une autre quantité est souvent utilisée pour mesurer la corrélation entre deux variables aléatoires.

Définition 3.6.6. *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 3.6.2 (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). Dans ce cas, on a, pour $a,b \in \mathbb{R}$,

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

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 $a,b \in \mathbb{R}$ 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 3.6.1.

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

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 3.6.8. *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. Puisque

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

on peut supposer sans perte de généralité que $\mathbb{E}(X) = \mathbb{E}(Y) = 0$. On obtient alors

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

et le membre de droite est minimum lorsque $b = 0$ et

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

□

Exemple 3.6.5. *En physiologie, la loi de Kleiber*[10] *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. Ceci peut se faire en calculant, à partir d'un échantillon, le coefficient de corrélation $\rho(X,Y)$. L'estimation des paramètres a et α à partir d'un échantillon est du ressort de la Statistique. Nous étudierons ce type de problèmes dans le Chapitre 6.

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

3.6.4 Vecteurs aléatoires

Les notions d'espérance et de covariance ont une extension naturelle aux vecteurs aléatoires.

Définition 3.6.7. *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 3.6.8. *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 $1 \leq i,j \leq n$*.*

Le lemme suivant justifie la notation $\mathcal{N}(\boldsymbol{\mu},\mathbf{C})$ du Lemme 3.5.11.

Lemme 3.6.9. *Soit* $\mathbf{X}$ *un vecteur gaussien de loi* $\mathcal{N}(\boldsymbol{\mu},\mathbf{C})$*. Alors*

$$\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}, \qquad \mathrm{Cov}(\mathbf{X},\mathbf{X}) = \mathbf{C}.$$

Démonstration. Elle se fait soit par simple intégration (exercice), soit, de façon plus élégante, à l'aide des fonctions caractéristiques, comme expliqué dans la Sous-section 4.2.2. □

3.6.5 Absence de corrélation et indépendance

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

Lemme 3.6.10. *Deux variables aléatoires indépendantes dont l'espérance existe sont non-corrélées.*

Démonstration. On applique le Lemme 3.6.4 avec la fonction $\varphi : \mathbb{R}^2 \to \mathbb{R}$, $\varphi(x,y) = xy$. Cela donne, dans le cas à densité,

$$\begin{aligned}\mathbb{E}(XY) = \mathbb{E}(\varphi(X,Y)) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(x,y) f_{(X,Y)}(x,y)\,\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(x,y) f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y = \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

La preuve est complètement similaire dans le cas discret. □

Il suit que si $X_1,\dots,X_n$ sont indépendantes, alors la variance de leur somme est égale à la somme de leurs variances.

Exemple 3.6.6. *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 3.6.7. *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.*

Il existe toutefois une classe importante de variables aléatoires pour lesquelles il y a équivalence entre ces deux notions.

Théorème 3.6.3. *Les composantes d'un vecteur aléatoire gaussien $\mathbf{X}$ sont indépendantes si et seulement si elles sont non-corrélées.*

Démonstration. Nous démontrerons ce théorème une fois la notion de fonction caractéristique introduite (*cf.* fin de la Sous-section 4.2.2) □

Exemple 3.6.8. *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.

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

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

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

pour tout $J \subseteq I$ fini.

Démonstration. Nous ne traitons que du cas continu. Le cas discret est analogue.
1. $\Longrightarrow$ *2.* Cela suit immédiatement du Lemme 3.6.4 et de la factorisation de la densité conjointe : pour tout $J = \{i_1, \dots, i_n\} \subseteq I$,

$$\begin{aligned}\mathbb{E}\big(\prod_{i\in J}\varphi_i(X_i)\big) &= \int_{\mathbb{R}^n} \varphi_{i_1}(x_1)\cdots\varphi_{i_n}(x_n) f_{(X_{i_1},\dots,X_{i_n})}(x_1,\dots,x_n)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \\ &= \int_{\mathbb{R}^n} \varphi_{i_1}(x_1)\cdots\varphi_{i_n}(x_n) f_{X_{i_1}}(x_1)\cdots f_{X_{i_n}}(x_n)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \\ &= \prod_{i\in J}\mathbb{E}(\varphi_i(X_i)).\end{aligned}$$

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

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

□

3.6.6 Espérance conditionnelle

Soient X et Y deux variables aléatoires discrètes sur $(\Omega,\mathcal{F},\mathbb{P})$. La notion de probabilité conditionnelle $\mathbb{P}(A\,|\,B)$, 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 3.6.9. *Soient X,Y deux variables aléatoires discrètes. La fonction de masse conditionnelle de Y sachant que $X = x$ est la fonction $f_{Y|X}(\,\cdot\,|\,x) : \mathbb{R} \to [0,1]$ définie par*

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

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

Soient X et Y deux variables aléatoires possédant la densité 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 $\mathbb{P}(Y \le y\,|\,X = x)$ 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,

$$\begin{aligned}\mathbb{P}(Y \le y\,|\,x \le X \le x + \mathrm{d}x) &= \frac{\mathbb{P}(Y \le y, x \le X \le x + \mathrm{d}x)}{\mathbb{P}(x \le X \le x + \mathrm{d}x)} \\ &\simeq \frac{\mathrm{d}x\,\int_{-\infty}^{y} f_{(X,Y)}(x,v)\mathrm{d}v}{f_X(x)\,\mathrm{d}x} \\ &= \int_{-\infty}^{y} \frac{f_{(X,Y)}(x,v)}{f_X(x)}\,\mathrm{d}v.\end{aligned}$$

En laissant $\mathrm{d}x \downarrow 0$, le membre de gauche converge vers ce que l'on aimerait définir comme $\mathbb{P}(Y \le y\,|\,X = x)$, et le membre de droite conduit donc à la définition suivante.

Définition 3.6.10. *Soient X,Y deux variables aléatoires avec densité conjointe $f_{(X,Y)}$. La densité 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 3.6.5. *Soient X_1 et X_2 deux variables aléatoires indépendantes de loi* exp(1). *Quelle est la densité 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 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 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 définir l'espérance conditionnelle, comme étant l'espérance sous la loi conditionnelle.

Définition 3.6.11. *Soient X,Y deux variables aléatoires discrètes. On appelle espérance conditionnelle de Y étant donné X la variable aléatoire*

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

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

Soient X et Y deux variables aléatoires de densité conjointe $f_{(X,Y)}$. L'espérance conditionnelle de Y sachant X est la variable aléatoire

$$\mathbb{E}(Y \,|\, X)(\cdot) \equiv \mathbb{E}(Y \,|\, X = \cdot) = \int_{\mathbb{R}} y\, f_{Y|X}(y \,|\, \cdot)\, \mathrm{d}y,$$

pourvu que $\int_{\mathbb{R}} |y|\, f_{Y|X}(y \,|\, \cdot)\, \mathrm{d}y < \infty$.

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

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

Plus généralement, 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 première affirmation est un cas particulier de la seconde : il suffit de choisir $\varphi \equiv 1$. Démontrons donc la seconde affirmation. Dans le cas discret, il suit du Lemme 3.6.4 que

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

Dans le cas à densité, la preuve est formellement identique,

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

□

En dehors de ses applications immédiates, l'intérêt de ce résultat est qu'on peut prendre cette propriété 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 3.6.9. *Soient* $X_1, X_2, \ldots$ *des variables aléatoires discrètes indépendantes d'espérance* μ*, et* N *une variable aléatoire prenant ses valeurs dans* $\mathbb{N}^*$ *et indépendante des* X_i*. Alors, si* $S = X_1 + \cdots + X_N$ *(somme d'un nombre aléatoire de termes), on a*

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

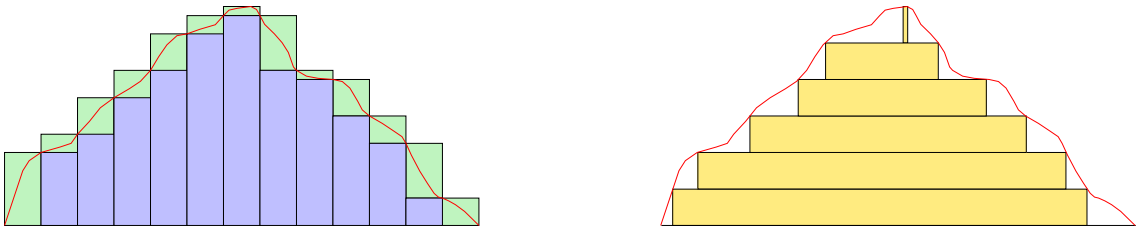


FIGURE 3.17: La construction des intégrales de Riemann et Lebesgue.

Par conséquent, $\mathbb{E}(S \mid N) = \mu N$*, et donc*

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

3.7 Détermination de la loi d'une variable aléatoire

Il existe plusieurs façons de déterminer la loi d'une variable aléatoire. : via le Corollaire 3.5.1, à travers sa fonction de répartition, ou encore par sa fonction caractéristique (voir Chapitre 4).

Comme exemple de l'emploi de la fonction de répartition, considérons une famille $X_1, \ldots, X_n$ de variables aléatoires i.i.d. de loi exponentielle de paramètre λ. On désire déterminer les lois de $\max_{i=1,\ldots,n} X_i$ et de $\min_{i=1,\ldots,n} X_i$. Manifestement, si $y \geq 0$,

$$\mathbb{P}(\max_{i=1,\ldots,n} X_i \leq y) = \mathbb{P}(X_1 \leq y, \ldots, X_n \leq y) = \mathbb{P}(X_1 \leq y)^n = (1 - e^{-\lambda y})^n.$$

La densité du maximum de n variables exponentielles de paramètre λ indépendantes est donc donnée par

$$\lambda n (1 - e^{-\lambda y})^{n-1} e^{-\lambda y}\, \mathbf{1}_{\{y \geq 0\}}.$$

De façon similaire,

$$\mathbb{P}(\min_{i=1,\ldots,n} X_i \leq y) = 1 - \mathbb{P}(X_1 \geq y, \ldots, X_n \geq y) = 1 - e^{-n\lambda y},$$

et donc la loi du minimum de n variables exponentielles de paramètre λ indépendantes est une loi exponentielle de paramètre λn.

3.8 Variables aléatoires générales

Le but de cette section est de décrire brièvement l'intégrale de Lebesgue, outil indispensable pour la formulation générale de la théorie des probabilités. Nous n'en ferons qu'un survol, car son étude fait partie du cours de théorie de la mesure (Analyse III).

3.8.1 Intégration au sens de Lebesgue

Considérons le cas d'une fonction f continue, positive, à support compact. L'intégrale de Riemann de f correspond précisément à l'aire comprise entre le graphe de f et l'axe des abscisses. La façon dont Riemann procède pour définir son intégrale est de partitionner le support de f en intervalles et de calculer les aires des sommes de Darboux correspondantes, obtenant ainsi une minoration et une majoration de l'aire (*cf.* Fig. 3.17). Lorsque la partition devient infiniment fine, on montre que les sommes de Darboux [11] supérieure et inférieure convergent vers une même limite, que l'on définit comme étant l'intégrale de Riemann de f.

L'idée de Lebesgue est de remplacer le découpage du support de f par un découpage de son image (*cf.* Fig. 3.17). À chaque intervalle $[\delta_i,\delta_{i+1}]$ de l'image, on associe l'ensemble $A_i = \{x : f(x) \geq \delta_{i+1}\}$. La contribution associée à l'intervalle $[\delta_i,\delta_{i+1}]$ est alors prise comme étant $(\delta_{i+1} - \delta_i)\ell(A_i)$, où $\ell(A_i)$ est la mesure de Lebesgue de A_i, qu'il convient d'interpréter comme étant sa « longueur ». Dans la limite d'une partition infiniment fine, on obtient à nouveau l'aire sous la courbe.

On voit donc qu'un premier problème avec cette approche est de donner un sens à la notion de longueur d'un sous-ensemble de $\mathbb{R}$, un problème tout à fait analogue à celui déjà rencontré d'associer une probabilité aux sous-ensembles de $\mathbb{R}$.

Mesures

Nous voulons pouvoir associer à tout sous-ensemble de $\mathbb{R}$ une « longueur ». Comme pour le problème analogue de la construction d'une probabilité, ceci n'est possible en général que pour un sous-ensemble strict des parties de $\mathbb{R}$, les boréliens de $\mathbb{R}$, $\mathcal{B}(\mathbb{R})$.

Définition 3.8.1. *Une mesure sur une tribu $\mathcal{F}$ est une application $\mu : \mathcal{F} \to \bar{\mathbb{R}}_+$ satisfaisant*

1. $\mu(\varnothing) = 0$;
2. *(σ-additivité) pour toute famille $A_1,A_2,\ldots$ d'éléments de $\mathcal{F}$ deux-à-deux disjoints, on a $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.*

Exemple 3.8.1. 1. *La mesure de Lebesgue ℓ est l'unique mesure sur $\mathcal{B}(\mathbb{R})$ telle que*

$$\ell([a,b]) = b - a, \qquad \forall -\infty < a \leq b < \infty.$$

2. *La masse de Dirac en $a \in \mathbb{R}$ est la mesure sur $\mathcal{B}(\mathbb{R})$ définie par*

$$\delta_a(A) = \mathbf{1}_{\{a \in A\}}.$$

Définition 3.8.2. *On dit qu'une propriété est vérifiée μ-presque partout si l'ensemble $\mathcal{N}$ des points où elle n'est pas vérifiée est de mesure nulle, $\mu(\mathcal{N}) = 0$. Si $\mu = \ell$, on dira simplement « presque partout » au lieu de « ℓ-presque partout ».*

Définition 3.8.3. *Étant donné deux mesures μ_1 et μ_2, on définit la mesure $\mu_1 + \mu_2$ par $(\mu_1 + \mu_2)(A) = \mu_1(A) + \mu_2(A)$, pour tout $A \in \mathcal{B}(\mathbb{R})$.*

11. Jean Gaston Darboux (1842, Nîmes - 1917, Paris), mathématicien français.

Fonctions mesurables

Comme déjà mentionné précédemment, une fonction mesurable $f : \Omega \to \mathbb{R}$ d'un espace probabilisable (dans le contexte général, on dit plutôt mesurable) $(\Omega,\mathcal{F})$ vers $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ est une fonction telle que $f^{-1}(A) \in \mathcal{F}$, pour tout borélien A. En fait, on peut montrer qu'il suffit que $f^{-1}\big((-\infty, x]\big) \in \mathcal{F}$, pour tout $x \in \mathbb{R}$. La classe des fonctions mesurables est très robuste.

Théorème 3.8.1. *1. Si $f,g : (\Omega,\mathcal{F}) \to (\mathbb{R},\mathcal{B}(\mathbb{R}))$ sont mesurables et $\lambda \in \mathbb{R}$, alors λf, $f + g$, et fg sont également mesurables.*

2. Si $f_n : (\Omega,\mathcal{F}) \to (\mathbb{R},\mathcal{B}(\mathbb{R}))$, $n \in \mathbb{N}$, sont mesurables, alors $\sup_{n\in\mathbb{N}} f_n$, $\inf_{n\in\mathbb{N}} f_n$, $\limsup_{n\to\infty} f_n$, $\liminf_{n\to\infty} f_n$ sont également mesurables, pourvu qu'elles prennent leurs valeurs dans $\mathbb{R}$.

3. Si $f_n : (\Omega,\mathcal{F}) \to (\mathbb{R},\mathcal{B}(\mathbb{R}))$, $n \in \mathbb{N}$, sont mesurables et convergent ponctuellement vers une fonction f à valeurs dans $\mathbb{R}$, alors f est mesurable.

4. Si $f : (\Omega,\mathcal{F}) \to (\mathbb{R},\mathcal{B}(\mathbb{R}))$ et $g : (\mathbb{R},\mathcal{B}(\mathbb{R})) \to (\mathbb{R},\mathcal{B}(\mathbb{R}))$ sont mesurables, alors $g \circ f$ est mesurable.

Construction de l'intégrale

Voyons à présent comment on construit réellement l'intégrale de Lebesgue. La construction se fait pour des classes de fonctions de plus en plus générales. Soit μ une mesure sur $\mathcal{B}(\mathbb{R})$.

Indicatrices. Soit $A \in \mathcal{B}(\mathbb{R})$. L'intégrale de la fonction indicatrice $\mathbf{1}_A$ est définie par

$$\int_{\mathbb{R}} \mathbf{1}_A(x)\,\mathrm{d}\mu(x) = \mu(A).$$

Fonctions étagées. On appelle fonction étagée une fonction qui s'écrit comme combinaison linéaire finie de fonctions indicatrices, $f(x) = \sum_{i=1}^n a_i \mathbf{1}_{A_i}(x)$, avec $A_i \in \mathcal{B}(\mathbb{R})$, $a_i \in \mathbb{R}$, $i = 1,\ldots,n$. Bien entendu, une fonction étagée admet de multiples représentations (on n'a pas exigé que les a_i soient tous distincts).

On montre facilement que toute fonction mesurable peut être obtenue comme limite croissante de fonctions étagées.

L'intégrale de la fonction étagée $\sum_{i=1}^n a_i \mathbf{1}_{A_i}(x)$ est définie par

$$\int_{\mathbb{R}} \Big(\sum_{i=1}^n a_i \mathbf{1}_{A_i}(x)\Big)\mathrm{d}\mu(x) = \sum_{i=1}^n a_i \mu(A_i).$$

On montre facilement que cette définition ne dépend pas de la représentation choisie pour la fonction étagée.

Fonctions mesurables positives. Soit $f : \mathbb{R} \to \bar{\mathbb{R}}$ une fonction mesurable positive. On définit

$$\int_{\mathbb{R}} f(x)\mathrm{d}\mu(x) = \sup\left\{\int_{\mathbb{R}} g(x)\,\mathrm{d}\mu(x) \;:\; g \leq f, g \text{ étagée}\right\}.$$

Observez que cette intégrale peut prendre la valeur ∞. Manifestement, cette définition coïncide avec la précédente si f est étagée.

Exemple 3.8.2.
1. *On vérifie facilement que lorsque $\mu = \ell$ et f est continue par morceaux, l'intégrale ainsi définie coïncide avec l'intégrale de Riemann (elles donnent toutes deux l'aire entre le graphe et l'axe des abscisses) (exercice).*
2. *Considérons le cas de la masse de Dirac en y, $\mu = \delta_y$. On vérifie facilement (exercice) que*

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}\delta_y(x) = f(y).$$

Fonctions mesurables. Pour étendre cette construction à une fonction mesurable quelconque f, on la décompose en sa partie positive et sa partie négative : $f = f^+ - f^-$, avec $f^+ = \max(f,0)$ et $f^- = \max(-f,0)$. Observez que f^+ et f^- sont toutes deux positives, et que $|f| = f^+ + f^-$. Si $\int_{\mathbb{R}} |f(x)|\,\mathrm{d}\mu(x) < \infty$, alors on dit que f est Lebesgue-intégrable. Dans ce cas, on a

$$\int_{\mathbb{R}} f^+(x)\,\mathrm{d}\mu(x) < \infty, \text{ et } \int_{\mathbb{R}} f^-(x)\,\mathrm{d}\mu(x) < \infty,$$

et on peut donc définir

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}\mu(x) = \int_{\mathbb{R}} f^+(x)\,\mathrm{d}\mu(x) - \int_{\mathbb{R}} f^-(x)\,\mathrm{d}\mu(x).$$

Fonctions à valeurs complexes. Les fonctions à valeurs complexes sont définis de la même façon, en traitant séparément leurs parties réelle et imaginaire,

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}\mu(x) = \int_{\mathbb{R}} \Re f(x)\,\mathrm{d}\mu(x) + \mathrm{i}\int_{\mathbb{R}} \Im f(x)\,\mathrm{d}\mu(x).$$

Notations. Nous emploierons les notations suivantes.
- $\int_{\mathbb{R}} f(x)\,\mathrm{d}\mu(x) = \int_{\mathbb{R}} f\mathrm{d}\mu$.
- Si $A \in \mathcal{F}$, $\int_A f\,\mathrm{d}\mu = \int_{\mathbb{R}} \mathbf{1}_A f\,\mathrm{d}\mu$.
- Si $\mu = \ell$, on écrit $\int_A f\mathrm{d}x$ au lieu de $\int_A f\mathrm{d}\ell$.

Propriétés de l'intégrale

Théorème 3.8.2. *L'intégrale de Lebesgue possède les propriétés suivantes.*

1. *(Linéarité) Soit f,g deux fonctions Lebesgue-intégrables, et a,b deux nombres réels. Alors, $\int_{\mathbb{R}}(af + bg)\,\mathrm{d}\mu = a\int_{\mathbb{R}} f\,\mathrm{d}\mu + b\int_{\mathbb{R}} g\,\mathrm{d}\mu$.*

2. *(Monotonicité) Si $f \leq g$ sont deux fonctions Lebesgue-intégrables, alors $\int_{\mathbb{R}} f \,\mathrm{d}\mu \leq \int_{\mathbb{R}} g \,\mathrm{d}\mu$.*
3. *(Linéarité en μ) Soient μ_1,μ_2 deux mesures. Pour tout a,$b \in \mathbb{R}^+$ et toute fonction f Lebesgue-intégrable par rapport à μ_1 et μ_2,*

$$\int_{\mathbb{R}} f \,\mathrm{d}(a\mu_1 + b\mu_2) = a \int_{\mathbb{R}} f \,\mathrm{d}\mu_1 + b \int_{\mathbb{R}} f \,\mathrm{d}\mu_2.$$

4. *Si f et g sont deux fonctions Lebesgue-intégrables égales μ-presque-partout, c'est-à-dire telles que $\mu(\{x \in \mathbb{R} : f(x) \neq g(x)\}) = 0$, alors $\int_{\mathbb{R}} f \,\mathrm{d}\mu = \int_{\mathbb{R}} g \,\mathrm{d}\mu$.*
5. *(Théorème de la convergence monotone) Soit $(f_k)_{k\in\mathbb{N}}$ une suite de fonctions mesurables positives telles que $f_k(x) \leq f_{k+1}(x)$, pour tout $k \in \mathbb{N}$ et $x \in \mathbb{R}$. Alors,*

$$\lim_{k\to\infty} \int_{\mathbb{R}} f_k \,\mathrm{d}\mu = \int \sup_k f_k \,\mathrm{d}\mu.$$

   (La valeur de ces intégrales peut être infinie).
6. *(Théorème de la convergence dominée) Soit $(f_k)_{k\in\mathbb{N}}$ une suite de fonctions mesurables convergeant ponctuellement vers une fonction f, et telles qu'il existe une fonction Lebesgue-intégrable g satisfaisant $|f_k| \leq g$, pour tout k. Alors f est Lebesgue-intégrable et*

$$\lim_{k\to\infty} \int_{\mathbb{R}} f_k \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu.$$

7. *(Lemme de Fatou) Soit $(f_k)_{k\in\mathbb{N}}$ une suite de fonctions mesurables positives. Alors*

$$\int_{\mathbb{R}} \liminf_{k\to\infty} f_k \,\mathrm{d}\mu \leq \liminf_{k\to\infty} \int f_k \,\mathrm{d}\mu.$$

Comparaison avec l'intégrale de Riemann

Discutons brièvement des avantages de l'intégrale de Lebesgue par rapport à l'intégrale de Riemann.

1. L'intégrale de Lebesgue permet d'intégrer beaucoup plus de fonctions que l'intégrale de Riemann. En effet, la classe des fonctions Lebesgue-intégrables contient toutes les fonctions dont la valeur absolue est Riemann-intégrable, mais permet aussi d'intégrer des fonctions beaucoup plus irrégulières. Par exemple, la fonction de Dirichlet $f(x) = \mathbf{1}_{\{x\in\mathbb{Q}\cap[0,1]\}}$ n'est pas Riemann-intégrable, car ses sommes de Darboux supérieure et inférieure sont égale à 1 et 0 respectivement, quelle que soit la finesse de la partition (les rationnels étant denses dans les réels), mais elle est Lebesgue-intégrable, d'intégrale donnée par $\ell(\mathbb{Q} \cap [0,1]) = 0$.
2. L'intégrale de Riemann se prête mal à l'interchange de limites et d'intégrales, opérations très fréquentes en mathématiques. L'intégrale de Lebesgue est beaucoup plus souple de ce point de vue, comme le montrent les Théorèmes de convergence monotone et dominée énoncés ci-dessus.

3.8.2 Espérance d'une variable aléatoire quelconque

Soit $X : \Omega \to \mathbb{R}$ une variable aléatoire sur un espace probabilisé $(\Omega,\mathcal{F},\mathbb{P})$. Alors, on définit l'intégrale

$$\int_\Omega X(\omega)\,\mathrm{d}\mathbb{P} = \int_{\mathbb{R}} x\,\mathrm{d}\mathbb{P}_X,$$

où $\mathbb{P}_X$ est la loi de X, c'est-à-dire la mesure (de probabilité) induite par $\mathbb{P}$ et X sur $(\mathbb{R},\mathcal{B}(\mathbb{R}))$. Cette dernière intégrale est bien de la forme étudiée plus haut.

Définition 3.8.4. *Soit $X : \Omega \to \mathbb{R}$ une variable aléatoire sur un espace probabilisé $(\Omega,\mathcal{F},\mathbb{P})$. L'espérance de X est définie par*

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

Il suit des propriétés de l'intégrale de Lebesgue que cette espérance possède toutes les propriétés discutées précédemment pour les variables aléatoires discrètes et à densité.

Exemple 3.8.3. *Voyons comment on peut retrouver les définitions antérieures données pour les variables aléatoires discrètes et à densité.*

– Variables aléatoires discrètes. *Une variable aléatoire discrète X est de la forme*

$$X = \sum_i a_i \mathbf{1}_{A_i},$$

où l'on peut supposer les a_i tous distincts et les A_i disjoints deux-à-deux. La loi $\mathbb{P}_X$ est caractérisée par les valeurs prises sur $X(\Omega) = \{a_i\}$, $f_X(a_i) = \mathbb{P}(A_i)$. Avec les notations de cette section, on peut donc écrire

$$\mathbb{P}_X(B) = \sum_{y\in X(\Omega)} f_X(y)\mathbf{1}_{\{y\in B\}} = \sum_{y\in X(\Omega)} f_X(y)\delta_y(B),$$

ce qui signifie que

$$\mathbb{P}_X = \sum_{y\in X(\Omega)} f_X(y)\delta_y.$$

Par conséquent,

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\,\mathrm{d}\mathbb{P}_X = \sum_{y\in X(\Omega)} f_X(y)\int_{-\infty}^{\infty} x\,\mathrm{d}\delta_y = \sum_{y\in X(\Omega)} y f_X(y).$$

– Variables aléatoires à densité. *Pour une variable aléatoire X avec densité f_X, on a*

$$\int_A \mathrm{d}\mathbb{P}_X = \mathbb{P}_X(A) = \int_A f_X\,\mathrm{d}x,$$

et, par conséquent, $\mathrm{d}\mathbb{P}_X(x) = f_X(x)\mathrm{d}x$. On a donc bien

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

3.8.3 Intégrales multiples

Exactement comme dans le cas des probabilités, étant donnés deux espaces $(E_1,\mathcal{E}_1,\mu_1)$ et $(E_2,\mathcal{E}_2,\mu_2)$, on peut construire leur produit $(E_1 \times E_2, \mathcal{E}_1 \times \mathcal{E}_2, \mu_1 \times \mu_2)$. Un résultat très utile dans ce contexte est le Théorème de Fubini, qui montre que l'ordre d'intégration d'une fonction réelle de plusieurs variables ne joue aucun rôle tant que toutes les quantités concernées existent.

Théorème 3.8.3 (Théorème de Fubini)**.** *Soit $(E_1,\mathcal{E}_1,\mu_1)$ et $(\mathbb{E}_2,\mathcal{E}_2,\mu_2)$ deux espaces mesurés, et $(E,\mathcal{E},\mu)$ l'espace produit correspondant. Une fonction mesurable $f(x) = f(x_1,x_2)$ mesurable sur $(E,\mathcal{E})$ peut être considérée comme une fonction $g_{x_1}(x_2)$ de x_2 pour chaque x_1 fixé, ou comme une fonction $h_{x_2}(x_1)$ de x_1 pour chaque x_2 fixé. Ces fonctions sont mesurables pour chaque x_1 et x_2. Si f est intégrable, alors ces deux fonctions sont également intégrables pour μ_1-presque tout x_1 et μ_2-presque tout x_2, respectivement. Leurs intégrales*

$$G(x_1) = \int_{E_2} g_{x_1}(x_2)\,\mathrm{d}\mu_2(x_2) \text{ et } H(x_2) = \int_{E_1} h_{x_2}(x_1)\,\mathrm{d}\mu_1(x_1)$$

sont mesurables, presque partout finies, et intégrables par rapport à μ_1 et μ_2, respectivement. Finalement,

$$\int_E f(x_1,x_2)\,\mathrm{d}\mu = \int_{E_1} G(x_1)\,\mathrm{d}\mu_1(x_1) = \int_{E_2} H(x_2)\,\mathrm{d}\mu_2(x_2).$$

Inversement, si f est mesurable et positive, et si soit G, soit H, qui sont toujours mesurables, a une intégrale finie, alors c'est également le cas de l'autre, et f est intégrable, et son intégrale est égale aux intégrales doubles correspondantes.

Ce théorème permet aisément de justifier tous les échanges d'intégrales effectués dans les sections précédentes (exercice).

Chapitre 4

Fonctions génératrices et fonctions caractéristiques

4.1 Fonctions génératrices

4.1.1 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.

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 différentiée 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 4.1.1. *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*

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

Exemple 4.1.1. *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*

$$\sum_{k=1}^{\infty} p(1 - p)^{k-1} s^k = \frac{ps}{1 - (1 - p)s}.$$

6. 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 $nb(N - b)(N - n)/(N^3 - N^2)$.

Puisque $G_X(1) = \mathbb{E}(1) = 1$, il suit que le rayon de convergence R de G_X est supérieur ou égal à 1. Le théorème d'Abel fournit une technique efficace pour calculer les moments de X ; par exemple ($G_X^{(k)}(1)$ étant un raccourci pour $\lim_{s\uparrow 1} G_X^{(k)}(s)$ 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 4.1.1. *Si $G_X(s)$ est la fonction génératrice de X, alors*

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

où les membres de droite doivent être compris comme des limites $s \uparrow 1$ lorsque le rayon de convergence de G_X est égal à 1.

Remarque 4.1.1. *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 $M_X(t)$.

Un type de question où les fonctions génératrices se révèlent particulièrement utiles est l'étude de sommes de variables aléatoires.

Proposition 4.1.2. *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 3.6.11, 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 4.1.2. *1.* 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*

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

Exemple 4.1.3. *1. Soient X et Y deux variables aléatoires indépendantes, suivant des lois binomiales de paramètres m et p, et n et p, respectivement. Alors*

$$\begin{aligned} G_{X+Y}(s) &= G_X(s)G_Y(s) = ((1-p)+ps)^m\,((1-p)+ps)^n \\ &= ((1-p)+ps)^{m+n} = \mathsf{binom}(m+n,p), \end{aligned}$$

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 $\lambda + \mu$:

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

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 4.1.3. *Soient X_1,X_2,... 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*

$$G_S(s) = G_N\big(G_X(s)\big).$$

Démonstration. En utilisant le Lemme 3.6.12,

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

□

Exemple 4.1.4. *En prenant la dérivée de G_S en 1, on retrouve immédiatement le résultat de l'Exemple 3.6.9.*

Exemple 4.1.5. *Une poule pond N oeufs, où N suit une loi de Poisson de paramètre λ. Chaque oeuf éclôt 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 ? Manifestement,*

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

Par conséquent,

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

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

Le théorème de continuité suivant, que l'on ne démontrera pas, montre que les fonctions génératrices permettent l'étude de la convergence de suites de variables aléatoires.

Théorème 4.1.1. *Soient $X, X_1, X_2, \ldots$ une suite de variables aléatoires à valeurs dans $\mathbb{N}$. Les deux propositions suivantes sont équivalentes :*

1. *$\lim_{n\to\infty} G_{X_n}(s) = G_X(s)$, pour tout $|s| \leq 1$;*
2. *la suite $(X_n)_{n\geq 1}$ converge en loi vers X, c'est-à-dire*

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

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

Exemple 4.1.6. *Soit $(X_n)_{n\geq 0}$ une suite de variables aléatoires de loi $\mathsf{binom}(n, p_n)$, 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 $\mathsf{poisson}(\lambda)$, on retrouve la loi des petits nombres.

4.1.2 Application aux processus de branchement

Dans cette sous-section, nous allons illustrer la puissance des fonctions génératrices dans l'étude d'une classe intéressante 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) : 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

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.

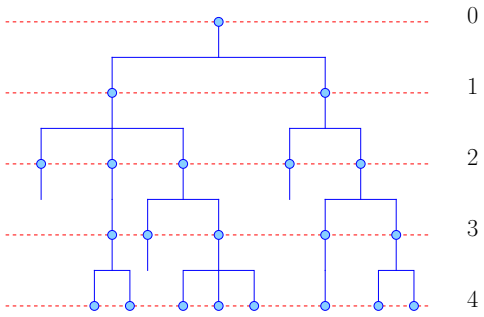


FIGURE 4.1: Une représentation du processus de branchement.

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 ;
- toutes les familles suivent la même loi, de fonction génératrice G.

Sous ces hypothèses, le processus est bien défini dès que la taille de la population initiale Z_0 est donnée ; on supposera ici que $Z_0 = 1$. 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. Soit $G_n(s) = \mathbb{E}(s^{Z_n})$ la fonction génératrice de Z_n.

Théorème 4.1.2. *$G_{m+n}(s) = G_m(G_n(s))$, et par conséquent, $G_n(s) = G(G(\ldots G(s)))$ est l'itéré n fois de G.*

Démonstration. Chacun des membres de la $(m+n)^{\text{ème}}$ génération possédant un unique ancêtre dans la génération m. On a donc

$$Z_{m+n} = X_1 + X_2 + \cdots + X_{Z_m}$$

où X_i représente le nombre de membres de la génération $m+n$ descendants du $i^{\text{ème}}$ individu de la génération m. Il s'agit donc d'une somme d'un nombre aléatoire de variables aléatoires

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

indépendantes, identiquement distribuées. Il suit donc de la Proposition 4.1.3 que

$$G_{m+n}(s) = G_m(G_{X_1}(s)) = G_m(G_n(s)),$$

puisque $G_{X_1}(s) = G_n(s)$. La seconde affirmation se démontre en itérant la première :

$$G_n(s) = G_1(G_{n-1}(s)) = G_1(G_1(G_{n-2}(s))) = \cdots = G_1(G_1(\ldots G_1(s))),$$

or G_1 est précisément ce que l'on avait appelé G. □

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 4.1.1. *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. En différentiant $G_n(s) = G(G_{n-1}(s))$ en $s = 1$, on obtient

$$\mathbb{E}(Z_n) = \mu\mathbb{E}(Z_{n-1}),$$

ce qui donne, après itération, $\mathbb{E}(Z_n) = \mu^n$. Similairement, en différentiant deux fois la relation $G_n(s) = G(G_{n-1}(s))$ en $s = 1$, on voit que

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

Par conséquent, la Proposition 4.1.1 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 ? L'événement « la population s'éteint après un temps fini » est donné par

$$\{\text{extinction}\} = \bigcup_{n\geq 1} \{Z_n = 0\}.$$

On observe que $\{Z_n = 0\} \subset \{Z_{n+1} = 0\}$; par conséquent, le Lemme 2.1.2 montre que la probabilité d'extinction est donnée par la limite $\lim_{n\to\infty} \mathbb{P}(Z_n = 0)$. Le théorème suivant montre que le destin de la population est étroitement lié à la taille moyenne des familles.

Théorème 4.1.3. *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 $\eta_n = \mathbb{P}(Z_n = 0)$. 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

$$\eta = G(\eta).$$

Vérifions à présent que si a est une racine positive de cette équation, alors $\eta \leq a$. En effet, puisque G est croissante sur $\mathbb{R}^+$,

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

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 $\mathbb{R}^+$; ceci est vrai, car

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

G est donc convexe (en fait, strictement convexe si $\mathbb{P}(Z_1 \geq 2) > 0$) et croissante sur [0,1], avec $G(1) = 1$. Un coup d'oeil sur la Figure 4.2 (et un argument plus analytique laissé en exercice), montre que les courbes $y = G(s)$ et $y = s$ ont généralement deux points d'intersection : en η et en 1. Lorsque $\mu < 1$, on a que $G'(1) = \mu < 1$, et donc les deux points d'intersection coïncident : $\eta = 1$. Lorsque $\mu > 1$, les deux points d'intersection sont distincts, et par conséquent $\eta < 1$ (η est toujours positif, et $\eta = 0$ si et seulement si $\mathbb{P}(Z_1 = 0) = 0$). Finalement, dans le cas $\mu = 1$, il faut considérer séparément le cas trivial où toutes les familles sont de taille 1, et donc évidemment $\eta = 0$, et celui où Z_1 possède une variance positive. Dans ce dernier cas, G est strictement convexe, ce qui implique que $G(s) > s$ pour tout $0 \leq s < 1$, et donc que $\eta = 1$. □

4.1.3 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.

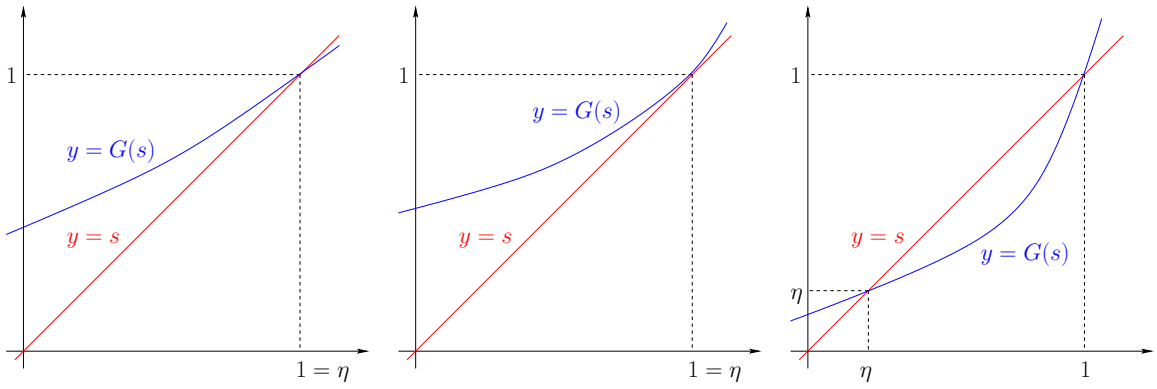


FIGURE 4.2: Solutions de l'équation $G(s) = s$. Gauche : $\mu < 1$. Milieu : $\mu = 1$ et $\mathrm{Var}(Z_1) > 0$. Droite : $\mu > 1$.

Définition 4.1.2. *La fonction génératrice conjointe du vecteur aléatoire* $\mathbf{X} = (X_1, \ldots, X_n)$ *prenant valeurs dans* $\mathbb{N}^n$ *est définie par*

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

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

Proposition 4.1.4. *$X_1, \ldots, X_n$, à valeurs dans $\mathbb{N}$, sont indépendantes si et seulement si*

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

pour tout $s_1, \ldots, s_n$.

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

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

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

$$\begin{aligned} &G_{(X_1,\ldots,X_n)}(s_1, \ldots, s_n) - G_{X_1}(s_1) \cdots G_{X_n}(s_n) = \\ &\qquad \sum_{x_1,\ldots,x_n} s_1^{x_1} \cdots s_n^{x_n} \big(\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) - \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n) \big). \end{aligned}$$

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

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

pour tout $x_1, \ldots, x_n$ (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. □

Remarque 4.1.2. *Dans ce qui précède, 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 Fonctions caractéristiques

Dans cette section, 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.

4.2.1 Définition et propriétés élémentaires

Définition 4.2.1. *La fonction caractéristique associée à une variable aléatoire* X *est la fonction* $\phi_X : \mathbb{R} \to \mathbb{C}$ *définie par*

$$\phi_X(t) = \mathbb{E}(e^{\mathrm{i}tX}).$$

Remarque 4.2.1. *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,*

$$\phi_X(t) = \mathbb{E}(\cos(tX)) + \mathrm{i}\mathbb{E}(\sin(tX)).$$

Théorème 4.2.1. *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 $|\phi(t)| \leq \mathbb{E}(|e^{\mathrm{i}tX}|) = 1$.

On a également

$$|\phi(t+s) - \phi(t)| = |\mathbb{E}(e^{\mathrm{i}(t+s)X} - e^{\mathrm{i}tX})| \leq \mathbb{E}(|e^{\mathrm{i}tX}(e^{\mathrm{i}sX} - 1)|) = \mathbb{E}(|e^{\mathrm{i}sX} - 1|).$$

Soit $Y(s) = |e^{\mathrm{i}sX} - 1|$; manifestement $0 \leq Y \leq 2$ et $\lim_{s\to 0} Y(s) = 0$. Par conséquent, le Théorème de convergence dominée (Théorème 3.8.2) 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 4.2.1. *Si X possède un moment d'ordre k, alors*

$$\phi_X(t) = \sum_{j=0}^{k} \frac{\mathbb{E}(X^j)}{j!} (\mathrm{i}t)^j + o(t^k),$$

lorsque $t \to 0$, et donc, en particulier, $\phi_X^{(k)}(0) = \mathrm{i}^k \mathbb{E}(X^k)$.

Démonstration. Cela suit du théorème de Taylor. La seule chose à vérifier est que si X admet un moment d'ordre n, alors ϕ_X est de classe $\mathcal{C}^n$ et $\phi_X^{(n)}(t) = \mathrm{i}^n \mathbb{E}(X^n e^{\mathrm{i}tX})$.

On procède par récurrence. Soit $k \leq n$, et supposons le résultat vérifié pour $1, \ldots, k-1$. On pose $F(t) = (\mathrm{i}X)^{k-1} e^{\mathrm{i}tX}$. On a alors $F'(t) = (\mathrm{i}X)^k e^{\mathrm{i}tX}$, et $|F'(t)| = |X|^k$. Par conséquent, il suit du Théorème de convergence dominée que

$$\begin{aligned}\phi_X^{(k)}(t) &= \lim_{\epsilon\to 0} \frac{\phi_X^{(k-1)}(t+\epsilon) - \phi_X^{(k-1)}(t)}{\epsilon} = \lim_{\epsilon\to 0} \mathbb{E}\left(\frac{F(t+\epsilon) - F(t)}{\epsilon}\right) \\ &= \mathbb{E}\left(\lim_{\epsilon\to 0} \frac{F(t+\epsilon) - F(t)}{\epsilon}\right) = \mathbb{E}(F'(t)).\end{aligned}$$

De plus, l'application $x \mapsto x^k e^{\mathrm{i}tx}$ étant continue et de module borné par $|x|^k$, il suit également du Théorème de convergence dominée que

$$\lim_{t\to t_0} \mathbb{E}(X^k e^{\mathrm{i}tX}) = \mathbb{E}(\lim_{t\to t_0} X^k e^{\mathrm{i}tX}) = \mathbb{E}(X^k e^{\mathrm{i}t_0 X}),$$

ce qui montre que $\phi_X^{(k)}$ est continue. □

Remarque 4.2.2. *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 4.2.1. *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 4.2.2. *Si a,$b \in \mathbb{R}$ et $Y = aX + b$, alors*

$$\phi_Y(t) = e^{\mathrm{i}tb}\phi_X(at).$$

Démonstration.

$$\phi_Y(t) = \mathbb{E}(e^{\mathrm{i}t(aX+b)}) = e^{\mathrm{i}tb}\mathbb{E}(e^{\mathrm{i}(at)X}) = e^{\mathrm{i}tb}\phi_X(at).$$

□

On peut également définir une notion de fonction caractéristique conjointe pour une famille de variables aléatoires.

Définition 4.2.2. *La fonction caractéristique conjointe du vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ est définie par*

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}(e^{\mathrm{i}\langle \mathbf{t}, \mathbf{X}\rangle}),$$

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 4.2.2. *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 4.2.1. La réciproque suit (de la version à n variables) du Théorème d'inversion énoncé plus bas. □

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 4.2.3 (Théorème d'inversion)**.** *Soit X une variable aléatoire de fonction de répartition F_X et de fonction caractéristique ϕ_X. Alors,*

$$F_X(b) - F_X(a) = \lim_{T\to\infty} \int_{-T}^{T} \frac{e^{-\mathrm{i}at} - e^{-\mathrm{i}bt}}{2\mathrm{i}\pi t} \phi_X(t)\,\mathrm{d}t.$$

en chaque point de continuité de F_X.

Démonstration. On écrit simplement F et ϕ. On a

$$\begin{aligned}
\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \mathrm{d}t\, \frac{e^{-\mathrm{i}at} - e^{-\mathrm{i}bt}}{\mathrm{i}t} \int e^{\mathrm{i}tx}\,\mathrm{d}\mathbb{P}_X & \\
&= \lim_{T\to\infty} \frac{1}{2\pi} \int \mathrm{d}\mathbb{P}_X \int_{-T}^{T} \frac{e^{\mathrm{i}t(x-a)} - e^{\mathrm{i}t(x-b)}}{\mathrm{i}t}\,\mathrm{d}t \\
&= \lim_{T\to\infty} \frac{1}{2\pi} \int \mathrm{d}\mathbb{P}_X \int_{-T}^{T} \frac{\sin\left(t(x-a)\right) - \sin\left(t(x-b)\right)}{t}\,\mathrm{d}t \\
&= \tfrac{1}{2} \int \left\{\mathrm{signe}(x-a) - \mathrm{signe}(x-b)\right\}\,\mathrm{d}\mathbb{P}_X \\
&= F(b) - F(a),
\end{aligned}$$

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 $\sup_{T,z} |u(T,z)| \le C$ et [3]

$$\lim_{T\to\infty} u(T,z) = \begin{cases} \pi/2 & \text{si } z > 0 \\ -\pi/2 & \text{si } z < 0 \\ 0 & \text{si } z = 0. \end{cases}$$

□

Corollaire 4.2.1. *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

$$F_X(b) - F_X(a) = F_Y(b) - F_Y(a),$$

3. Poser, pour $n \ge 1$, $u_n = \int_0^{\pi/2} \sin\big((2n-1)x\big)/\sin(x)\,\mathrm{d}x$ et $v_n = \int_0^{\pi/2} \sin(2nx)/x\,\mathrm{d}x$. Montrer que : (i) $u_{n+1} = u_n$, $\forall n \ge 1$ (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) $\lim_{T\to\infty} u(T,1) = \lim_{n\to\infty} v_n = \pi/2$.

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 de répartition 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 5).

Définition 4.2.3. *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 4.2.4 (Théorème de continuité de Lévy[4])**.** *Soient F_1,F_2,... une suite de fonctions de répartition, et ϕ_1,ϕ_2,... 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. □

4.2.2 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

Si X suit une loi exponentielle de paramètre λ, alors le changement de variable $y = (\lambda - \mathrm{i}t)x$ donne

$$\phi_X(t) = \lambda \int_0^\infty e^{-\lambda x + \mathrm{i}tx}\,\mathrm{d}x = \frac{\lambda}{-\lambda + \mathrm{i}t}\int_0^\infty e^{-y}\mathrm{d}y = \frac{\lambda}{\lambda - \mathrm{i}t}.$$

4. Paul Pierre Lévy (1886, Paris – 1971, Paris), mathématicien français.

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 4.2.2 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

$$\phi_Y(t) = e^{-\frac{1}{2}\sigma^2t^2+\mathrm{i}\mu t}.$$

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 $\mathbf{C} = \mathbf{U}^{\mathrm{t}}\mathbf{D}\mathbf{U}$. 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

$$\phi_{\mathbf{Z}}(\mathbf{U}\mathbf{t}) = e^{-\frac{1}{2}\langle \mathbf{U}\mathbf{t},\mathbf{D}\mathbf{U}\mathbf{t} \rangle} = e^{-\frac{1}{2}\langle \mathbf{t},\mathbf{U}^{\mathrm{t}}\mathbf{D}\mathbf{U}\mathbf{t} \rangle} = e^{-\frac{1}{2}\langle \mathbf{t},\mathbf{C}\mathbf{t} \rangle}.$$

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 $\mathrm{Cov}(\mathbf{X},\mathbf{X}) = \mathbf{C}$.

De plus, 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. Ceci démontre le Théorème 3.6.3.

Chapitre 5

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.

5.1 Un point technique

Les 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 puisse être définie n'est pas évidente, et nous allons brièvement discuter cette question à présent.

Soit $(\Omega,\mathcal{F},\mathbb{P})$ un espace probabilisé, 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 :

$$\lim_{x_{n+1}\to\infty} F_{(t_1,\ldots,t_n,t_{n+1})}(x_1,\ldots,x_n,x_{n+1}) = F_{(t_1,\ldots,t_n)}(x_1,\ldots,x_n), \tag{5.1}$$

$$F_{\pi\mathbf{t}}(\pi\mathbf{x}) = F_{\mathbf{t}}(\mathbf{x}), \tag{5.2}$$

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 5.1.1 (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 2.5 pour construire un espace probabilisé 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 $(X_t)_{t\in T}$. □

5.2 Quelques outils

5.2.1 Les lemmes de Borel-Cantelli

Soit $A_1, A_2 \ldots$, une suite infinie d'événements sur un espace probabilisé $(\Omega,\mathcal{F},\mathbb{P})$. 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 cet événement est réalisé.

Théorème 5.2.1 (Lemmes de Borel-Cantelli). *Soit $A_1, A_2 \ldots$, une suite infinie d'événements sur un espace probabilisé $(\Omega,\mathcal{F},\mathbb{P})$, 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 que $A_1, A_2, \ldots$ sont des événements indépendants.*

Démonstration. *1.* Pour tout n,

$$A \subseteq \bigcup_{m=n}^{\infty} A_m,$$

et par conséquent

$$\mathbb{P}(A) \leq \sum_{m=n}^{\infty} \mathbb{P}(A_m),$$

et le membre de droite tend vers 0 lorsque n tend vers l'infini.

2. On vérifie aisément que

$$A^{\mathrm{c}} = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m^{\mathrm{c}}.$$

Cependant,

$$\begin{aligned}
\mathbb{P}\Big(\bigcap_{m=n}^{\infty} A_m^{\mathrm{c}}\Big) &= \lim_{N\to\infty} \mathbb{P}\Big(\bigcap_{m=n}^{N} A_m^{\mathrm{c}}\Big) && \text{(Lemme 2.1.2)} \\
&= \lim_{N\to\infty} \prod_{m=n}^{N} (1 - \mathbb{P}(A_m)) && \text{(indépendance)} \\
&\leq \lim_{N\to\infty} \prod_{m=n}^{N} \exp\big(-\mathbb{P}(A_m)\big) && (1 - x \leq e^{-x}) \\
&= \lim_{N\to\infty} \exp\Big(-\sum_{m=n}^{N} \mathbb{P}(A_m)\Big) \\
&= 0
\end{aligned}$$

dès que $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$. Manifestement $(\bigcap_{m=n}^{\infty} A_m^{\mathrm{c}})_{n\geq 1}$ est une suite croissante d'événements ; il suit donc du Lemme 2.1.2 que

$$\mathbb{P}(A^{\mathrm{c}}) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcap_{m=n}^{\infty} A_m^{\mathrm{c}}\Big) = 0.$$

□

Remarque 5.2.1. *Sans l'hypothèse d'indépendance, la seconde partie peut être fausse : il suffit de considérer la suite d'événements $A_k = B$, pour tout $k \geq 1$. 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).*

5.2.2 Quelques inégalités

Supposons que $X_1, \ldots, X_n$ 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. Avant de l'énoncer, démontrons une inégalité extrêmement utile.

Théorème 5.2.2. *Soit $\varphi : \mathbb{R} \to [0,\infty)$. Alors*

$$\mathbb{P}(\varphi(X) \geq a) \leq \frac{\mathbb{E}(\varphi(X))}{a}, \qquad \forall a > 0.$$

Démonstration. Soit $A = \{\varphi(X) \geq a\}$. Trivialement,

$$\varphi(X) \geq a\mathbf{1}_A,$$

et donc, en prenant l'espérance,

$$\mathbb{E}(\varphi(X)) \geq a\mathbb{E}(\mathbf{1}_A) = a\mathbb{P}(A).$$

□

Corollaire 5.2.1. *Soit X une variable aléatoire.*

1. *(Inégalité de Markov[1]) 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[2]-Tchebychev[3]) Si X possède une variance, alors*

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

3. *(Inégalité de Chernoff[4]) 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 $a \in \mathbb{R}$,

$$\mathbb{P}(X \geq a) \leq \exp\big(-\sup_{t \geq 0}\{ta - H(t)\}\big).$$

1. Andrei Andreevitch Markov (1856, Riazan - 1922, Petrograd), mathématicien russe.
2. Irénée-Jules Bienaymé (1796, Paris - 1878, Paris), probabiliste et statisticien français.
3. Pafnouti Lvovitch Tchebychev (1821, Okatovo - 1894, Saint-Petersbourg), mathématicien russe. Son nom est aussi translittéré comme Chebyshov, Chebyshev, ou Tschebyscheff.
4. Herman Chernoff (1923, New York -), mathématicien et statisticien américain.

Démonstration. *1.* Il suffit de prendre $\varphi(x) = |x|$ dans le Théorème 5.2.2.

2. Par le Théorème 5.2.2, avec $\varphi(x) = x^2$, appliqué à la variable aléatoire $Y = X - \mathbb{E}(X)$, on a

$$\mathbb{P}\big(|X - \mathbb{E}(X)| \geq a\big) = \mathbb{P}\big(Y^2 \geq a^2\big) \leq \frac{\mathbb{E}\big(Y^2\big)}{a^2} = \frac{\mathrm{Var}(X)}{a^2}.$$

3. En appliquant le Théorème 5.2.2 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 $t \geq 0$. □

Remarque 5.2.2. *Soit Y une variable aléatoire possédant une variance finie. L'inégalité de Bienaymé-Tchebychev montre que la probabilité qu'une variable aléatoire s'éloigne de son espérance d'une distance grande par rapport à son écart-type est très faible. En d'autres termes, la variable aléatoire Y « est concentrée dans un intervalle d'ordre $\sigma(Y)$ autour de son espérance ». Nous verrons des formes plus fortes et plus précises de cette observation plus tard.*

5.3 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 5.3.1. *Soient $X_1, X_2, \ldots$ et X des variables aléatoires sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$. 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 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 5.3.1. *Lorsque $X_n \xrightarrow{1} X$, on parle de* convergence en moyenne.
Lorsque $X_n \xrightarrow{2} X$, on parle de convergence en moyenne quadratique.

Notons le résultat suivant, qui montre quelles sont les implications entre ces différents modes de convergence.

Théorème 5.3.1. *Les implications suivantes sont vérifiées :*

$$\begin{array}{ccc} (X_n \xrightarrow{\text{p.s.}} X) & & \\ \Downarrow & & \\ (X_n \xrightarrow{\mathbb{P}} X) & \Rightarrow & (X_n \xrightarrow{\mathcal{L}_{\mathbb{P}}} X) \\ \Uparrow & & \\ (X_n \xrightarrow{s} X) & & \\ \Uparrow & & \\ (X_n \xrightarrow{r} X) & & \end{array}$$

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 5.3.2. *1. Si $X_n \xrightarrow{\mathcal{L}_{\mathbb{P}}} c$, avec c une constante, alors $X_n \xrightarrow{\mathbb{P}} c$.*

2. Si $X_n \xrightarrow{\mathbb{P}} X$ et $\exists k$ tel que $\mathbb{P}(|X_n| \leq k) = 1$, pour tout n, alors $X_n \xrightarrow{r} X$, pour tout $r \geq 1$.

3. Si $\sum_n \mathbb{P}(|X_n - X| > \epsilon) < \infty$, pour tout $\epsilon > 0$, alors $X_n \xrightarrow{\text{p.s.}} X$.

Démonstration. 1. $\mathbb{P}(|X_n - c| > \epsilon) = \mathbb{P}(X_n < c - \epsilon) + \mathbb{P}(X_n > c + \epsilon) \to 0$, si $X_n \xrightarrow{\mathcal{L}_{\mathbb{P}}} c$.

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

$$\mathbb{E}\big(|X_n - X|^r\big) \leq \epsilon^r + (2k)^r \mathbb{P}(A_n(\epsilon)) \to \epsilon^r,$$

lorsque $n \to \infty$. La conclusion suit puisque ϵ était arbitraire.

3. L'affirmation est une conséquence du lemme suivant.

Lemme 5.3.1. *Soit $A_n(\epsilon) = \{|X_n - X| > \epsilon\}$ et $B_m(\epsilon) = \bigcup_{n\geq m} A_n(\epsilon)$. Alors $X_n \xrightarrow{\text{p.s.}} X$ si et seulement si $\lim_{m\to\infty} \mathbb{P}(B_m(\epsilon)) = 0, \forall \epsilon > 0$.*

Démonstration du Lemme 5.3.1. Soit $C = \{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$, et

$$A(\epsilon) = \{\omega : \omega \in A_n(\epsilon) \text{ pour une infinité de valeurs de } n\}.$$

Manifestement, $X_n(\omega) \to X(\omega)$ si et seulement si $\omega \notin A(\epsilon)$, pour tout $\epsilon > 0$. Par conséquent, $\mathbb{P}(C) = 1$ implique que $\mathbb{P}(A(\epsilon)) = 0$, pour tout $\epsilon > 0$. D'autre part, si $\mathbb{P}(A(\epsilon)) = 0$ pour tout $\epsilon > 0$, alors

$$\mathbb{P}(C^c) = \mathbb{P}(\bigcup_{\epsilon>0} A(\epsilon)) = \mathbb{P}(\bigcup_{m\geq 1} A(1/m)) \leq \sum_{m\geq 1} \mathbb{P}(A(1/m)) = 0,$$

puisque $\epsilon \geq \epsilon' \implies A(\epsilon) \subseteq A(\epsilon')$. Ceci montre que $\mathbb{P}(C) = 1$ si et seulement si $\mathbb{P}(A(\epsilon)) = 0$ pour tout $\epsilon > 0$.

La première affirmation suit puisque $A(\epsilon) = \bigcap_m B_m(\epsilon)$ et donc $\mathbb{P}(A(\epsilon)) = 0$ si et seulement si $\lim_{m\to\infty} \mathbb{P}(B_m(\epsilon)) = 0$. □

Pour démontrer *3.*, il suffit alors d'observer que

$$\mathbb{P}(B_m(\epsilon)) \leq \sum_{n=m}^{\infty} \mathbb{P}(A_n(\epsilon)),$$

et donc $\lim_{m\to\infty} \mathbb{P}(B_m(\epsilon)) = 0$ dès que $\sum_n \mathbb{P}(A_n(\epsilon)) < \infty$, par le premier lemme de Borel-Cantelli, *cf.* Théorème 5.2.1. □

5.4 La loi des grands nombres

5.4.1 Loi faible des grands nombres

Définition 5.4.1. *Soient X_1,X_2,…, X_n une famille de variables aléatoires. Leur moyenne empirique est la variable aléatoire*

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

La Figure 5.1 montre le comportement d'une réalisation de la moyenne empirique d'une famille de variables aléatoires de loi $\mathsf{U}(-1,1)$ (pour n allant de 1 à 10000). On voit que la moyenne empirique semble converger vers son espérance (nulle dans ce cas). Que cela a bien lieu est le contenu de la loi des grands nombres.

Théorème 5.4.1 (Loi faible des grands nombres)**.** *Pour tout entier $n \geq 1$, on se donne des variables $X_1,\ldots,X_n$, non-corrélées, de même espérance μ et de même variance σ^2. Alors la moyenne empirique S_n converge en moyenne quadratique vers μ, lorsque $n \to \infty$:*

$$\mathbb{E}\big(|S_n - \mu|^2\big) = \frac{\sigma^2}{n}.$$

En particulier,

$$\mathbb{P}(|S_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2 n} \to 0, \qquad n \to \infty$$

pour tout $\epsilon > 0$.

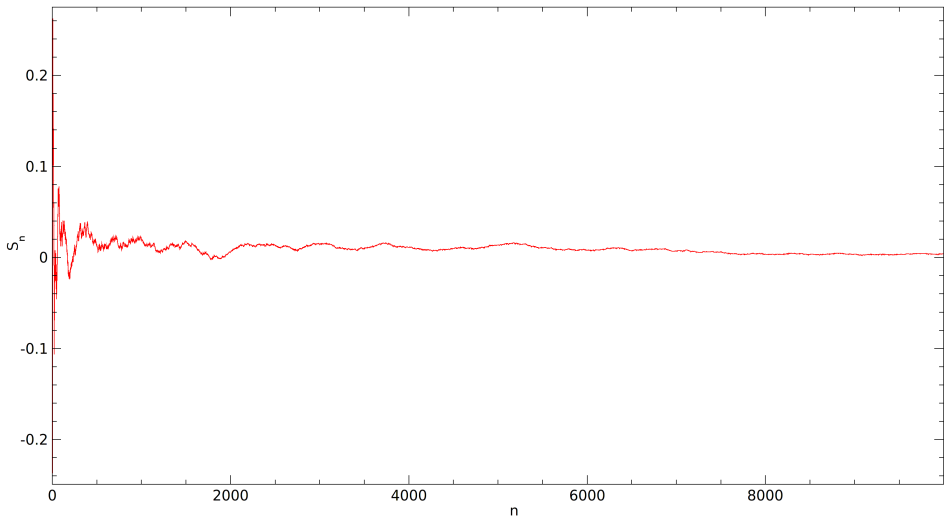


FIGURE 5.1: La moyenne empirique d'une famille de variables aléatoires de loi $\mathsf{U}(-1,1)$ (n allant de 1 à 10000).

Démonstration. On a

$$\mathbb{E}\big(|S_n-\mu|^2\big)=\mathbb{E}\big((S_n-\mathbb{E}(S_n))^2\big)=\mathrm{Var}(S_n)=\frac{\sigma^2}{n}.$$

La seconde affirmation suit alors de l'inégalité de Bienaymé-Tchebychev,

$$\mathbb{P}(|S_n-\mu|\geq\epsilon)\leq\frac{\mathrm{Var}(S_n)}{\epsilon^2}.$$

□

Exemple 5.4.1. *On effectue* 10000 *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* $\mathbb{P}(X_1=1)=\mathbb{P}(X_1=-1)=\frac{1}{2}$ *(au lieu de* 0 *et* 1*). La loi faible des grands nombres énoncée ci-dessus affirme que* $S_n\in[-\epsilon,\epsilon]$ *avec grande probabilité lorsque* n *est suffisamment grand. L'estimée dans la preuve du théorème nous donne*

$$\mathbb{P}\big(|S_n|\geq\epsilon\big)\leq\frac{1}{n\epsilon^2}.$$

Par exemple, pour 10000 *jets et* $\epsilon=0{,}1$*, on a*

$$\mathbb{P}(|S_{10000}|\geq 0{,}1)\leq\frac{1}{100}.$$

Notez que ce n'est qu'une borne supérieure sur cette probabilité. On verra plus bas qu'elle est en fait très mauvaise dans le cas présent.

Dans le cas où les variables aléatoires sont indépendantes, et pas seulement non-corrélées, la version suivante de la loi des grands nombres montre qu'il suffit d'avoir une espérance finie.

Théorème 5.4.2 (Loi faible des grands nombres)**.** *Soient $X_1,X_2,\dots$ des variables aléatoires indépendantes de même espérance μ. Alors $S_n \xrightarrow{\mathcal{L}_{\mathbb{P}}} \mu$:*

$$\lim_{n\to\infty} F_{S_n}(x) = \begin{cases} 1 & si\ x > \mu, \\ 0 & si\ x < \mu. \end{cases}$$

Démonstration. Il suit du Lemme 4.2.1 que

$$\phi_X(t) = 1 + \mathrm{i}t\mu + o(t).$$

Par conséquent, la Proposition 4.2.1 et le Lemme 4.2.2 impliquent que la fonction caractéristique de la variable aléatoire $S_n = \frac{1}{n}\sum_{i=1}^n X_i$ satisfait

$$\phi_{S_n}(t) = \big(\phi_X(t/n)\big)^n = \big(1 + \frac{\mathrm{i}t\mu}{n} + o(\frac{t}{n})\big)^n \to e^{\mathrm{i}t\mu},$$

lorsque $n \to \infty$. Comme $e^{\mathrm{i}t\mu}$ est la fonction caractéristique de la variable aléatoire constante μ, le résultat suit du Théorème de continuité 4.2.4. □

Remarque 5.4.1. *On ne peut pas affaiblir davantage les hypothèses : 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*

$$\phi_{S_n}(t) = \big(\phi_X(t/n)\big)^n = e^{-|t|},$$

ce qui montre que S_n suit également une loi de Cauchy, et ne peut donc pas converger vers une constante ! La Figure 5.2 montre le comportement d'une réalisation de S_n pour n allant de 1 à 10000.

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 diffère 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 triplet $(\Omega,\mathcal{F},\mathbb{P})$, que l'on répète N fois, de façon indépendante, obtenant une suite de résultats $(\omega_1,\omega_2,\dots,\omega_N)$. Alors, pour tout événement $A \in \mathcal{F}$, les variables aléatoires $Y_k(\omega_1,\dots,\omega_N) = \mathbf{1}_A(\omega_k)$ sont i.i.d., avec $\mathbb{E}(Y_k) = \mathbb{P}(A)$. Par conséquent, si l'on note $N(A) = \#\{1 \le k \le N : \omega_k \in A\}$ le nombre d'expériences lors desquelles l'événement A est réalisé, on a

$$\frac{N(A)}{N} = \frac{1}{N}\sum_{k=1}^N Y_k \xrightarrow{\mathcal{L}_{\mathbb{P}}} \mathbb{P}(A),$$

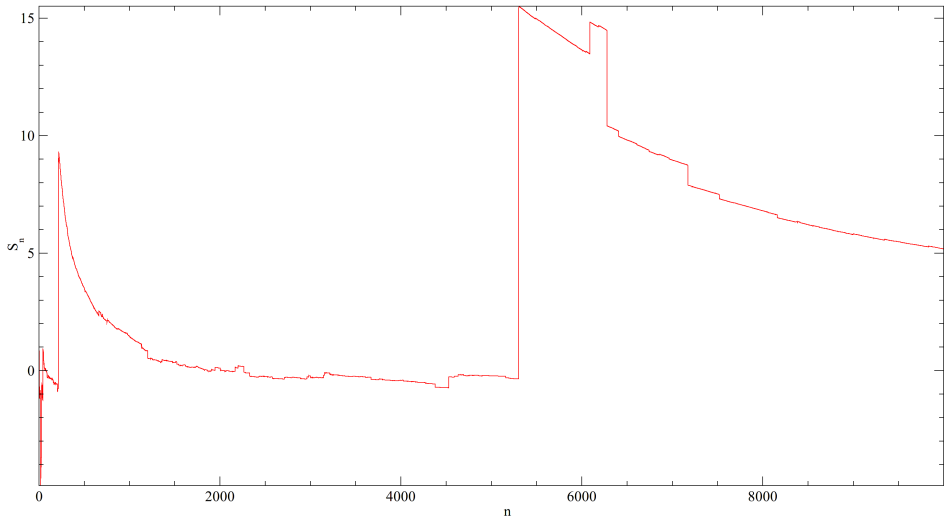


FIGURE 5.2: La moyenne empirique d'une famille de variables aléatoires suivant une loi de Cauchy (n allant de 1 à 10000).

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

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 plus précises de la vitesse de convergence.

Exemple 5.4.2. *Pour illustrer ce point, reprenons l'exemple des* 10000 *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*

$$\mathbb{P}\big(S_n \geq x\big) \leq \inf_{t\geq 0} e^{(n\log\cosh(t/n) - tx)}.$$

Un petit calcul[5] *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*

$$I(x) = -f(t^*) = \frac{1}{2}\big\{(1+x)\log(1+x) + (1-x)\log(1-x)\big\},$$

et en utilisant la symétrie pour estimer $\mathbb{P}\big(S_n \leq -x\big)$*, on a finalement*

$$\mathbb{P}\big(|S_n| \geq x\big) \leq 2\,e^{-nI(x)}. \tag{5.3}$$

5. Se rappeler que $\cosh(u) = 1/\sqrt{1-\tanh^2(u)}$ et que $\operatorname{argtanh}(u) = \frac{1}{2}\log\{(1+x)/(1-x)\}$.

En posant $n = 10000$ *et* $\epsilon = 0{,}1$, *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 5.4.1.

Un résultat du type (5.3) est ce qu'on appelle une estimée de grande déviation. *La théorie des grandes déviations est un domaine important de la théorie des probabilités, et a été récemment récompensée du prix Abel par l'intermédiaire de l'un de ses principaux artisans, S.R.S. Varadhan*[6].

5.4.2 La loi forte des grands nombres

Si la loi faible des grands nombres montre que pour tout grand n *fixé*, S_n est typiquement proche de μ, elle n'affirme *pas* que S_n 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 $|S_{n_k} - \mu| > \epsilon$, pour tout $k \geq 1$. La loi forte des grands nombres montre que ceci a probabilité nulle : pour tout $\epsilon > 0$, avec probabilité 1, seul un nombre fini des événements

$$\big|S_n - \mu\big| > \epsilon$$

sont réalisés.

Théorème 5.4.3. *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,* $\mu = \mathbb{E}(X_1)$.

Démonstration. Nous nous contenterons de démontrer la convergence 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 5.2.2 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

$$\mathbb{E}(S_n^4) = n^{-3}\mathbb{E}(X_1^4) + 12n^{-3}(n-1)\mathbb{E}(X_1^2)\mathbb{E}(X_2^2),$$

et il existe donc une constante C telle que

$$\sum_{n \geq 1} \mathbb{P}(|S_n| > \epsilon) \leq \sum_{n \geq 1} \frac{C}{n^2} < \infty.$$

6. Sathamangalam Ranga Iyengar Srinivasa Varadhan (1940, Chennai -), probabiliste américain d'origine indienne. Lauréat du prix Abel en 2007.

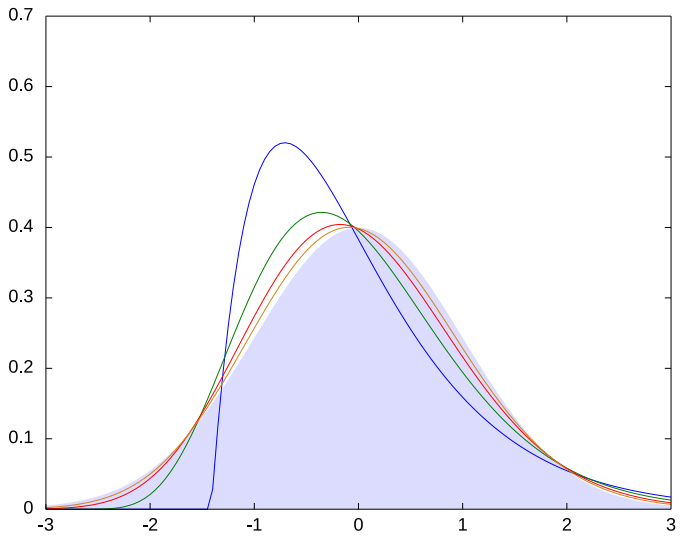


FIGURE 5.3: Convergence vers une loi normale pour une suite de variables aléatoires X_i de loi exp(1). Les courbes correspondent aux densités des variables $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - 1)$, pour n = 2,8,32,128. La densité de la loi $\mathcal{N}(0,1)$ est aussi tracée.

Ceci implique, via le Lemme de Borel-Cantelli [7] (Théorème 5.2.1), que, $\mathbb{P}$-presque sûrement, $|S_n| \leq \epsilon$ pour tout n suffisamment grand. La convergence presque-sûre suit alors du point *3.* du Théorème 5.3.2. □

À présent que l'on sait que la moyenne empirique d'une suite de variables aléatoirs 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 $\sigma(S_n - \mu)\sqrt{n/\sigma^2}$ 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 !

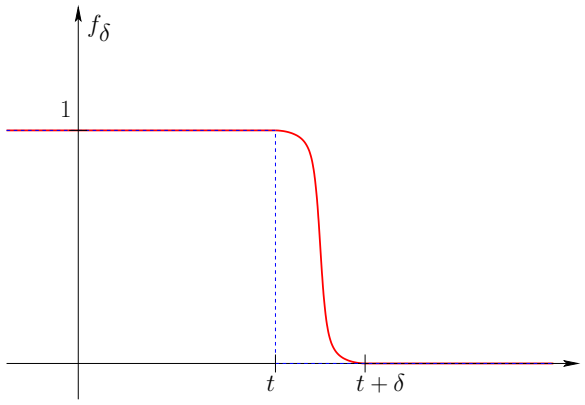


FIGURE 5.4: La fonction f_δ (en rouge) et l'indicatrice qu'elle approxime (traitillé).

5.5 Le Théorème Central Limite

Théorème 5.5.1 (Théorème Central Limite)**.** *Soit X_1,X_2,... 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*

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \Big|\mathbb{P}\Big(\frac{1}{\sqrt{n\sigma^2}} \sum_{k=1}^{n} (X_k - \mu) \le x\Big) - \Phi(x)\Big| = 0.$$

Si, de plus, $\mathbb{E}(|X_1 - \mathbb{E}(X_1)|^3) < \infty$, alors

$$\sup_{x\in\mathbb{R}} \Big|\mathbb{P}\Big(\frac{1}{\sqrt{n\sigma^2}} \sum_{k=1}^{n} (X_k - \mu) \le x\Big) - \Phi(x)\Big| \le C \frac{\mathbb{E}(|X_1 - \mathbb{E}(X_1)|^3)}{\sigma^3 \sqrt{n}},$$

pour une certaine constante universelle $C \le 0{,}7655$.

Remarque 5.5.1. *L'estimée explicite de l'erreur dans le théorème central limite donnée ci-dessus est appelée inégalité de Berry*[8]*–Esséen*[9]*. 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,... une suite de variables aléatoires i.i.d. de loi $\mathcal{N}(0,1)$, indépendantes des variables aléatoires X_k. On pose

$$\widehat{S}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i, \qquad T_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i.$$

7. Francesco Paolo Cantelli (1875, Palerme - 1966, Rome), mathématicien italien.
8. Andrew C. Berry (1???, -????,), mathématicien...
9. Carl-Gustav Esséen (1918, ??? - 2001, ???), mathématicien suédois.

(Observez que T_n suit une loi $\mathcal{N}(0,1)$.) Soit $h : \mathbb{R} \to [0,1]$ une fonction de classe $\mathcal{C}^3$, telle que $h(s) = 1$ si $s \leq 0$, et $h(s) = 0$ si $s \geq 1$. Étant donné $t \in \mathbb{R}$ et $0 < \delta \leq 1$, on définit une nouvelle fonction $f_\delta : \mathbb{R} \to [0,1]$ par (voir Fig. 5.4)

$$f_\delta(x) = h(\delta^{-1}(x - t)).$$

Par construction, $\mathbf{1}_{(-\infty,t]}(x) \leq f_\delta(x)$, pour tout $x \in \mathbb{R}$, et donc

$$\mathbb{P}(\widehat{S}_n \leq t) = \mathbb{E}(\mathbf{1}_{(-\infty,t]}(\widehat{S}_n)) \leq \mathbb{E}(f_\delta(\widehat{S}_n)).$$

Puisque $\Phi(t) = \mathbb{E}(\mathbf{1}_{(-\infty,t]}(T_n))$, on obtient donc

$$\mathbb{P}(\widehat{S}_n \leq t) - \Phi(t) \leq \mathbb{E}(f_\delta(\widehat{S}_n)) - \mathbb{E}(f_\delta(T_n)) + \mathbb{E}(f_\delta(T_n)) - \mathbb{E}(\mathbf{1}_{(-\infty,t]}(T_n)).$$

Manifestement, T_n suivant une loi $\mathcal{N}(0,1)$,

$$\mathbb{E}(f_\delta(T_n)) - \mathbb{E}(\mathbf{1}_{(-\infty,t]}(T_n)) = \frac{1}{\sqrt{2\pi}} \int_t^{t+\delta} h(\delta^{-1}(x - t)) e^{-x^2/2} \mathrm{d}x \leq \frac{\delta}{\sqrt{2\pi}}.$$

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 :

$$\mathbb{E}(f_\delta(\widehat{S}_n)) - \mathbb{E}(f_\delta(T_n)) = \sum_{k=1}^{n} \big\{ \mathbb{E}(f_\delta(U_k + \frac{X_k}{\sqrt{n}})) - \mathbb{E}(f_\delta(U_k + \frac{Z_k}{\sqrt{n}})) \big\},$$

où $U_k = (Z_1 + Z_2 + \cdots + Z_{k-1} + X_{k+1} + X_{k+2} + \cdots + X_n)/\sqrt{n}$. 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 $U_k \leq Y \leq U_k + (X_k/\sqrt{n})$. On traite de la même façon le terme $f_\delta(U_k + (Z_k/\sqrt{n}))$. On obtient ainsi

$$\mathbb{E}(f_\delta(U_k + \frac{X_k}{\sqrt{n}})) - \mathbb{E}(f_\delta(U_k + \frac{Z_k}{\sqrt{n}})) \leq \frac{A}{\delta^3 n^{3/2}} \big(\mathbb{E}(|X_k|^3) + \mathbb{E}(|Z_k|^3) \big),$$

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

$$\mathbb{P}(\widehat{S}_n \leq t) - \Phi(t) \leq C n^{-1/8}.$$

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 $g_\delta(x) \leq \mathbf{1}_{(-\infty,t]}(x)$ pour tout x.

Méthode utilisant la fonction caractéristique. On ne démontre que la première affirmation. La preuve est presque identique à celle du Théorème 5.4.2. On peut à nouveau

supposer, sans perte de généralité, que $\mu = 0$ et $\sigma^2 = 1$. Dans ce cas, il suit du Lemme 4.2.1 que

$$\phi_X(t) = 1 - \tfrac{1}{2}t^2 + o(t^2).$$

D'autre part, la Proposition 4.2.1 et le Lemme 4.2.2 impliquent que la fonction caractéristique de la variable aléatoire $\widehat{S}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ satisfait

$$\phi_{\widehat{S}_n}(t) = \{\phi_X(t/\sqrt{n})\}^n = \big\{1 - \frac{t^2}{2n} + o(\frac{t^2}{n})\big\}^n,$$

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 4.2.4. □

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

$$\mathbb{P}\big(\sum_{i=1}^n X_i \in [\widehat{a},\widehat{b}]\big) \simeq \Phi(\frac{\widehat{b} - n\mu}{\sqrt{n\sigma^2}}) - \Phi(\frac{\widehat{a} - n\mu}{\sqrt{n\sigma^2}}).$$

Exemple 5.5.1. *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*

$$\mathbb{P}(N \geq 50) = \mathbb{P}(N \in [50, 400]) \simeq \Phi(\infty) - \Phi\big(\frac{50 - 40}{\sqrt{36}}\big) \simeq 0{,}05.$$

Il y a environ 5% *de chances d'obtenir au moins* 50 *pièces défectueuses.*

À titre de comparaison, N *suivant une loi* 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.

5.6 La loi 0-1 de Kolmogorov

L'énoncé précis de ce résultat nécessite un peu de terminologie.

Définition 5.6.1. *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 5.6.2. *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 $\mathcal{T}_\infty \overset{\text{déf}}{=} \bigcap_n \mathcal{T}_n$ 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 5.6.1 (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 5.6.3. *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 $\bigcup_n \sigma(X_1, X_2, \ldots, X_n)$. Il suit [10] que 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 5.6.4. *Une variable aléatoire mesurable par rapport à la tribu asymptotique $\mathcal{T}_\infty$ est dite* asymptotique.

Corollaire 5.6.1. *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.

□

10. Ceci requiert en fait un argument classique de théorie de la mesure. On observe 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 $\sigma(X_1, X_2, \ldots)$.

Chapitre 6

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.

6.1 Estimateurs

6.1.1 Définition, consistance, biais

Soit $\mathbb{P}$ une mesure de probabilité sur $\mathbb{R}^d$.

Définition 6.1.1. *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 6.1.1.
- *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 6.1.2. *Soit $X_1,\ldots,X_n$ un n-échantillon.*

On appelle statistique toute fonction mesurable $F(X_1,\ldots,X_n)$.

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 6.1.3. *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 6.1.2. *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 6.1.4. *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 6.1.5. *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 6.1.1. *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 5.2.2,

$$\mathbb{P}_\theta(|T_n - f(\theta)| \geq \epsilon) = \mathbb{P}_\theta((T_n - f(\theta))^2 \geq \epsilon^2) \leq \epsilon^{-2}\mathbb{E}_\theta\big((T_n - f(\theta))^2\big),$$

pour tout $\theta \in \Theta$. Puisque $\mathbb{E}_\theta\big((T_n - f(\theta))^2\big) = \mathrm{Var}_\theta(T_n) + \big(\mathbb{E}_\theta(T_n - f(\theta))\big)^2$, et que chacun de ces deux termes tend vers 0 par hypothèse, la conclusion suit. □

6.1.2 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

$$\tilde{\sigma}_n^2 = \frac{1}{n}\big(X_1^2 + \cdots + X_n^2\big) - \big(\frac{1}{n}(X_1 + \cdots + X_n)\big)^2.$$

La loi des grands nombres implique sa consistance, puisque le premier terme converge vers $\mathbb{E}_\theta(X_1^2)$ et le second vers $\mathbb{E}_\theta(X_1)^2$. Calculons le biais de cet estimateur. On a

$$\mathbb{E}_\theta\left(\frac{1}{n}\big(X_1^2 + \cdots + X_n^2\big)\right) = \mathbb{E}_\theta(X_1^2),$$

$$\mathbb{E}_\theta\left(\big(\frac{1}{n}(X_1 + \cdots + X_n)\big)^2\right) = \frac{1}{n}\mathbb{E}_\theta(X_1^2) + \frac{n-1}{n}\mathbb{E}_\theta(X_1)^2,$$

et donc

$$\mathbb{E}_\theta(\tilde{\sigma}_n^2) = \frac{n-1}{n}\big(\mathbb{E}_\theta(X_1^2) - \mathbb{E}_\theta(X_1)^2\big) = \frac{n-1}{n}\sigma^2.$$

Cet estimateur est donc biaisé. On 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 $(X_1,Y_1),\dots,(X_n,Y_n)$, 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

$$\tilde{\tau}_n = \frac{1}{n}\big(X_1Y_1 + \cdots + X_nY_n\big) - \big(\frac{1}{n}(X_1 + \cdots + X_n)\big)\big(\frac{1}{n}(Y_1 + \cdots + Y_n)\big)$$

est consistant et biaisé, mais que l'estimateur

$$\hat{\tau}_n = \frac{n}{n-1}\tilde{\tau}_n$$

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

$$\hat{I} = (b-a)\frac{1}{n}\big(h(U_1) + \cdots + h(U_n)\big),$$

où $U_1,\dots,U_n$ est un n-échantillon de loi uniforme sur $[a,b]$. $\hat{I}$ est un estimateur sans biais et consistant de I.

6.1.3 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 $X_1,\dots,X_n$ 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 6.1.3. *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 $X_1,\ldots,X_n$ de loi $\mathbb{P}_\theta$. Étant en possession d'une réalisation $x_1,\ldots,x_n$ 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 $x_1,\ldots,x_n$; 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 $X_1,\ldots,X_n$ discrètes.

Définition 6.1.6. *La vraisemblance (ou fonction de vraisemblance), notée $L(\theta; x_1,\ldots,x_n)$, d'un modèle en $x_1,\ldots,x_n$ est la probabilité d'observer $\{X_1 = x_1,\ldots,X_n = x_n\}$ lorsque le paramètre est θ.*

Remarque 6.1.1. *Insistons sur le fait que la variable est θ ; $x_1,\ldots,x_n$ sont des paramètres.*

Par indépendance des observations, on peut écrire

$$L(\theta; x_1,\ldots,x_n) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_i = x_i).$$

La définition ci-dessus n'a de sens que pour des variables aléatoires discrètes. Dans le cas continu, on travaille avec les densités :

$$L(\theta; x_1,\ldots,x_n) = \prod_{i=1}^{n} f_\theta(x_i),$$

où f_θ est la densité associée à la loi $\mathbb{P}_\theta$.

Définition 6.1.7. *On appelle* estimateur du maximum de vraisemblance *de θ la variable aléatoire correspondant à la valeur $\hat{\theta}(X_1,\ldots,X_n)$ en laquelle la fonction de vraisemblance atteint son maximum.*

Proposition 6.1.2. *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 ($x_i > 0$, $i = 1,\ldots,n$)

$$L(\lambda; x_1,\ldots,x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda(x_1+\cdots+x_n)}.$$

Pour trouver le maximum, on considère la log-vraisemblance,

$$\log L(\lambda; x_1,\ldots,x_n) = n\log\lambda - \lambda(x_1+\cdots+x_n).$$

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

$$\hat{\theta} = \max\{X_1, \ldots, X_n\}.$$

6.1.4 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 6.1.8. *Le risque quadratique de l'estimateur* $\hat{\theta}$ *de* θ *est défini par*

$$\mathcal{R}_{\hat{\theta}}(\theta) = \mathbb{E}_\theta\big((\hat{\theta} - \theta)^2\big).$$

Définition 6.1.9. *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

$$\mathcal{R}_T(\theta) = \mathbb{E}_\theta\big((T - f(\theta))^2\big).$$

Lemme 6.1.1. *Soit* $\hat{\theta}$ *un estimateur de* θ*. Alors*

$$\mathcal{R}_{\hat{\theta}}(\theta) = \mathrm{Var}_\theta(\hat{\theta}) + \big(\mathbb{E}_\theta(\hat{\theta} - \theta)\big)^2.$$

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 6.1.4. *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* $\tilde{\theta}$ *:*

$$\mathbb{P}_\theta(\tilde{\theta} \leq x) = \mathbb{P}_\theta(X_1 \leq x, \ldots, X_n \leq x) = \big(\mathbb{P}_\theta(X_1 \leq x)\big)^n = (\frac{x}{\theta})^n,$$

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* $\mathcal{R}_{\tilde{\theta}}(\theta) \geq \mathcal{R}_{\tilde{\tilde{\theta}}}(\theta)$, *pour tout* $\theta > 0$, *et tout* $n \geq 1$, *l'inégalité étant stricte dès que* $n \geq 3$. *L'estimateur* $\tilde{\tilde{\theta}}$ *est donc plus performant, malgré son biais. Remarquons qu'on peut facilement corriger le biais en considérant l'estimateur*

$$\frac{n+1}{n}\tilde{\theta}.$$

6.2 Intervalles de confiance

6.2.1 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 6.2.1. *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 6.2.1. *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 6.2.2. *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

$$I = [\hat{\theta},\alpha^{-1/n}\hat{\theta}].$$

6.2.2 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 6.2.2. *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 6.2.3. *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 5.2.2 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 6.2.3. *Pour un n-échantillon* $X_1,\ldots,X_n$, *un intervalle de confiance* asymptotique *pour* θ *au niveau* $1-\alpha$ *est un intervalle* $I_n = I_n(X_1,\ldots,X_n)$ *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* $I_n = I_n(X_1,\ldots,X_n)$ *tel que*

$$\lim_{n\to\infty} \mathbb{P}_\theta(I_n \ni \theta) \geq 1-\alpha, \quad \forall \theta \in \Theta.$$

Exemple 6.2.4. *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* $X_1,\ldots,X_n$. *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 6.2.5. *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,*

$$\liminf_{n\to\infty} \mathbb{P}_\theta(J_n \ni \mu) \geq \mathbb{P}(Z \in [-a,a]),$$

et l'affirmation est démontrée.

6.2.3 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 6.2.4. *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 6.2.1. *Un estimateur de* θ *asymptotiquement normal est nécessairement consistant.*

Démonstration. Soit $\epsilon > 0$. On a

$$\mathbb{P}_\theta(|T_n - \theta| \geq \epsilon) = \mathbb{P}_\theta(\sqrt{n}(T_n-\theta) \notin [-\epsilon\sqrt{n},\epsilon\sqrt{n}]) \leq \mathbb{P}_\theta(\sqrt{n}(T_n-\theta) \notin [-A,A]),$$

pour tout $n \geq A^2\epsilon^{-2}$. 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 6.2.5. *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.

6.3 Tests d'hypothèses

6.3.1 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 $1^{\text{ère}}$ espèce consiste à rejeter l'hypothèse nulle alors qu'elle est vraie.
- L'erreur de 2^{nde} 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.

6.3.2 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 6.3.1. *Une région de rejet est un événement $D = D(X_1,\ldots,X_n)$.*

Définition 6.3.2. *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 6.3.3. *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 6.3.4. *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 6.3.5. *Une hypothèse H est dite* simple *si l'ensemble Θ correspondant est réduit à un seul élément, sinon elle est dite* composite.

Exemple 6.3.1. *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*

$$\mathbb{P}_{\theta_0}(I \not\ni \theta_0) = \alpha.$$

6.3.3 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

$$D = \big\{|\bar{X}_n - \mu_0| \geq C\big\}.$$

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

$$D = \big\{\bar{X}_n > C\big\}.$$

On veut un niveau de risque de 5%, c'est-à-dire

$$\sup_{\mu\leq\mu_0} \mathbb{P}_\mu(D) = \sup_{\mu\leq\mu_0} \mathbb{P}\big(\frac{\sigma}{\sqrt{n}}Z > C-\mu\big) = 0{,}05,$$

où Z est normale standard. La borne supérieure est atteinte pour $\mu = \mu_0$, et on obtient donc $C \simeq \mu_0 + 1{,}64\sigma/\sqrt{n}$.

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

$$D=\Big\{\frac{|\bar{X}_n-\mu_0|}{S_n}\geq C\Big\},$$

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

$$D=\Big\{\frac{|\bar{X}_n|}{S_n}\geq\frac{2{,}093}{\sqrt{20}}\Big\}.$$

6.3.4 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]. L'objet central est le rapport de vraisemblance,

$$R(\theta_0,\theta_1;x_1,\dots,x_n)=\frac{L(\theta_1;x_1,\dots,x_n)}{L(\theta_0;x_1,\dots,x_n)}.$$

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.

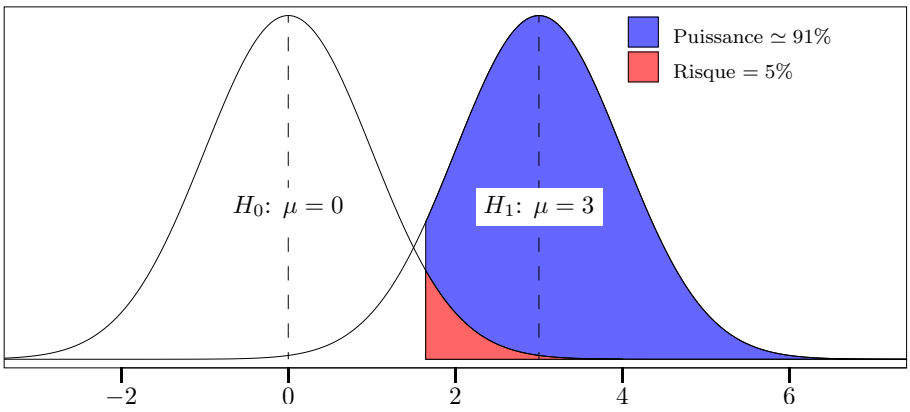


FIGURE 6.1: Test de deux hypothèses simples.

On prend pour région de rejet

$$D = \{R(\theta_0,\theta_1; X_1, \ldots, X_n) > C\},$$

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 6.3.2. *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 6.3.1 (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*

$$D = \left\{ (x_1, \ldots, x_n) \,:\, \prod_{i=1}^{n} f_{\theta_1}(x_i) > C \prod_{i=1}^{n} f_{\theta_0}(x_i) \right\},$$

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 $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathrm{d}\mathbf{x} = \mathrm{d}x_1 \cdots \mathrm{d}x_n$, 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 6.3.1. *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 $\{(x_1,\dots,x_n) : p_{\theta_1}(x_1)\cdots p_{\theta_1}(x_n) = Cp_{\theta_0}(x_1)\cdots p_{\theta_0}(x_n)\}$ 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 $\mathbb{P}_{\theta_0}(D > C)+\rho\mathbb{P}_{\theta_0}(D = C) = \alpha$.*

6.3.5 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 6.3.1. *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

$$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}.$$

Remarque 6.3.2. *La raison pour laquelle on a $k-1$ degrés de liberté et non k est que seuls $k-1$ des N_i sont indépendants puisque $N_1 + \cdots + N_k = n$.*

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 6.3.3. *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_k soient supérieurs à 5.*

Exemple 6.3.3. *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 :*

3. Johann Gregor Mendel (1822, Heinzendorf – 1884, Brünn), moine et botaniste Autrichien. Il est communément reconnu comme le père fondateur de la génétique.

	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,

$$Z_{556} = \frac{\left(315 - 556 \cdot \frac{9}{16}\right)^2}{556 \cdot \frac{9}{16}} + \frac{\left(101 - 556 \cdot \frac{3}{16}\right)^2}{556 \cdot \frac{3}{16}} + \frac{\left(108 - 556 \cdot \frac{3}{16}\right)^2}{556 \cdot \frac{3}{16}} + \frac{\left(32 - 556 \cdot \frac{1}{16}\right)^2}{556 \cdot \frac{1}{16}}$$
$$\simeq 0{,}47.$$

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 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.

Chapitre 7

Marches aléatoires

Les marches aléatoires forment une classe très importante de processus stochastiques, avec 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 des marches aléatoires sur $\mathbb{Z}^d$, en nous concentrant principalement sur le cas des marches aléatoires simples.

7.1 Quelques généralités sur les processus stochastiques

Un processus stochastique est une collection $(Y_t)_{t\in\mathbb{T}}$ de variables aléatoires à valeurs dans un ensemble E et indexée par les éléments d'un ensemble $\mathbb{T} \subset \mathbb{R}$. Dans cette section, nous allons brièvement décrire comment un tel processus est construit, dans le cas où $\mathbb{T} = \mathbb{N}$ (processus en temps discret) et E est un ensemble au plus dénombrable.

On suppose données les lois fini-dimensionnelles du processus, c'est-à-dire les fonctions de masse conjointes f_n des variables aléatoires $Y_0,\ldots,Y_n$, pour tout $n \geq 0$. Évidemment, ces fonctions doivent être consistantes, dans le sens que

$$\sum_{y\in E} f_{n+1}(y_0,\ldots,y_n,y) = f_n(y_0,\ldots,y_n). \tag{7.1}$$

Notre but est de construire un espace probabilisé $(\Omega,\mathcal{F},\mathbb{P})$ et une collection de variables aléatoires $(Y_n)_{n\geq 0}$ sur cet espace, à valeurs dans E et telle que $\mathbb{P}(Y_0 = y_0,\ldots,Y_n = y_n) = f_n(y_0,\ldots,y_n)$ pour tout $n \geq 0$.

Univers. L'univers $\Omega = E^{\mathbb{N}}$ est formé de toutes les trajectoires possibles $\omega = (y_0,y_1,y_2,\ldots)$ du processus.

Variables aléatoires. Les variables aléatoires $Y_n : \Omega \to E$, $n \in \mathbb{N}$, sont définies par

$$\omega = (y_0,y_1,y_2,\ldots) \mapsto Y_n(\omega) = y_n.$$

Tribu. On introduit la tribu $\mathcal{F}_n$ des événements antérieurs au temps n : celle-ci est engendrée par les ensembles

$$[y_0,y_1,\ldots,y_n] \equiv \{\omega \in \Omega \,:\, Y_0(\omega) = y_0, Y_1(\omega) = y_1,\ldots,Y_n(\omega) = y_n\}$$

décrivant les n premiers pas de la trajectoires. On appelle cylindres les éléments de $\mathcal{F}_n$ de la forme

$$\{\omega \in \Omega : Y_0(\omega) \in A_0, Y_1(\omega) \in A_1, \ldots, Y_n(\omega) \in A_n\} = \bigcup_{\substack{y_0 \in A_0 \\ \vdots \\ y_n \in A_n}} [y_0, \ldots, y_n],$$

où $A_0, A_1, \ldots, A_n \subset E$.

Manifestement, $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ (on dit que $(\mathcal{F}_n)_{n\geq 0}$ est une filtration), ce qui permet de conclure que $\bigcup_{n\geq 0} \mathcal{F}_n$ est une algèbre. La tribu associée au processus stochastique est alors définie par

$$\mathcal{F} = \sigma(\bigcup_{n\geq 0} \mathcal{F}_n).$$

Mesure de probabilité. On introduit à présent une mesure de probabilité sur $\mathcal{F}_n$ en posant

$$\mathbb{P}_n([y_0, \ldots, y_n]) = f_n(y_0, \ldots, y_n).$$

Les mesures $\mathbb{P}_n$ induisent une mesure de probabilité $\mathbb{P}$ sur l'algèbre $\bigcup_{n\geq 0} \mathcal{F}_n$: si $A \in \bigcup_{n\geq 0} \mathcal{F}_n$, il existe $m \in \mathbb{N}$ tel que $A \in \mathcal{F}_m$, et on peut donc poser

$$\mathbb{P}(A) = \mathbb{P}_m(A).$$

On vérifie grâce à (7.1) que cette définition ne dépend pas du m choisi. À ce stade, le Théorème d'extension de Carathéodory implique que $\mathbb{P}$ peut être étendue de façon unique en une mesure de probabilité sur $\mathcal{F}$.

7.2 Marche aléatoire simple unidimensionnelle

Soit $X_1, X_2, \ldots$ une suite de variables aléatoires i.i.d. telles que

$$\mathbb{P}(X_1 = 1) = p \text{ et } \mathbb{P}(X_1 = -1) = 1 - p \equiv q,$$

pour un certain $p \in [0,1]$. On appelle marche aléatoire simple [1] partant de $a \in \mathbb{Z}$ la suite de variables aléatoires $(S_n)_{n\geq 1}$ définie par

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

On notera $\mathbb{P}_a$ la loi de la marche aléatoire simple partant de a. Cette marche aléatoire est dite symétrique lorsque $p(= q) = \frac{1}{2}$.

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.

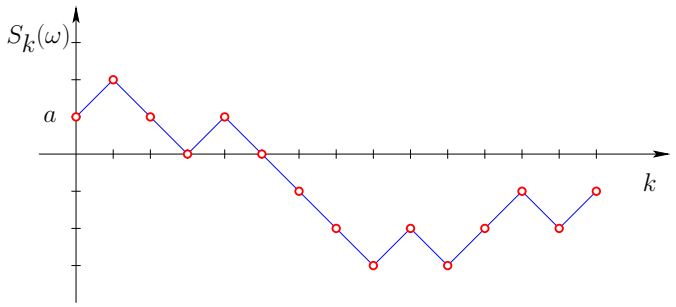


FIGURE 7.1: Le début de la trajectoire associée à une réalisation d'une marche aléatoire simple partant de a.

Remarque 7.2.1. *On a défini la marche aléatoire directement à partir d'une suite de variables aléatoires i.i.d. $(X_k)_{k\geq 1}$. Alternativement, on aurait pu la construire comme esquissé dans la section précédente, en spécifiant ses lois fini-dimensionnelles : pour tout $n \geq 0$ et toute suite $s_0,\ldots,s_n$ telle que $|s_k - s_{k-1}| = 1$, $1 \leq k \leq n$,*

$$\mathbb{P}_a(S_0 = s_0, S_1 = s_1, \ldots, S_n = s_n) = \mathbf{1}_{\{s_0=a\}} p^{n_+} q^{n-n_+},$$

où $n_+ = \#\{1 \leq k \leq n : s_k - s_{k-1} = 1\}$ est le nombre de pas « vers la droite » dans la portion de trajectoire $s_0,\ldots,s_n$.

Une réalisation de la trajectoire de la marche est donnée par la suite des couples $\big((k,S_k(\omega))\big)_{k\geq 0}$, avec la convention que $S_0 = a$ (voir Figure 7.1). Le paramètre k est souvent interprété comme le temps, et la ligne polygonale passant par chacun des points $(k,S_k(\omega))$ est appelée la trajectoire de la marche. Le processus $(S_n)_{n\geq 0}$ possède les importantes propriétés suivantes.

Lemme 7.2.1. *1. (Homogénéité spatiale) $\mathbb{P}_a(S_n = s) = \mathbb{P}_{a+b}(S_n = s+b)$, $\forall a,b \in \mathbb{Z}$.*

2. (Propriété de Markov) Soit $B \in \sigma(S_0,\ldots,S_n)$ 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

$$\mathbb{P}_a\big((S_n,S_{n+1},S_{n+2},\ldots) \in A \;\big|\; S_n = s, B\big) = \mathbb{P}_s\big((S_0,S_1,S_2,\ldots) \in A\big),$$

pour tout ensemble de trajectoires $A \in \mathcal{F}$.

Ce qu'affirme la propriété de Markov, c'est que 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. Pour la première affirmation,

$$\mathbb{P}_a(S_n = s) = \mathbb{P}(\sum_{i=1}^n X_i = s - a) = \mathbb{P}(\sum_{i=1}^n X_i = s + b - (a+b)) = \mathbb{P}_{a+b}(S_n = s+b).$$

Pour la seconde propriété, on observe que

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

où l'on a utilisé le fait que $\{S_n=s\}\cap B\in\sigma(X_1,\ldots,X_n)$ est indépendant de $(X_{n+k})_{k\geq 1}$ pour la deuxième égalité, et le fait que $(s,s+X_{n+1},s+X_{n+1}+X_{n+2},\ldots)$ et $(s,s+X_1,s+X_1+X_2,\ldots)$ ont même loi, puisque les $(X_i)_{i\geq 1}$ sont i.i.d.. □

7.2.1 Ruine du joueur

Parmi les nombreuses interprétations de ce processus, 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

$$\begin{cases}f(a)=p\,f(a+1)+q\,f(a-1), & 1\leq a\leq N-1\\ f(0)=1, f(N)=0.\end{cases}\tag{7.2}$$

Lemme 7.2.2. *L'équation (7.2) possède une unique solution.*

Démonstration. Si f et g sont solutions de (7.2), 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 $\bar{x}\in\{1,\ldots,N-1\}$ tel que $|h(\bar{x})|$ soit maximum ; on suppose sans perte de généralité que $h(\bar{x})\geq 0$ (sinon il suffit de considérer $g-f$). On a alors

$$h(\bar{x}+1)=\frac{1}{p}\big(h(\bar{x})-q\,h(\bar{x}-1)\big)\geq\frac{1}{p}\big(h(\bar{x})-q\,h(\bar{x})\big)=h(\bar{x}),$$

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 (7.2) est donnée par [2]

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

Lorsque $p \neq q$, on vérifie aisément qu'elle est donnée par [3]

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

7.2.2 Propriétés trajectorielles : approche combinatoire

La méthode utilisée dans la section précédente peut être étendue à des situations beaucoup plus générales, comme nous le verrons plus tard. Dans cette section, nous allons utiliser une autre approche, de nature combinatoire. De cette manière, nous établirons un certain nombre de propriétés trajectorielles pour la marche aléatoire simple unidimensionnelle, dont certaines peuvent sembler surprenantes au premier abord.

Lemme 7.2.3. *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 $n \to \infty$,

$$\mathbb{P}_0(S_{2n} = 0) = \frac{1 + o(1)}{\sqrt{\pi n}}. \tag{7.4}$$

2. Une façon de trouver cette solution est d'observer que (7.2) 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 $f(a) = 1 - a/N$.

3. Trouver la solution lorsque $p \neq q$ conduit à des calculs plus pénibles, mais est conceptuellement très simple. On peut dans ce cas écrire (7.2) sous la forme suivante : pour tout $1 \leq a \leq N-1$,

$$\begin{pmatrix} f(a+1) \\ f(a) \end{pmatrix} = \begin{pmatrix} 1/p & -q/p \\ 1 & 0 \end{pmatrix} \begin{pmatrix} f(a) \\ f(a-1) \end{pmatrix} = \begin{pmatrix} 1/p & -q/p \\ 1 & 0 \end{pmatrix}^a \begin{pmatrix} f(1) \\ f(0) \end{pmatrix}. \tag{7.3}$$

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 (7.3) avec $a = N - 1$ que

$$f(1) = \frac{(q/p) - (q/p)^N}{1 - (q/p)^N},$$

et donc, en appliquant à nouveau (7.3),

$$f(a) = \frac{(q/p)^a - (q/p)^N}{1 - (q/p)^N}.$$

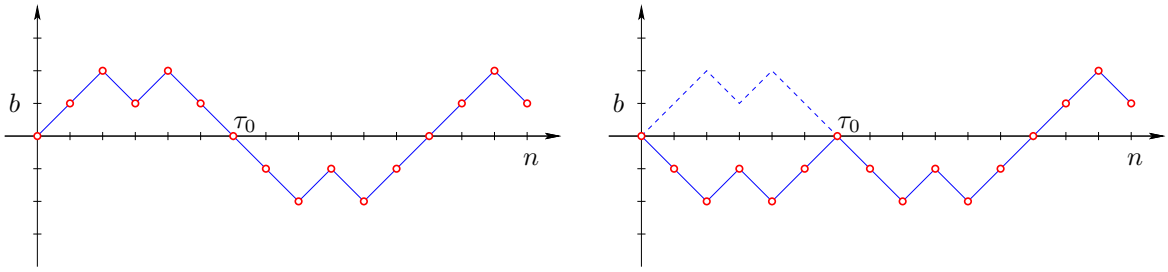


FIGURE 7.2: Le principe de réflexion.

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 $\binom{n}{n_+}$. □

Nous allons à présent étudier le temps mis par la marche pour retourner à son point de départ. Nous noterons $\tau_0 = \min\{n \geq 1 : S_n = 0\}$ la variable aléatoire correspondante, avec la convention habituelle que $\min \varnothing = +\infty$.

Lemme 7.2.4. *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. Chacune des portions de trajectoire joignant (0,0) à (n,b) a probabilité $p^{(n+b)/2}q^{(n-b)/2}$. Il ne reste donc plus qu'à déterminer le nombre de ces portions de trajectoires ne revisitant pas 0.

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 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 (*cf.* Fig. 7.2) : 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

$$\#\mathcal{T}_{\pm}[(1,1),(n,b)] = \binom{n-1}{\frac{n+b}{2}-1} - \binom{n-1}{\frac{n+b}{2}} = \frac{b}{n}\binom{n}{\frac{n+b}{2}}. \tag{7.5}$$

Par conséquent,

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

par le Lemme 7.2.3. □

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

Lemme 7.2.5. *Dans le cas symétrique,*

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

Démonstration. En appliquant le résultat du lemme précédent, on obtient

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

la troisième ligne suivant de (7.5). □

Le résultat précédent montre que, dans le cas symétrique, $\mathbb{P}_0(\tau_0 > 2n)$ tend vers 0 plutôt lentement ($\mathbb{P}_0(\tau_0 > 2n) = (1 + o(1))/\sqrt{\pi n}$, par Stirling). Bien entendu, puisque $\mathbb{P}_0(\tau_0 = \infty) = \mathbb{P}_0(\bigcap_{n\geq 1}\{\tau_0 > n\}) = \lim_{n\to\infty}\mathbb{P}_0(\tau_0 > n) = 0$, la marche retourne à l'origine presque sûrement. Intuitivement, il semble clair que cela devrait impliquer qu'elle y retourne infiniment souvent, ce que confirme le lemme suivant.

Lemme 7.2.6. *Soit N le nombre de retours de la marche aléatoire à l'origine. Alors, dans le cas symétrique,*

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

Démonstration. Soit $\tau_0^{(n)}$ le temps du $n^{\text{ème}}$ retour en 0 (avec $\tau_0^{(n)} = \infty$ si $N < n$). Pour tout $k \in \mathbb{N}$,

$$\mathbb{P}_0(N = k) = \sum_{\ell \geq k} \mathbb{P}_0(\tau_0^{(k)} = 2\ell)\mathbb{P}_0(N = k \,|\, \tau_0^{(k)} = 2\ell).$$

La propriété de Markov implique donc, puisque $\{\tau_0^{(k)} = 2\ell\}$ ne dépend que des 2ℓ premiers pas de la trajectoire et implique que $S_{2\ell} = 0$,

$$\begin{aligned}\mathbb{P}_0(N = k \,|\, \tau_0^{(k)} = 2\ell) &= \mathbb{P}_0(S_j \neq 0, \forall j > 2\ell \,|\, \tau_0^{(k)} = 2\ell) \\ &= \mathbb{P}_0(S_j \neq 0, \forall j > 2\ell \,|\, S_{2\ell} = 0, \tau_0^{(k)} = 2\ell) \\ &= \mathbb{P}_0(S_j \neq 0, \forall j > 0) = \mathbb{P}_0(N = 0) = 0,\end{aligned}$$

et donc $\mathbb{P}_0(N = k) = 0$, pour tout $k \in \mathbb{N}$. La conclusion suit donc, puisque

$$\mathbb{P}_0(N < \infty) = \sum_{k \geq 1} \mathbb{P}_0(N = k) = 0.$$

□

Une autre conclusion intéressante du Lemme 7.2.5 est que l'espérance du temps τ_0 du premier retour à l'origine est infinie :

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

la dernière identité suivant de (7.4). Ainsi, s'il est certain que la marche symétrique passera par l'origine infiniment souvent, elle le fera très rarement. Une autre façon de voir cela est de réaliser que l'espérance du nombre de retours en 0 jusqu'au temps n est donnée par

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

et la fréquence des retours tend donc vers 0 comme $n^{-1/2}$.

Nous allons à présent obtenir une formule explicite pour le temps de premier retour en 0.

Lemme 7.2.7. *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.)

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

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

où l'on a utilisé le résultat du Lemme 7.2.4. □

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 Lemme 7.2.5.)

On peut également s'intéresser au moment de la *dernière* visite en 0 au cours des $2n$ premiers pas, $\nu_0(2n) = \max\{0 \leq k \leq 2n \,:\, S_k = 0\}$.

Lemme 7.2.8 (Loi de l'arcsinus pour la dernière visite en 0)**.** *On suppose que $p = 1/2$. Pour tout $0 \leq k \leq n$,*

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

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

$$\lim_{n\to\infty} \mathbb{P}_0(\nu_0(2n) \leq 2\alpha n) = \frac{2}{\pi}\arcsin\sqrt{\alpha}.$$

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 Lemme 7.2.5.

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

$$\mathbb{P}_0(S_{2k} = 0)\mathbb{P}_0(S_{2n-2k} = 0) = \binom{2k}{k}\binom{2n-2k}{n-k}2^{-2n} = \frac{1 + o(1)}{\pi\sqrt{k(n-k)}}, \tag{7.6}$$

lorsque k et $n-k$ tendent vers l'infini.

D'autre part, il suit de la première affirmation que, pour tout $0 \leq m \leq n$,

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

Ceci implique que, pour tout $0 < \alpha \leq \frac{1}{2}$,

$$\lim_{n\to\infty} \mathbb{P}_0(\nu_0(2n) \leq 2\alpha n) = \lim_{n\to\infty} \mathbb{P}_0(\nu_0(2n) \geq 2(1-\alpha)n) = 1 - \lim_{n\to\infty} \mathbb{P}_0(\nu_0(2n) \leq 2(1-\alpha)n), \tag{7.7}$$

puisque $\lim_{n\to\infty} \sup_{0\leq m\leq 2n} \mathbb{P}_0(\nu_0(2n) = m) = 0$. On en déduit en particulier que

$$\lim_{n\to\infty} \mathbb{P}_0(\nu_0(2n)/2n \in [0,\tfrac{1}{2}]) = \tfrac{1}{2}.$$

Supposons à présent que $\alpha \in (\frac{1}{2},1)$. Dans ce cas, on peut appliquer (7.6) pour obtenir

$$\lim_{n\to\infty} \mathbb{P}_0(\frac{\nu_0(2n)}{2n} \in (\tfrac{1}{2},\alpha]) = \lim_{n\to\infty} \frac{1}{\pi n} \sum_{\alpha n > k > n/2} \big(\frac{k}{n}(1-\frac{k}{n})\big)^{-1/2}$$
$$= \frac{1}{\pi} \int_{\frac{1}{2}}^{\alpha} \big(x(1-x)\big)^{-1/2}\, \mathrm{d}x$$
$$= \frac{2}{\pi} \arcsin\sqrt{\alpha} - \tfrac{1}{2},$$

et donc $\lim_{n\to\infty} \mathbb{P}_0(\frac{\nu_0(2n)}{2n} \leq \alpha) = \frac{2}{\pi} \arcsin\sqrt{\alpha}$.

Le cas $\alpha \in (0,\frac{1}{2})$ suit facilement du précédent, en utilisant (7.7) et l'identité

$$\arcsin\sqrt{1-\alpha} = \frac{\pi}{2} + \arcsin\sqrt{\alpha}.$$

□

Le lemme précédent a des conséquences peut-être assez surprenantes au premier abord : si l'on procède à un grand nombre de lancers à pile ou face, la dernière fois que le nombre de « pile » et le nombre de « face » obtenus ont coïncidé est proche du début ou de la fin de la série avec une probabilité substantielle : on a, par exemple (voir également la Figure 7.3),

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

Nous allons à présent nous intéresser au temps de première visite en un sommet $b \neq 0$, $\tau_b = \min\{n \geq 1 \,:\, S_n = b\}$.

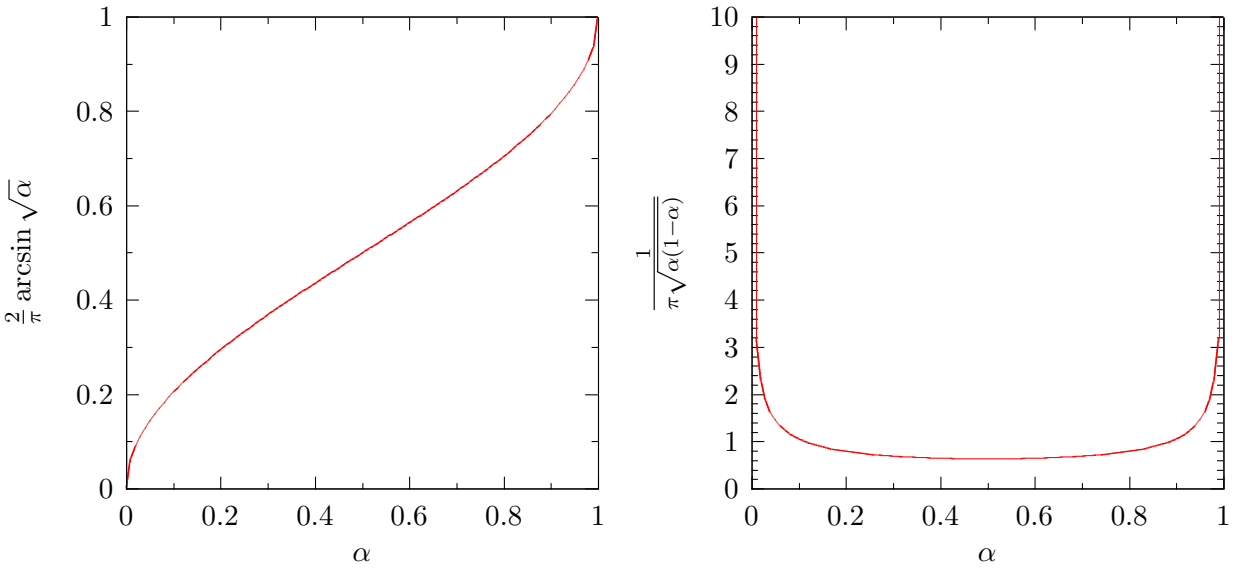


FIGURE 7.3: La fonction de répartition (gauche) et la densité (droite) de la loi de l'arcsinus.

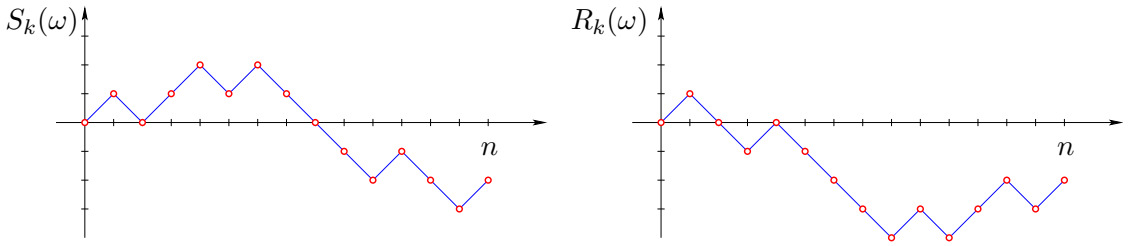


FIGURE 7.4: Une trajectoire et la trajectoire retournée.

Lemme 7.2.9. *Pour tout $b \neq 0$,*

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

Démonstration. Cette preuve repose sur une autre idée : le 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. 7.4)

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

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 Lemme 7.2.4. □

Il suit du lemme précédent que le nombre moyen de visites au site $b \neq 0$ avant le premier retour en 0 est égal à

$$\mathbb{E}_0(\sum_{n \geq 1} \mathbf{1}_{\{\tau_0 > n, S_n = b\}}) = \sum_{n \geq 1} \mathbb{P}_0(\tau_0 > n, S_n = b) = \sum_{n \geq 1} \mathbb{P}_0(\tau_b = n) = \mathbb{P}_0(\exists n \geq 0 \,:\, S_n = b).$$

Ce résultat a une conséquence assez surprenante.

Lemme 7.2.10. *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. 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 donc que, par symétrie,

$$\mathbb{P}_0(\exists n \geq 0 \,:\, S_n = b) = 1, \qquad \forall b \in \mathbb{Z}.$$

□

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 lemme ci-dessus montre que celle-ci est de 1 franc, *quelle que soit la valeur de m* !

Nous allons à présent établir un autre résultat classique, également contre-intuitif lorsqu'on le rencontre pour la première fois.

Lemme 7.2.11 (loi de l'arcsinus pour les temps de séjour)**.** *On suppose que $p = 1/2$. Soit (cf. Fig. 7.5)*

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

le temps pendant lequel la marche est positive. Alors,

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

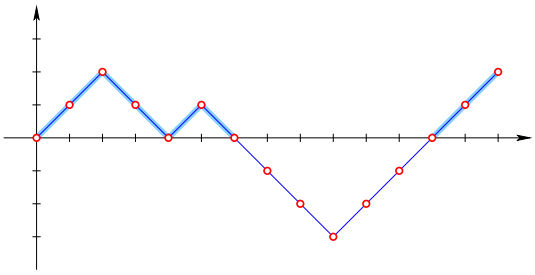


FIGURE 7.5: Sur cette réalisation, le temps total passé au-dessus de 0 pendant les 14 premiers pas est $t_{14}^+ = 8$.

(Observez que t_{2n}^+ est nécessairement pair.) En particulier, pour tout $0 < \alpha < 1$,

$$\lim_{n\to\infty} \mathbb{P}_0(\frac{t_{2n}^+}{2n} \leq \alpha) = \frac{2}{\pi} \arcsin\sqrt{\alpha}.$$

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

$$f_{2n}(2k) = g_{2k} g_{2n-2k}. \tag{7.8}$$

La première observation est que, par le Lemme 7.2.5,

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

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 (7.8) 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 (7.8) pour tous les $0 \leq k \leq n$ lorsque $n = 1$. Supposons donc (7.8) 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 (7.8). La dernière identité suit de l'observation que, pour tout $\ell \geq 1$,

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

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 Lemme 7.2.8. □

Discutons à présent quelques conséquences de la loi de l'arcsinus. 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 (voir Fig. 7.3) : avec probabilité $1/5$, la marche passera près de 97,6% de son temps du même côté de l'origine ; avec probabilité $1/10$, elle le fera pendant 99,4% de son temps. La figure 7.6 montre cinq trajectoires typiques d'une marche aléatoire simple symétrique sur $\mathbb{Z}$.

De façon plus imagée, supposons que deux joueurs jouent à pile ou face. On suppose que la pièce est jetée 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 dominera l'autre joueur *pendant plus de* 364 *jours et* 10 *heures* !

7.2.3 Propriétés trajectorielles : fonctions génératrices

Nous allons à présent donner une illustration de l'utilisation des fonctions génératrices dans ce contexte.

Nous nous intéressons à nouveau à la loi des temps de retour à l'origine. Évidemment il suffit de considérer la loi du temps de premier retour τ_0, puisque les intervalles entre retours consécutifs suivent la même loi. 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 (c'est-à-dire que $\mathbb{P}_0(\tau_0 = \infty) > 0$), et dans ce cas $\mathbb{H}(1) = \mathbb{P}_0(\tau_0 < \infty) < 1$.

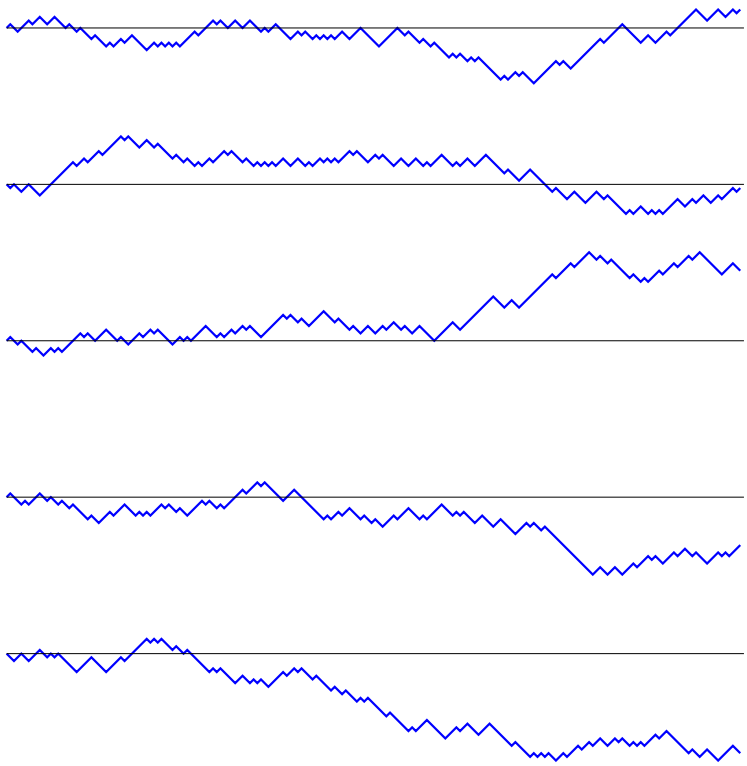


FIGURE 7.6: Cinq trajectoires de la marche aléatoire simple symétrique sur $\mathbb{Z}$ (seuls les 200 premiers pas sont affichés). Observez la présence de très longues excursions (morceaux de trajectoires compris entre deux zéros successifs). L'espérance de la longueur de ces dernières est infinie.

Lemme 7.2.12. *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 2.2.5) 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 7.2.1. *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 7.2.12 (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 7.2.1. *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 et seulement si $p = \frac{1}{2}$.

7.3 Marche aléatoire simple sur $\mathbb{Z}^d$

Nous allons à présent brièvement décrire la généralisation du processus étudié dans la section précédente de $\mathbb{Z}$ à $\mathbb{Z}^d$. Le type de processus ainsi obtenu (et leurs généralisations) jouent 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 [4] en 1905 afin d'expliquer le mouvement erratique des particules de pollen dans l'eau observé en 1827 par Brown [5], et de cette façon confirmer la théorie atomiste alors encore controversée en permettant à Perrin [6] de déterminer expérimentalement la constante d'Avogadro [7]).

Soit $X_1, X_2, \ldots$ une suite de variables 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.

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.

4. Albert Einstein (1879, Ulm – 1955, Princeton), physicien allemand, puis apatride (1896), suisse (1899), et enfin suisse-américain (1940). Prix Nobel de physique en 1921.

5. Robert Brown (1773, Montrose – 1858, Londres), botaniste britannique.

6. Jean Baptiste Perrin (1870, Lille – 1942, New York), physicien français. Prix Nobel de Physique en 1926.

7. Lorenzo Romano Amedeo Carlo Avogadro, Comte de Quaregna et Cerreto (1776, Turin – 1856, Turin). Physicien et chimiste italien.

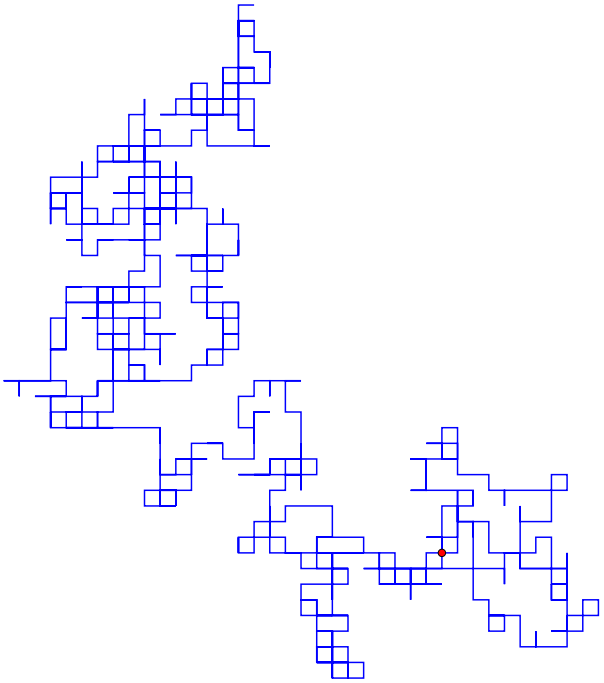


FIGURE 7.7: Les 1000 premiers pas d'une marche aléatoire simple symétrique sur $\mathbb{Z}^2$ partant du point rouge.

On vérifie aisément que les propriétés énoncées dans le Lemme 7.2.1 sont également vérifiées ici (la structure étant identique).

7.3.1 Probabilités de sortie

Le but de cette sous-section est de montrer que l'approche utilisée dans le cas unidimensionnel dans la Sous-section 7.2.1 s'étend sans autre à cette situation plus générale (Figure 7.8)).

Lemme 7.3.1. *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

$$\Delta_d f(x) = \frac{1}{2d} \sum_{\substack{y \in \mathbb{Z}^d \\ |x-y|=1}} f(y) - f(x).$$

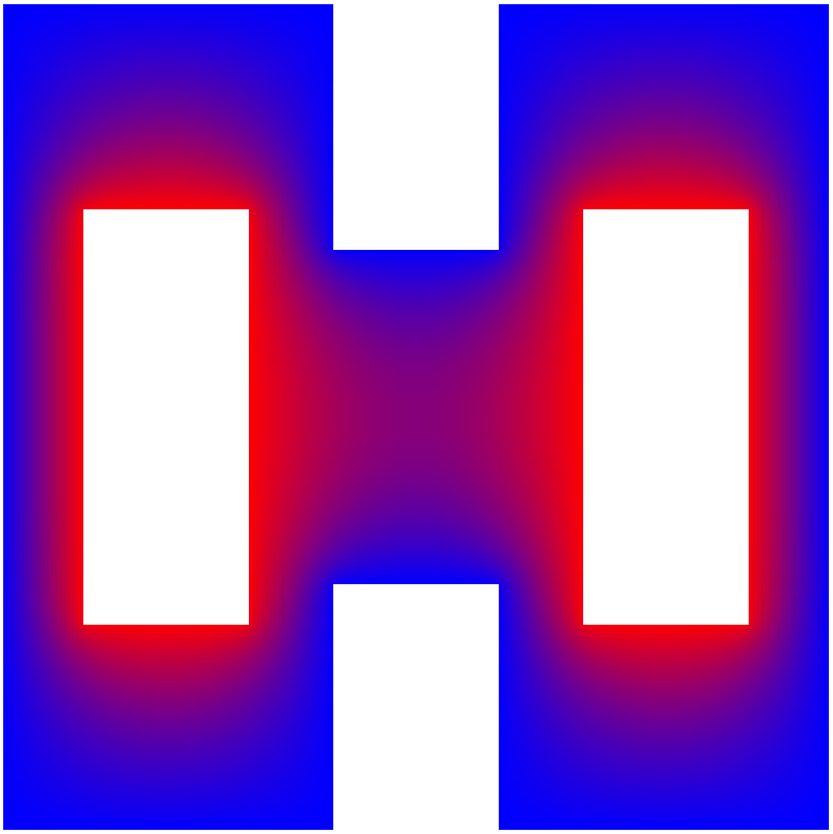


FIGURE 7.8: Probabilités de pénétrer dans un des deux trous avant de sortir du domaine : la couleur passe du bleu au rouge lorsque la probabilité passe de 0 à 1.

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 (7.3.1), 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 $f \equiv g$. □

7.3.2 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 [8] en 1921 ; il montre que la dimension du réseau affecte crucialement le comportement de la marche aléatoire.

Théorème 7.3.1. *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 7.3.2. *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

$$\mathbb{E}_0(N) = \mathbb{E}_0(\sum_{n\geq 1} \mathbf{1}_{\{S_n=0\}}) = \sum_{n\geq 1} \mathbb{P}_0(S_n = 0),$$

le lemme est démontré. □

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

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, le théorème de Fubini permet d'écrire, 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 $d > 2$. □

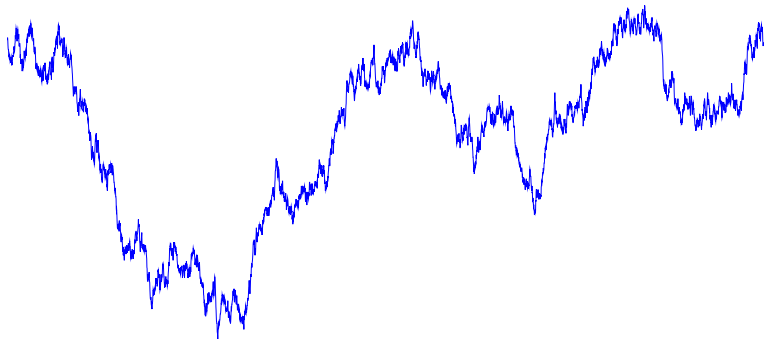


FIGURE 7.9: Partie d'une trajectoire du mouvement brownien en dimension 1.

Remarque 7.3.1. *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.

Lemme 7.3.3. *La marche aléatoire simple symétrique bidimensionnelle est récurrente-nulle.*

Démonstration. Notons $S_n = (S_n(1), S_n(2))$ la marche aléatoire simple symétrique bidimensionnelle, 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

$$\mathbb{E}_0(\tau_0) = \mathbb{E}_0\big(\inf\{n \geq 1 \,:\, S_n(1) = S_n(2) = 0\}\big) \geq \mathbb{E}_0\big(\inf\{n \geq 1 \,:\, \widetilde{S}_n = 0\}\big) = \mathbb{E}_0(\widetilde{\tau}_0) = \infty,$$

puisque la marche aléatoire simple symétrique unidimensionnelle est récurrente-nulle. □

7.3.3 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

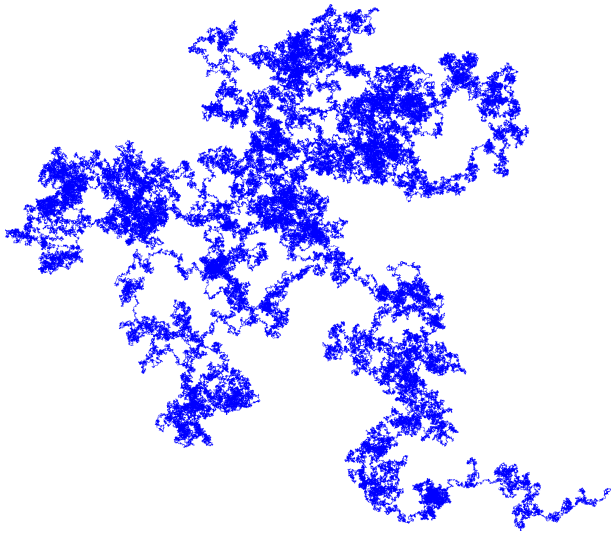


FIGURE 7.10: Partie d'une trajectoire du mouvement brownien en dimension 2 (tous les temps considérés sont superposés).

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[9]. Une partie d'une trajectoire de ce processus est donnée sur la Figure 7.9.

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 7.10, on a tracé une portion de trajectoire dans le cas bidimensionnel.

9. Norbert Wiener (1894, Columbia – 1964, Stockholm), mathématicien américain.

Chapitre 8

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.

8.1 Définition et exemples

Soit $X_0, X_1, X_2, \ldots$ un suite de variables aléatoires prenant valeur dans un ensemble S dénombrable. Nous noterons X le processus stochastique correspondant et $\mathbb{P}$ sa loi.

Définition 8.1.1. *Le processus X est une chaîne de Markov s'il possède la propriété de Markov,*

$$\mathbb{P}(X_n = s_n \,|\, X_0 = s_0, X_1 = s_1, \ldots, X_{n-1} = s_{n-1}) = \mathbb{P}(X_n = s_n \,|\, X_{n-1} = s_{n-1}),$$

pour tout $n \geq 1$ et tout $s_0, s_1, \ldots, s_n \in S$.

S est appelé espace des états de la chaîne.

Les marches aléatoires du chapitre 7 fournissent un exemple de chaîne de Markov, avec $S = \mathbb{Z}^d$. La taille de la population dans le processus de branchement étudié dans la Sous-section 4.1.2 est un autre exemple de processus de Markov, cette fois avec $S = \mathbb{N}$.

Définition 8.1.2. *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, i, j.

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

$$p(i,j) = \mathbb{P}(X_1 = j \,|\, X_0 = i).$$

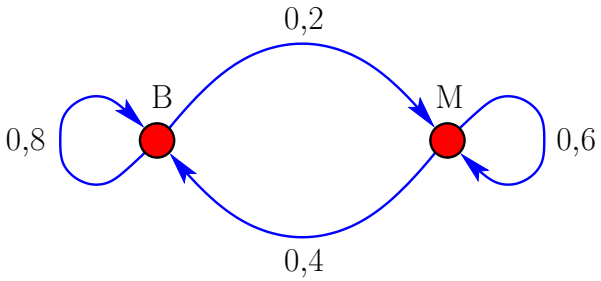


FIGURE 8.1: La représentation graphique de la chaîne de l'exemple 8.1.1.

Définition 8.1.3. *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*).*

Lemme 8.1.1. *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 8.1.4. *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 $\mathbb{E}_i$.

Remarque 8.1.1. *À nouveau, la construction du processus peut se faire comme esquissé dans la section 7.1, les moments fini-dimensionnels associés à la chaîne de Markov* $(\mathbf{P}, \boldsymbol{\mu})$ *étant donnés par*

$$\mathbb{P}_{\boldsymbol{\mu}}(X_0 = s_0, X_1 = s_1, \ldots, X_n = s_n) = \mu(s_0)p(s_0, s_1)\cdots p(s_{n-1}, s_n),$$

pour toute suite $s_0, \ldots, s_n \in S$.

Exemple 8.1.1. *Après une longue collecte de données, Robinson a conçu le modèle suivant pour décrire approximativement le temps qu'il fera sur son île :*

$$S = \{\textit{beau temps, mauvais temps}\}, \qquad \textit{et} \qquad \mathbf{P} = \begin{pmatrix} 0,8 & 0,2 \\ 0,4 & 0,6 \end{pmatrix}.$$

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 8.1.*

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}*.*

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 $n \geq 1$,

$$\begin{aligned} p_n(i,j) &= \mathbb{P}_i(X_n = j) \\ &= \sum_{k \in S} \mathbb{P}_i(X_n = j, X_{n-1} = k) \\ &= \sum_{k \in S} \mathbb{P}_i(X_n = j \,|\, X_{n-1} = k)\, \mathbb{P}_i(X_{n-1} = k) \\ &= \sum_{k \in S} \mathbb{P}_k(X_1 = j)\, \mathbb{P}_i(X_{n-1} = k) \\ &= \sum_{k \in S} p(k,j) p_{n-1}(i,k). \end{aligned}$$

Cette relation est connue sous le nom d'équation de Chapman-Kolmogorov. On en déduit facilement le résultat fondamental suivant.

Théorème 8.1.1. *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.

Théorème 8.1.2. *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.

$$\begin{aligned}\mu_n(i) = \mathbb{P}_{\boldsymbol{\mu}_0}(X_n = i) &= \sum_{j \in S} \mathbb{P}_{\boldsymbol{\mu}_0}(X_n = i \,|\, X_0 = j)\mathbb{P}_{\boldsymbol{\mu}_0}(X_0 = j) \\ &= \sum_{j \in S} p_n(j,i)\mu_0(j) = (\boldsymbol{\mu}_0 \mathbf{P}^n)_i.\end{aligned}$$

□

Nous nous intéresserons principalement à deux classes particulières, mais très importantes, de chaînes de Markov.

Définition 8.1.5. *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.

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 8.1.2. *On positionne un cavalier sur une des cases d'un échiquier (Fig. 8.2). 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, et on développera (Théorème 8.3.2 et exercices) des méthodes permettant de répondre très facilement à ce type de question.*

Exemple 8.1.3. ***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. 8.3). Leur modèle, très simplifié, de cette situation peut être formulé*

1. Paul Ehrenfest (1880, Vienne – 1933, Amsterdam), physicien théoricien autrichien.
2. Tatiana Alexeyevna Afanaseva (1876, Kiev – 1964, Leiden), mathématicienne russe et danoise.

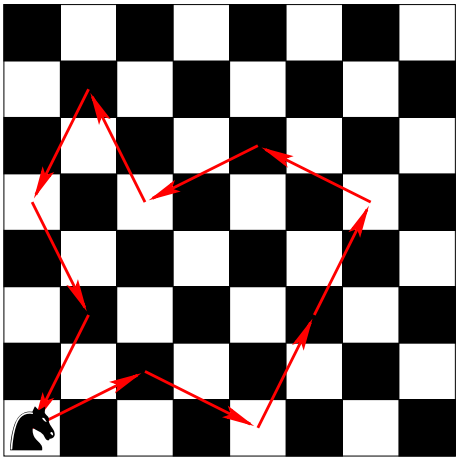


FIGURE 8.2: Quel est le nombre moyen de pas nécessaires pour que le cavalier se déplaçant au hasard sur l'échiquier se retrouve à son point de départ ?

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 8.4.* X *est clairement irréductible.*

Exemple 8.1.4 (Modèle du votant)**.** *Le type de modèle que nous allons considérer à présent a été utilisé entre autres en génétique. Il 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. 8.5). 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*.*

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.

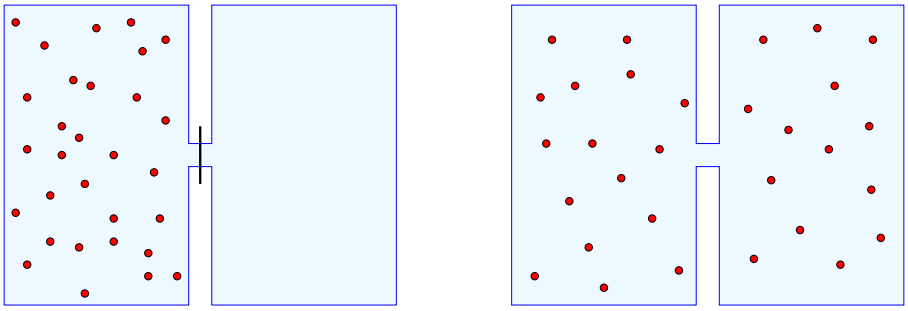


FIGURE 8.3: Au début de l'expérience, toutes les molécules du gaz sont confinées dans le récipient de gauche. Lorsque l'on retire la paroi séparant les deux récipients, les molécules se répartissent uniformément dans tout le volume disponible. Comment une telle irréversibilité peut-elle être compatible avec la réversibilité des équations d'évolution microscopiques ?

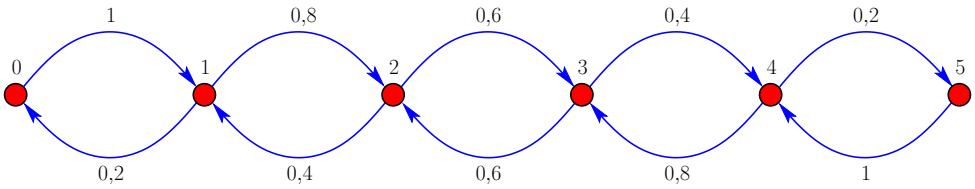


FIGURE 8.4: La représentation graphique du modèle des urnes d'Ehrenfest.

La figure 8.6 montre l'état initial de la chaîne, et deux états ultérieurs. Nous démontrerons plus tard qu'à chaque instant, la probabilité que la chaîne soit 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.

8.2 Chaînes de Markov absorbantes

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 il y a r états absorbants, $\mathbf{Q}$ 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 8.2.1. *Soit* $\mathbf{P}$ *une matrice de transition sous sa forme canonique. Alors, pour*

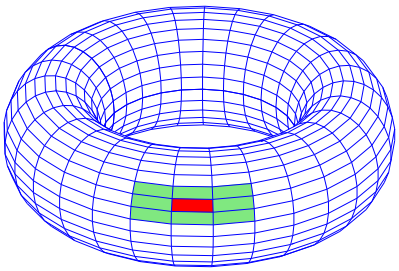


FIGURE 8.5: Une grille 30×30 enroulée en un tore. Chaque case possède 8 voisines.

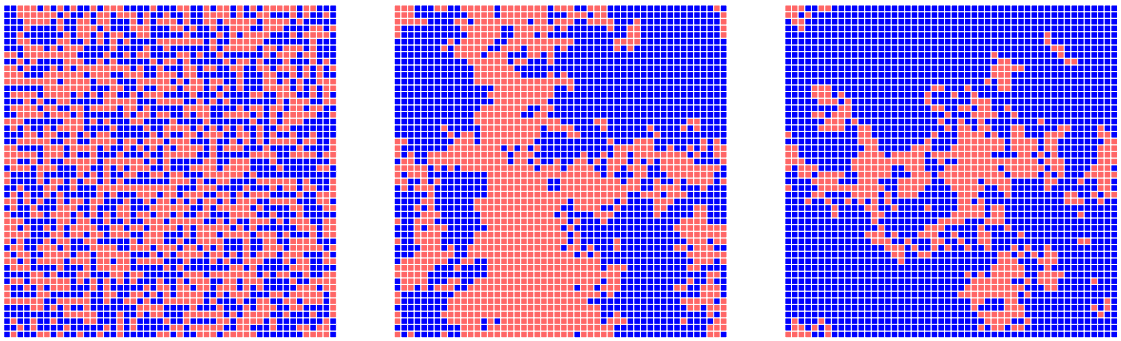


FIGURE 8.6: Le modèle du votant (Exemple 8.1.4), pour $k = 2$, sur une grille 50×50 (représentée « à plat »). Gauche : état initial ; milieu : après 1000000 de pas ; droite : après 10000000 de pas.

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 8.2.1. *Soit* $\mathbf{P}$ *une matrice de transition mise sous forme canonique. Alors,*

$$\lim_{n \to \infty} \mathbf{Q}^n = \mathbf{0}.$$

Démonstration. Soit $\mathcal{A}$ l'ensemble des états absorbants, et $i, j \notin \mathcal{A}$. On a

$$(\mathbf{Q}^n)_{ij} = \mathbb{P}_i(X_n = j) \leq \mathbb{P}_i(X_n \notin \mathcal{A}).$$

Soient $M = \max_{i \in S} \min\{k : \mathbb{P}_i(X_k \in \mathcal{A}) > 0\}$ et

$$p = \min_{i \in S} \mathbb{P}_i(X_M \in \mathcal{A}) (> 0).$$

On a alors,

$$\max_{i \in S} \mathbb{P}_i(X_M \notin \mathcal{A}) = 1 - p,$$

et par conséquent, on déduit de la propriété de Markov que

$$\max_{i \in S} \mathbb{P}_i(X_n \notin \mathcal{A}) \leq \left(\max_{i \in S} \mathbb{P}_i(X_M \notin \mathcal{A}) \right)^{\lfloor \frac{n}{M} \rfloor} = (1-p)^{\lfloor \frac{n}{M} \rfloor},$$

et le résultat suit en prenant la limite $n \to \infty$. □

Corollaire 8.2.1. *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 8.2.1 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 8.2.1. *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 8.2.1. *Soit* $\mathbf{N}$ *la matrice fondamentale de la chaîne,* $\mathcal{A}$ *l'ensemble des états absorbants, et* $\tau = \min\{n \geq 0 : X_n \in \mathcal{A}\}$. *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. $\mathbb{E}_i(\tau) = \sum_{j\notin\mathcal{A}} \mathbf{N}_{ij}$, *pour tout* $i \notin \mathcal{A}$;
3. $\mathbb{P}_i(X_\tau = j) = (\mathbf{NR})_{ij}$, *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,

$$\mathbb{E}_i(\tau) = \mathbb{E}_i(\sum_{n\geq 0} \mathbf{1}_{\{X_n\notin\mathcal{A}\}}) = \sum_{j\notin\mathcal{A}} \mathbb{E}_i(\sum_{n\geq 0} \mathbf{1}_{\{X_n=j\}}) = \sum_{j\notin\mathcal{A}} \mathbf{N}_{ij}.$$

3. On a, pour tout $i \notin \mathcal{A}$ et $j \in \mathcal{A}$,

$$\begin{aligned}
\mathbb{P}_i(X_\tau = j) &= \sum_{n\geq 1} \mathbb{P}_i(X_n = j, X_{n-1} \notin \mathcal{A}) \\
&= \sum_{n\geq 1} \sum_{k\notin\mathcal{A}} \mathbb{P}_i(X_n = j, X_{n-1} = k) \\
&= \sum_{n\geq 1} \sum_{k\notin\mathcal{A}} \mathbb{P}(X_n = j \,|\, X_{n-1} = k) \mathbb{P}_i(X_{n-1} = k) \\
&= \sum_{n\geq 1} \sum_{k\notin\mathcal{A}} \mathbf{R}_{kj} (\mathbf{Q}^{n-1})_{ik} \\
&= \sum_{n\geq 1} (\mathbf{Q}^{n-1}\mathbf{R})_{ij} \\
&= (\mathbf{NR})_{ij}.
\end{aligned}$$

□

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 8.2.2. *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, $\mathbf{f} = \mathbf{P}\mathbf{f}$, où $\mathbf{f} = (f(i))_{i\in S}$.

Théorème 8.2.2. *Soient $(X_n)_{n\geq 0}$ une chaîne de Markov absorbante de matrice de transition $\mathbf{P}$, τ le temps d'absorption, et $\mathcal{A}$ l'ensemble des états absorbants. Alors, pour toute fonction f $\mathbf{P}$-harmonique, et tout $i \in S$,*

$$f(i) = \sum_{j\in\mathcal{A}} f(j)\mathbb{P}_i(X_\tau = j).$$

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 8.2.1. □

Ce théorème peut se révéler particulièrement utile, dans certaines circonstances.

Exemple 8.2.1. *Retournons au modèle introduit dans l'Exemple 8.1.4. 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,

$$\sum_{j\in S} p(i,j)\big(N_1(j) - N_1(i)\big) = \sum_{\substack{(x,y)\\ voisins}} \frac{1}{8n^2}\left(\mathbf{1}_{\{i(x)\neq 1, i(y)=1\}} - \mathbf{1}_{\{i(x)=1, i(y)\neq 1\}}\right),$$

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 τ 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

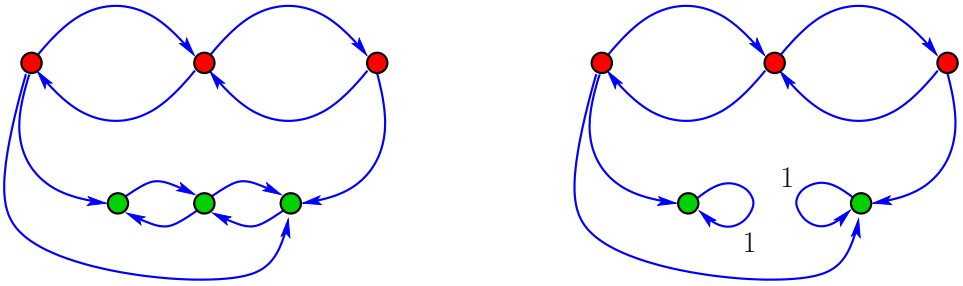


FIGURE 8.7: Une fois la chaîne entrée dans un des états représentés en vert, elle ne peut plus retourner vers les états représentés en rouge (gauche). Ce type de chaîne peut être étudié de la même façon que les chaînes absorbantes, en rendant les points d'entrée de la sous-chaîne absorbant (droite).

de cases de couleur 1 *dans l'état initial* i_0 *soit égale à* ρ. *Le théorème précédent implique donc que*

$$\rho = f(i_0) = \sum_{\ell=1}^{k} f(a_\ell)\mathbb{P}_{i_0}(X_\tau = a_\ell).$$

Or, $f(a_1) = 1$ *(puisque toutes les cases de* a_1 *sont de couleur* 1*), et* $f(a_\ell) = 0$ *pour* $\ell = 2, \ldots, k$. *On a donc*

$$\rho = \mathbb{P}_{i_0}(X_\tau = a_1).$$

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 8.2.1.*

Une remarque s'impose avant de conclure cette section. Considérons la chaîne de Markov de la Figure 8.7 (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 8.7 (droite), et celle-ci est absorbante.

8.3 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 8.3.1.

– *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 8.3.1. *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 $\sum_n p_n(i,j) < \infty$, $\forall i \in S$.*

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 7.2.12, 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 8.3.2. *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 8.3.1

$$1 = \lim_{n\to\infty} \sum_{j\in S} p_n(i,j) = \sum_{j\in S} \lim_{n\to\infty} p_n(i,j) = 0,$$

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 8.3.3. *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

$$\begin{aligned}\mathbb{P}_i(X_n \neq i, \forall n \geq 1) &\geq \mathbb{P}_i(X_n \neq i, \forall n \geq 1, X_{n(i,j)} = j)\\ &= \mathbb{P}_i(X_n \neq i, \forall n > m, X_{n(i,j)} = j)\\ &= p_{n(i,j)}(i,j)\,\mathbb{P}_j(X_n \neq i, \forall n \geq 1).\end{aligned}$$

On sait du Lemme 8.3.2 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 $\mathbb{P}_j(X_n \neq i, \forall n > 1) = 0$. □

8.3.1 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 8.3.2. *Un vecteur $\boldsymbol{\pi} = \big(\pi(i)\big)_{i\in S}$ est appelé distribution stationnaire associé 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 Théorème 8.1.2 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 8.3.1. *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 8.3.2 qu'avec probabilité 1, $T_k<\infty$ et donc $X_0=X_{T_k}=k$. D'autre part, l'événement $\{T_k\geq n\}=\{X_1\neq k, X_2\neq k,\dots,X_{n-1}\neq k\}$ ne dépendant que de $X_1,\dots,X_{n-1}$, la propriété de Markov au temps $n-1$ (et le fait que les deux membres sont nuls lorsque $i=k$) donne

$$\mathbb{P}_k(X_{n-1}=i,X_n=j,T_k\geq n)=\mathbb{P}_k(X_{n-1}=i,T_k\geq n)\,p(i,j),\qquad\forall i,j\in S.$$

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 :

$$\sum_{i\in S}\gamma^k(i)=\sum_{i\in S}\mathbb{E}_k\Big(\sum_{n=0}^{T_k-1}\mathbf{1}_{\{X_n=i\}}\Big)=\mathbb{E}_k\Big(\sum_{n=0}^{T_k-1}\sum_{i\in S}\mathbf{1}_{\{X_n=i\}}\Big)=\mathbb{E}_k\big(T_k\big)=\rho_k.$$

(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

$$\gamma^k(i)=\sum_{j\in S}\gamma^k(j)\,p_{n(k,i)}(j,i)\geq\gamma^k(k)\,p_{n(k,i)}(k,i)=p_{n(k,i)}(k,i)>0.\tag{8.1}$$

□

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 8.3.2. *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 (8.1) 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 8.3.1 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. □

8.3.2 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 8.3.3.
- *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 Théorème 8.3.2 montre qu'une chaîne de Markov X sur S est ergodique si et seulement si elle est irréductible et apériodique.

Lemme 8.3.4. *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 $t_1, t_2, \ldots, t_\ell$ 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.

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.

Théorème 8.3.3. *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 8.3.2 et 8.3.3 impliquent que $\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\mu}}(X_n = i) = 1/\rho_i$, pour tout $i \in S$.

Remarque 8.3.1. *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* $Y^{(r)}$*,* $0 \le r < d$*, 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 $n \ge N$ (voir le Lemme 8.3.4).

Notons $\mathbb{P}_{(i,j)}$ la loi de la chaîne Z partant de $Z_0 = (i,j)$. Fixons $s \in S$, et introduisons $T = \min\{n \ge 1 : Z_n = (s,s)\}$. Z étant irréductible, $\mathbb{P}_{(i,j)}(T < \infty) = 1$, pour tout $i, j \in S$. L'observation cruciale est que, pour tout $m \ge 0$, 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

$$\begin{aligned} p_n(i,k) &= \mathbb{P}_{(i,j)}(X_n = k) \\ &= \mathbb{P}_{(i,j)}(X_n = k, T \le n) + \mathbb{P}_{(i,j)}(X_n = k, T > n) \\ &= \mathbb{P}_{(i,j)}(Y_n = k, T \le n) + \mathbb{P}_{(i,j)}(X_n = k, T > n) \\ &\le \mathbb{P}_{(i,j)}(Y_n = k) + \mathbb{P}_{(i,j)}(T > n) \\ &= p_n(j,k) + \mathbb{P}_{(i,j)}(T > n). \end{aligned}$$

On obtient donc

$$|p_n(i,k) - p_n(j,k)| \le \mathbb{P}_{(i,j)}(T > n) \overset{n\to\infty}{\longrightarrow} 0,$$

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

$$\lim_{n\to\infty}\Big|\sum_{j\in S}\mu(j)p_n(j,i)-\pi(i)\Big| = \lim_{n\to\infty}\Big|\sum_{j\in S}\mu(j)(p_n(j,i)-\pi(i))\Big|$$
$$\leq \sum_{j\in S}\mu(j)\lim_{n\to\infty}|p_n(j,i)-\pi(i)| = 0.$$

□

8.3.3 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 $\boldsymbol{\pi}$ pour tout $n \in \mathbb{Z}$. On définit la chaîne renversée Y par

$$Y_n = X_{-n}, n \in \mathbb{Z}.$$

Définition 8.3.4. *La chaîne X est réversible (à l'équilibre) si les matrices de transition de X et Y sont identiques.*

Théorème 8.3.4. *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 8.3.5. *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 8.3.1. *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 8.3.3 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 9

Modèle de percolation

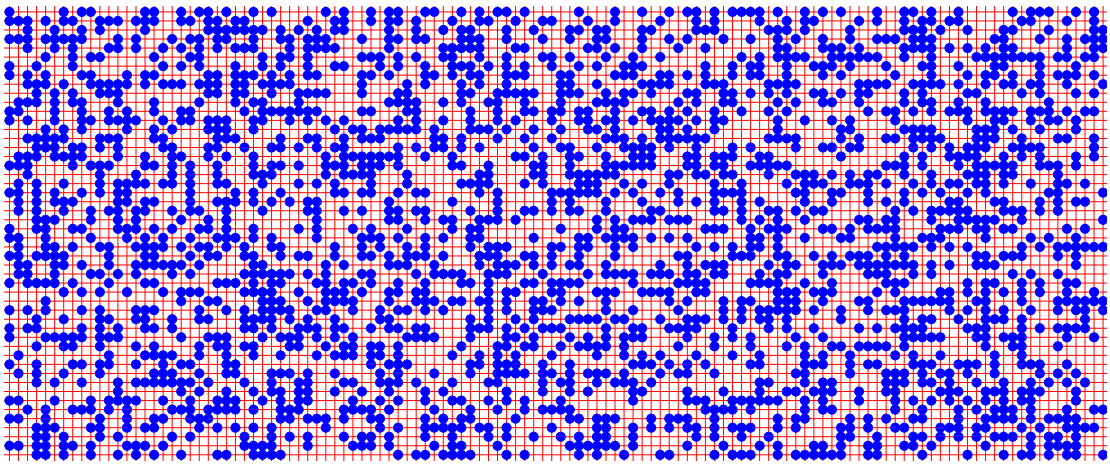
Dans ce chapitre, nous allons introduire un autre processus très important en théorie des probabilités : le modèle de percolation. Contrairement à la marche aléatoire et aux chaînes de Markov, il ne s'agit plus d'une famille de variables aléatoires indicées par un paramètre que l'on peut interpréter 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 $\mathbb{Z}^2$.

9.1 Définition

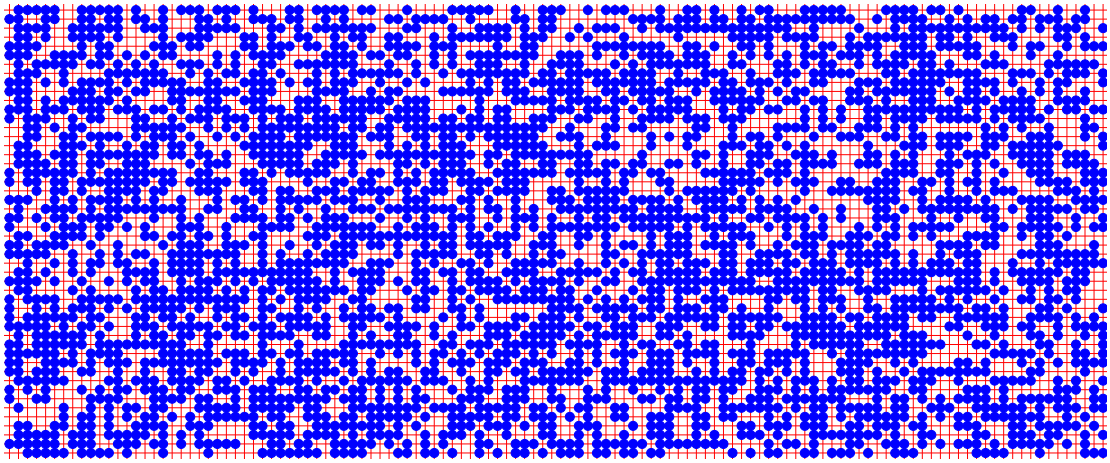
Soit $p \in [0,1]$, et soit $(X_i)_{i\in\mathbb{Z}^2}$ une famille de variables aléatoires indépendantes suivant une loi de Bernoulli de paramètre p, indicées par les sommets de $\mathbb{Z}^2$. On note $\mathbb{P}_p$ la loi de ce champ.

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 9.1 trois réalisations de ce processus pour des valeurs diverses de p.

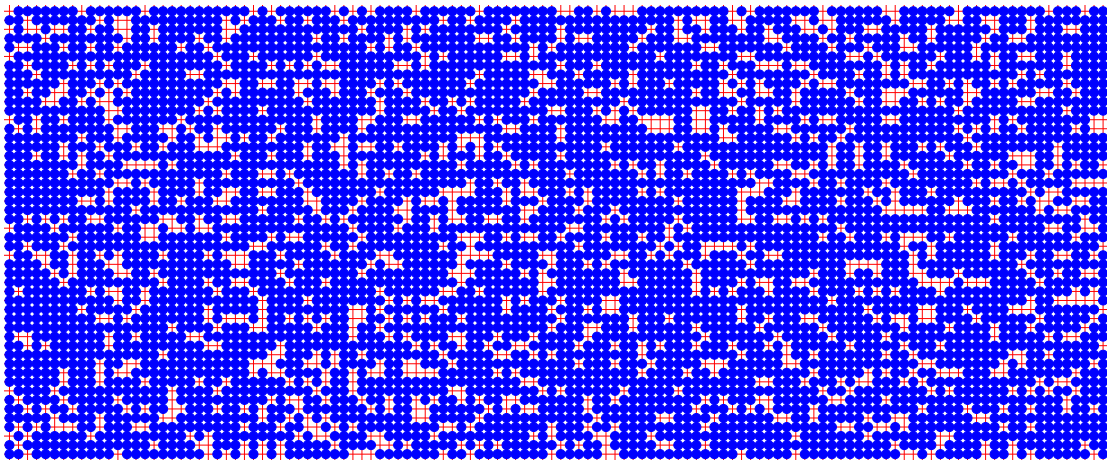
L'interprétation originelle de ce processus est comme modèle d'un matériau poreux. Un tel matériau contient un grand nombre de trous microscopiques. La question de base que l'on se pose alors est si cette porosité locale induit une porosité globale : si l'on plonge une pierre poreuse dans de l'eau, quelle 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.



$p = 0,4$



$p = 0,6$



$p = 0,8$

FIGURE 9.1: Trois réalisations du processus de percolation.

9.2 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.

Théorème 9.2.1. *1. Il existe $0 < p_c < 1$ tel que*

$$\begin{aligned} \Theta(p) &= 0 \qquad \forall p < p_c, \\ \Theta(p) &> 0 \qquad \forall p > p_c. \end{aligned}$$

2. La probabilité qu'il existe (au moins) un amas infini est égale à 1 si $\Theta(p) > 0$, et 0 sinon.

Remarque 9.2.1. *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,

$$\mathbb{P}_p\big(\bigcup_{i\in\mathbb{Z}^2}\{|C(i)| = \infty\}\big) \geq \mathbb{P}_p(\{|C(0)| = \infty\}) > 0,$$

si $\Theta(p) > 0$, et donc $\mathbb{P}_p(\bigcup_{i\in\mathbb{Z}^2}\{|C(i)| = \infty\}) = 1$. Réciproquement,

$$\mathbb{P}_p\big(\bigcup_{i\in\mathbb{Z}^2}\{|C(i)| = \infty\}\big) \leq \sum_{i\in\mathbb{Z}^2}\mathbb{P}_p(\{|C(i)| = \infty\}),$$

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

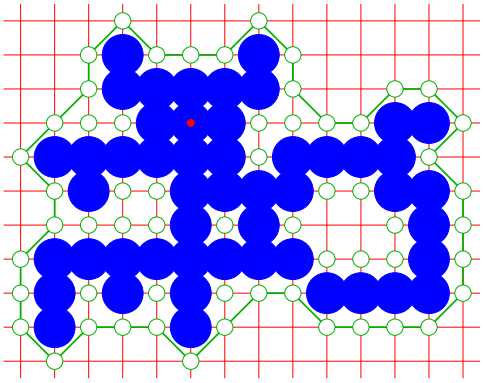


FIGURE 9.2: Lorsque $X_0 = 1$ mais que l'amas contenant l'origine est fini, il y a toujours un $\star$-circuit vide entourant l'origine (le point rouge). On a représenté par des cercles verts les sommets qui sont nécessairement vides si $C(0)$ est l'amas représenté en bleu.

le constituant soient tous occupés est exactement donnée par $\prod_{i\in\gamma} \mathbb{P}_p(X_i = 1) = p^n$. Par conséquent,

$$\mathbb{E}_p(N_{\mathrm{o}}(n)) = \mathbb{E}_p\Big(\sum_{\gamma\in\mathcal{N}(n)} \mathbf{1}_{\{\gamma \text{ occupé}\}}\Big) = \sum_{\gamma\in\mathcal{N}(n)} \mathbb{P}_p(\gamma \text{ occupé}) = p^n N(n) \leq (4p)^n.$$

Lorsque l'événement $\{|C(0)| = \infty\}$ est réalisé, il existe de tels chemins occupés de toutes les longueurs. On obtient donc, pour tout $n \geq 1$,

$$\mathbb{P}_p(|C(0)| = \infty) \leq \mathbb{P}_p(N_{\mathrm{o}}(n) \geq 1) \leq \mathbb{E}_p(N_{\mathrm{o}}(n)) \leq (4p)^n,$$

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 de sommets $i_1, \ldots, i_n$ tous distincts tels que $\|i_k - i_{k-1}\|_2 \leq \sqrt{2}$, $k = 2, \ldots, n$, 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 9.2). 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,

$$\mathbb{E}_p(N^\star_{\mathrm{v}}(n)) = (1-p)^n N^\star(n) < \tfrac{1}{2}n(8(1-p))^n.$$

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.

Comme on l'a vu, lorsque $X_0 = 1$ et $|C(0)| < \infty$, il exite un entier n tel que $N_{\mathrm{v}}^{\star}(n) \geq 1$. À présent,

$$\begin{aligned}\mathbb{P}_p(X_0 = 1, |C(0)| < \infty) &\leq \mathbb{P}_p\big(\bigcup_{n\geq 4}\{N_{\mathrm{v}}^{\star}(n) \geq 1\}\big) \leq \sum_{n\geq 4}\mathbb{P}_p\big(N_{\mathrm{v}}^{\star}(n) \geq 1\big)\\ &\leq \sum_{n\geq 4}\mathbb{E}_p(N_{\mathrm{v}}^{\star}(n)) \leq \sum_{n\geq 4}\tfrac{1}{2}n(8(1-p))^n,\end{aligned}$$

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 10

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.

10.1 Définition et propriétés élémentaires

Il y a trois façons naturelles de décrire un tel processus (*cf.* Fig 10.1) :

- 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 10.1.1. *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.

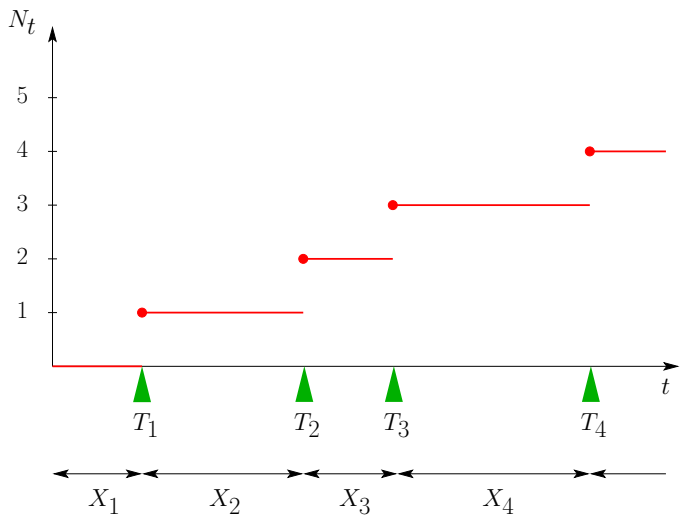


FIGURE 10.1: Une réalisation d'un processus de Poisson.

Remarque 10.1.1. *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 $(N_t)_{t\geq 1}$.

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 10.1.2. *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 10.1.3. *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 10.1.1. *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. *$\forall\ 0 < t_1 < t_2 < \cdots < t_n$, les variables aléatoires $(N_{t_{i+1}} - N_{t_i})_{i=1,\ldots,n-1}$ sont indépendantes.*

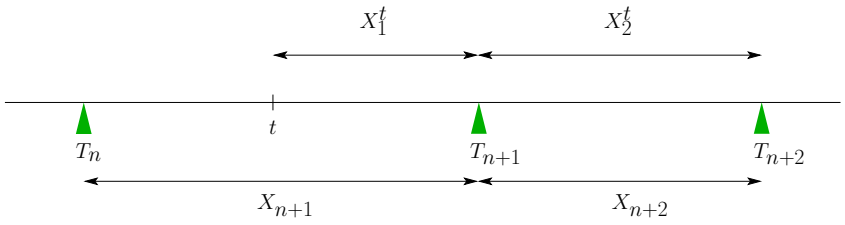


FIGURE 10.2: Définition des variables aléatoires X_k^t (lorsque $N_t = n$).

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 $x > 0$,

$$\begin{aligned}
\mathbb{P}_\lambda(X_1^t > x \,|\, N_t = n)\mathbb{P}_\lambda(N_t = n) &= \mathbb{P}_\lambda(X_1^t > x, T_n \leq t < T_{n+1}) \\
&= \mathbb{P}_\lambda(T_n \leq t, X_{n+1} > t + x - T_n) \\
&= \int_0^t \mathrm{d}y \int_{t+x-y}^\infty \mathrm{d}z \, f_{(T_n, X_{n+1})}(y, z) \\
&= \int_0^t \mathrm{d}y \, f_{T_n}(y) \int_{t+x-y}^\infty \mathrm{d}z \, f_{X_{n+1}}(z) \\
&= \int_0^t \mathbb{P}_\lambda(X_{n+1} > t + x - y) f_{T_n}(y) \mathrm{d}y \\
&= e^{-\lambda x} \int_0^t \mathbb{P}_\lambda(X_{n+1} > t - y) f_{T_n}(y) \mathrm{d}y \\
&= e^{-\lambda x} \, \mathbb{P}_\lambda(T_n \leq t, X_{n+1} > t - T_n) \\
&= e^{-\lambda x} \, \mathbb{P}_\lambda(N_t = n). \qquad (10.1)
\end{aligned}$$

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 (10.1). 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 $\mathsf{exp}(\lambda)$. 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

$$\begin{aligned}
&\mathbb{P}_\lambda(N_{t_{i+1}} - N_{t_i} = k_i, i = 1, \ldots, n-1) \\
&\qquad = \sum_{\ell\geq 0} \mathbb{P}_\lambda(N_{t_{i+1}} - N_{t_i} = k_i, i = 1, \ldots, n-1 \,|\, N_{t_1} = \ell)\, \mathbb{P}_\lambda(N_{t_1} = \ell) \\
&\qquad = \sum_{\ell\geq 0} \mathbb{P}_\lambda(N_{t_{i+1}} - N_{t_1} = m_i, i = 1, \ldots, n-1 \,|\, N_{t_1} = \ell)\, \mathbb{P}_\lambda(N_{t_1} = \ell) \\
&\qquad = \sum_{\ell\geq 0} \mathbb{P}_\lambda(N_{t_{i+1}-t_1} = m_i, i = 1, \ldots, n-1)\, \mathbb{P}_\lambda(N_{t_1} = \ell) \\
&\qquad = \mathbb{P}_\lambda(N_{t_{i+1}-t_1} = m_i, i = 1, \ldots, n-1) \\
&\qquad = \mathbb{P}_\lambda(N_{t_2-t_1} = k_1, N_{t_{i+1}-t_1} - N_{t_2-t_1} = m_i - m_1, i = 2, \ldots, n-1).
\end{aligned}$$

De la même façon, puisque $t_{i+1} - t_1 - (t_2 - t_1) = t_{i+1} - t_2$,

$$\begin{aligned}
&\mathbb{P}_\lambda(N_{t_2-t_1} = k_1, N_{t_{i+1}-t_1} - N_{t_2-t_1} = m_i - m_1, i = 2, \ldots, n-1) \\
&\qquad = \mathbb{P}_\lambda(N_{t_{i+1}-t_1} - N_{t_2-t_1} = m_i - m_1, i = 2, \ldots, n-1 \,|\, N_{t_2-t_1} = k_1) \\
&\qquad\qquad \times \mathbb{P}_\lambda(N_{t_2-t_1} = k_1) \\
&\qquad = \mathbb{P}_\lambda(N_{t_{i+1}-t_2} = m_i - m_1, i = 2, \ldots, n-1)\, \mathbb{P}_\lambda(N_{t_2-t_1} = k_1). \qquad (10.2)
\end{aligned}$$

Mais

$$\begin{aligned}
&\mathbb{P}_\lambda(N_{t_{i+1}-t_2} = m_i - m_1, i = 2, \ldots, n-1) \\
&\qquad = \mathbb{P}_\lambda(N_{t_3-t_2} = k_2, N_{t_{i+1}-t_2} - N_{t_3-t_2} = m_i - m_2, i = 3, \ldots, n-1),
\end{aligned}$$

et l'on peut donc répéter la procédure (10.2), pour obtenir finalement

$$\mathbb{P}_\lambda(N_{t_{i+1}} - N_{t_i} = k_i, i = 1, \ldots, n-1) = \prod_{i=1}^{n-1} \mathbb{P}_\lambda(N_{t_{i+1}-t_i} = k_i) = \prod_{i=1}^{n-1} \mathbb{P}_\lambda(N_{t_{i+1}} - N_{t_i} = k_i),$$

la dernière identité résultant de la première partie du théorème. □

Lemme 10.1.1. *Soit $(N_t)_{t\geq 0}$ un processus de Poisson d'intensité λ. Alors, T_n suit une loi* $\mathsf{gamma}(\lambda, n)$,

$$f_{T_n}(x) = \frac{1}{(n-1)!}\lambda^n x^{n-1} e^{-\lambda x}\mathbf{1}_{[0,\infty)}(x).$$

Démonstration. T_n est une somme de n variables aléatoires i.i.d. de loi $\mathsf{exp}(\lambda)$. Manifestement, T_1 suit une loi $\mathsf{exp}(\lambda)$, et celle-ci coïncide avec la loi $\mathsf{gamma}(\lambda, 1)$. On procède par récurrence. Supposons l'énoncé vrai pour T_n. On a alors,

$$\begin{aligned} f_{T_{n+1}}(x) = f_{T_n+X_{n+1}}(x) &= \int_{-\infty}^{\infty} f_{T_n}(u) f_{X_{n+1}}(x-u)\mathrm{d}u \\ &= \int_0^x \frac{1}{(n-1)!}\lambda^n u^{n-1} e^{-\lambda u}\,\lambda e^{-\lambda(x-u)}\mathrm{d}u \\ &= \frac{\lambda^{n+1}}{(n-1)!} e^{-\lambda x}\int_0^x u^{n-1}\mathrm{d}u \\ &= \frac{\lambda^{n+1}}{n!} e^{-\lambda x} x^n, \end{aligned}$$

et le lemme est démontré. □

Théorème 10.1.2. *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* $\mathsf{poisson}(\lambda(t-s))$.

Démonstration. Il suit du Théorème 10.1.1 qu'il suffit de considérer le cas $s = 0$. Puisque $N_t = n \iff T_n \leq t < T_{n+1}$, on a immédiatement

$$\begin{aligned} \mathbb{P}_\lambda(N_t = n) = \mathbb{P}_\lambda(T_n \leq t < T_{n+1}) &= \mathbb{P}_\lambda(T_{n+1} > t) - \mathbb{P}_\lambda(T_n > t) \\ &= \frac{\lambda^n}{n!}\int_t^\infty \big(x^n \lambda e^{-\lambda x} - n x^{n-1} e^{-\lambda x}\big)\mathrm{d}x \\ &= \frac{\lambda^n}{n!}\int_t^\infty \frac{\mathrm{d}}{\mathrm{d}x}\big(-x^n e^{-\lambda x}\big)\mathrm{d}x \\ &= \frac{\lambda^n}{n!}\, t^n e^{-\lambda t}. \end{aligned}$$

□

Définition 10.1.4. *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 10.1.1 et 10.1.2 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 10.1.3. *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* $\mathsf{poisson}(\lambda s)$.

Remarque 10.1.2. *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

$$\begin{aligned}
&\mathbb{P}(T_1 \in (s_1, t_1], \ldots, T_n \in (s_n, t_n]) \\
&\qquad = \prod_{i=1}^{n} \mathbb{P}(N_{s_i} - N_{t_{i-1}} = 0) \prod_{i=1}^{n-1} \mathbb{P}(N_{t_i} - N_{s_i} = 1)\mathbb{P}(N_{t_n} - N_{s_n} \geq 1) \\
&\qquad = \prod_{i=1}^{n} e^{-\lambda(s_i - t_{i-1})} \prod_{i=1}^{n-1} \lambda(t_i - s_i) e^{-\lambda(t_i - s_i)} (1 - e^{-\lambda(t_n - s_n)}) \\
&\qquad = \lambda^{n-1}(e^{-\lambda s_n} - e^{-\lambda t_n}) \prod_{i=1}^{n-1} (t_i - s_i) \\
&\qquad = \int_{s_1}^{t_1} \cdots \int_{s_n}^{t_n} \lambda^n e^{-\lambda u_n} \mathrm{d}u_n \cdots \mathrm{d}u_1.
\end{aligned}$$

La loi conjointe de $(T_1, \ldots, T_n)$ possède donc la densité

$$f_{(T_1,\ldots,T_n)}(u_1, \ldots, u_n) = \begin{cases} \lambda^n e^{-\lambda u_n} & \text{si } 0 < u_1 < \cdots < u_n, \\ 0 & \text{sinon.} \end{cases}$$

Déterminons à présent la densité de la loi conjointe de $(X_1, \ldots, X_n)$. La fonction de répartition conjointe est donnée par

$$\begin{aligned}
\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) &= \mathbb{P}(T_1 \leq x_1, T_2 - T_1 \leq x_2, \ldots, T_n - T_{n-1} \leq x_n) \\
&= \int_0^{x_1} \int_{u_1}^{u_1 + x_2} \cdots \int_{u_{n-1}}^{u_{n-1} + x_n} f_{(T_1,\ldots,T_n)}(u_1, \ldots, u_n) \mathrm{d}u_n \cdots \mathrm{d}u_1,
\end{aligned}$$

et la densité conjointe est donc donnée par

$$\frac{\partial^n}{\partial x_1 \cdots \partial x_n}\mathbb{P}(X_1 \le x_1, \ldots, X_n \le x_n) = f_{(T_1,\ldots,T_n)}(x_1, x_1+x_2, \ldots, x_1+\cdots+x_n)$$
$$= \lambda^n e^{-\lambda(x_1+\cdots+x_n)}.$$

On reconnaît la densité conjointe de n variables aléatoires i.i.d. de loi $\mathsf{exp}(\lambda)$. □

Nous allons voir à présent une troisième définition du processus, de nature plus dynamique.

Lemme 10.1.2. *Soit $(N_t)_{t\ge 0}$ un processus de Poisson d'intensité λ, et $0 < t_1 < \cdots < t_k$. Alors, pour $0 \le n_1 \le \cdots \le n_k$ des entiers, on a, lorsque $\epsilon \downarrow 0$,*

$$\begin{aligned}\mathbb{P}_\lambda(N_{t_k+\epsilon} - N_{t_k} = 0 \,|\, N_{t_j} = n_j, 1 \le j \le k) &= 1 - \lambda\epsilon + o(\epsilon),\\ \mathbb{P}_\lambda(N_{t_k+\epsilon} - N_{t_k} = 1 \,|\, N_{t_j} = n_j, 1 \le j \le k) &= \lambda\epsilon + o(\epsilon),\\ \mathbb{P}_\lambda(N_{t_k+\epsilon} - N_{t_k} \ge 2 \,|\, N_{t_j} = n_j, 1 \le j \le k) &= o(\epsilon).\end{aligned} \tag{10.3}$$

Démonstration. Posons $n_0 = 0$. Puisque $\{N_{t_j} = n_j, 1 \le j \le k\} = \{N_{t_j} - N_{t_{j-1}} = n_j - n_{j-1}, 1 \le j \le k\}$, 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 $\mathsf{poisson}(\lambda\epsilon)$: 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 10.1.4. *Un processus de comptage est un processus de Poisson d'intensité λ si et seulement s'il satisfait* (10.3).

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 10.1.3. En utilisant l'inégalité $|\mathbb{P}(B) - \mathbb{P}(C)| \le \mathbb{P}(B \triangle C)$, on obtient que [1]

$$|p_n(t) - p_n(s)| \le \mathbb{P}(N_{t_k+s} \ne N_{t_k+t})/\mathbb{P}(A),$$

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.

ce qui montre que $p_n(t)$ est une fonction continue de t, puisque $\lim_{s\to t} \mathbb{P}(N_s \neq N_t) = 0$.

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 $m \leq n$. On a donc

$$\begin{aligned} p_n(t+\epsilon) = p_n(t)\,\mathbb{P}(D_{t+\epsilon} - D_t = 0 \,|\, A, D_t = n) \\ + \mathbf{1}_{\{n\geq 1\}}\, p_{n-1}(t)\,\mathbb{P}(D_{t+\epsilon} - D_t = 1 \,|\, A, D_t = n-1) \\ + \mathbf{1}_{\{n\geq 2\}} \sum_{m=0}^{n-2} p_m(t)\,\mathbb{P}(D_{t+\epsilon} - D_t = n-m \,|\, A, D_t = m). \end{aligned}$$

Par (10.3), on obtient

$$p_n(t+\epsilon) = p_n(t)(1-\lambda\epsilon) + \mathbf{1}_{\{n\geq 1\}}\, p_{n-1}(t)\lambda\epsilon + o(\epsilon).$$

En divisant par ϵ et en prenant la limite $\epsilon \downarrow 0$, on obtient [2]

$$p_n'(t) = -\lambda p_n(t) + \mathbf{1}_{\{n\geq 1\}}\, \lambda p_{n-1}(t), \tag{10.4}$$

avec condition au bord $p_n(0) = \delta_{n0}$.

Il reste à intégrer (10.4). 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 $n \geq 0$. □

10.2 Autres propriétés

10.2.1 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

2. Il y a une petite subtilité ici : *a priori*, la dérivée dans le membre de gauche de (10.4) 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 $t \geq 0$, 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 $G^+(s_0) \leq 0$, puisque s_0 est un maximum, ce qui est une contradiction.

lampe dont on change immédiatement l'ampoule lorsque celle-ci est défaillante ; la durée de vie d'une ampoule est supposée suivre une loi $\mathsf{exp}(\lambda)$. 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 $\mathsf{exp}(\lambda)$; 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 $\mathsf{exp}(\lambda)$. 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 $\mathsf{exp}(\lambda)$.

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.

10.2.2 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 $s \in (0, t]$,

$$\begin{aligned}\mathbb{P}_\lambda(T_1 < s \,|\, N_t = 1) &= \frac{\mathbb{P}_\lambda(T_1 < s, N_t = 1)}{\mathbb{P}_\lambda(N_t = 1)} = \frac{\mathbb{P}_\lambda(N_s = 1, N_t - N_s = 0)}{\mathbb{P}_\lambda(N_t = 1)} \\ &= \frac{(\lambda s e^{-\lambda s})(e^{-\lambda(t-s)})}{\lambda t e^{-\lambda t}} = \frac{s}{t}.\end{aligned}$$

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.

10.2.3 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 10.2.1. *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 10.1.3.

Pour tout $0 < s < t$ et $n \geq 0$, 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 de N_t sont stationnaires et suivent une loi de Poisson de paramètre λ. 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 10.2.1. *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 10.2.2. *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

$$\begin{aligned}
\mathbb{P}(N_t^{(1)} - N_s^{(1)} = k) &= \sum_{n=k}^{\infty} \mathbb{P}(N_t - N_s = n) \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{n=k}^{\infty} \frac{\lambda^n (t-s)^n}{n!} e^{-\lambda(t-s)} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{(\lambda p\,(t-s))^k}{k!} e^{-\lambda(t-s)} \sum_{n \geq k} \frac{(\lambda(t-s)(1-p))^{n-k}}{(n-k)!} \\
&= \frac{(\lambda p\,(t-s))^k}{k!} e^{-\lambda p\,(t-s)}.
\end{aligned}$$

On montre de la même façon que $N_t^{(2)} - N_s^{(2)}$ est $\mathsf{poisson}(\lambda(1-p))$.

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

$$\begin{aligned}
&\mathbb{P}(\Delta_i^{(1)} = n_i, \Delta_i^{(2)} = m_i, 1 \leq i \leq n) \\
&\quad = \mathbb{P}(\Delta_i^{(1)} = n_i, 1 \leq i \leq n \,|\, \Delta_i = m_i + n_i, 1 \leq i \leq n) \mathbb{P}(\Delta_i = m_i + n_i, 1 \leq i \leq n) \\
&\quad = \prod_{i=1}^{n} \binom{n_i + m_i}{n_i} p^{n_i} (1-p)^{m_i} \prod_{i=1}^{n} \mathbb{P}(\Delta_i = m_i + n_i) \\
&\quad = \prod_{i=1}^{n} \binom{n_i + m_i}{n_i} p^{n_i} (1-p)^{m_i} \prod_{i=1}^{n} \frac{(\lambda(t_i - s_i))^{n_i + m_i}}{(n_i + m_i)!} e^{-\lambda(t_i - s_i)} \\
&\quad = \prod_{i=1}^{n} \frac{(\lambda p\,(t_i - s_i))^{n_i}}{n_i!} e^{-\lambda p (t_i - s_i)} \frac{(\lambda(1-p)\,(t_i - s_i))^{m_i}}{m_i!} e^{-\lambda(1-p)(t_i - s_i)} \\
&\quad = \prod_{i=1}^{n} \mathbb{P}(\Delta_i^{(1)} = n_i)\, \mathbb{P}(\Delta_i^{(2)} = m_i),
\end{aligned}$$

et les processus $(N_t^{(1)})_{t \geq 0}$ et $(N_t^{(2)})_{t \geq 0}$ sont donc à accroissements indépendants, et sont indépendants l'un de l'autre. □

Définition 10.2.2. *On dit que les processus $(N_t^{(1)})_{t \geq 0}$ et $(N_t^{(2)})_{t \geq 0}$ ci-dessus sont des* amincissements *du processus $(N_t)_{t \geq 0}$.*

Remarque 10.2.1. *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 $(N_t^{(i)})_{t \geq 0}$.*

Exemple 10.2.1. *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* $\mathsf{exp}(\lambda_1)$ *et* $\mathsf{exp}(\lambda_2)$ *respectivement. On désire déterminer la probabilité que la première caissière ait fini de*

s'occuper de son $n^{ème}$ client avant que la seconde ait fini de s'occuper de son $m^{ème}$ 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 $\mathsf{gamma}(\lambda_1, n)$ *et* $\mathsf{gamma}(\lambda_2, m)$ *respectivement, et faire un calcul laborieux. Nous allons à la place utiliser les résultats de cette section. Soit $(N_t)_{t\geq 0}$ un processus de Poisson de paramètre $\lambda = \lambda_1 + \lambda_2$. On a vu que les processus $(N_t^{(1)})_{t\geq 0}$ et $(N_t^{(2)})_{t\geq 0}$ peuvent être obtenus en coloriant les points de $(N_t)_{t\geq 0}$ 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*

$$\mathbb{P}(T_n^{(1)} < T_m^{(2)}) = \sum_{k=n}^{n+m-1} \binom{n+m-1}{k} \Big(\frac{\lambda_1}{\lambda_1+\lambda_2}\Big)^k \Big(\frac{\lambda_2}{\lambda_1+\lambda_2}\Big)^{n+m-1-k}.$$

10.2.4 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 10.2.3. *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 10.2.3. *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 10.2.4. *La fonction $m(t)$ dans le Théorème 10.2.3 est appelée* fonction de valeur moyenne *du processus.*

Démonstration. La preuve est semblable à celle du Théorème 10.1.4, 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

$$p_0(s,t) = \exp(-\int_s^t \lambda(u)\mathrm{d}u) = e^{-(m(t)-m(s))}, \qquad s,t \geq 0.$$

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

$$\frac{\partial}{\partial t}p_1(s,t) = \left(e^{-(m(t)-m(s))} - p_1(s,t)\right)\frac{\partial}{\partial t}\big(m(t)-m(s)\big).$$

On voit alors facilement que la solution est donnée par

$$p_1(s,t) = e^{-(m(t)-m(s))}\big(m(t)-m(s)\big), \qquad s,t \geq 0.$$

Par récurrence, on montre ensuite que

$$p_n(s,t) = e^{-(m(t)-m(s))}\frac{1}{n!}\big(m(t)-m(s)\big)^n, \qquad s,t \geq 0, n \in \mathbb{N}.$$

□

10.2.5 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.

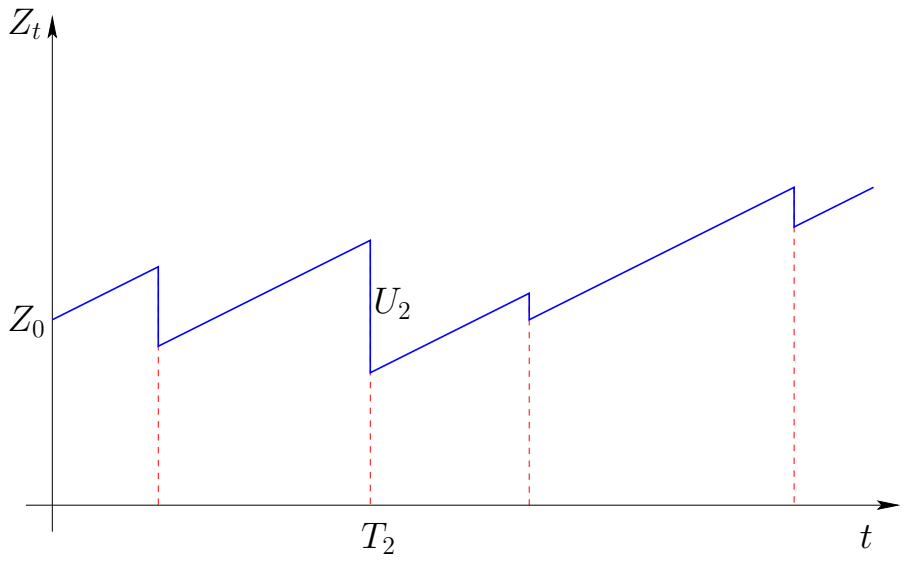


FIGURE 10.3: Évolution des réserves d'une compagnie d'assurance.

Définition 10.2.5. *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 10.2.2. *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 $n^{ème}$ 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.

10.2.6 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 10.2.6. *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 $B \in \mathcal{B}(\mathbb{R}^d)$;*
- *$N(B_1), \ldots, N(B_n)$ sont indépendantes lorsque $B_1, \ldots, B_n$ 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 10.2.4. *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 $x_1, \ldots, x_n$ 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 $A : X_1, \ldots, X_n$. Alors, pour $\epsilon > 0$ suffisamment petit,

$$\begin{aligned}
\mathbb{P}\big(X_i \in B_\epsilon(x_i),\, i = 1, \ldots, n \;\big|\; N(A) = n\big) & \\
&= \frac{1}{n!}\mathbb{P}\big(N(B_\epsilon(x_i)) = 1,\, i = 1, \ldots, n \;\big|\; N(A) = n\big) \\
&= \frac{1}{n!}\frac{\mathbb{P}(N(A \setminus \bigcup_{j=1}^n B_\epsilon(x_j)) = 0)\prod_{i=1}^n \mathbb{P}(N(B_\epsilon(x_i)) = 1)}{\mathbb{P}(N(A) = n)} \\
&= \frac{1}{n!}\frac{e^{-\lambda(|A| - n\epsilon^d)}\prod_{i=1}^n \lambda\epsilon^d e^{-\lambda\epsilon^d}}{(\lambda|A|)^n e^{-\lambda|A|}/n!} \\
&= \epsilon^{nd}|A|^{-n}.
\end{aligned}$$

Par conséquent, conditionnellement à $N(A) = n$, la densité conjointe de $(X_1, \ldots, X_n)$ en $(x_1, \ldots, x_n)$ est donnée par

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon^{nd}}\,\mathbb{P}(X_i \in B_\epsilon(x_i),\, i = 1, \ldots n \,|\, N(A) = n) = |A|^{-n},$$

et coïncide donc bien avec la densité conjointe de n points tirés indépendamment uniformément dans A. □

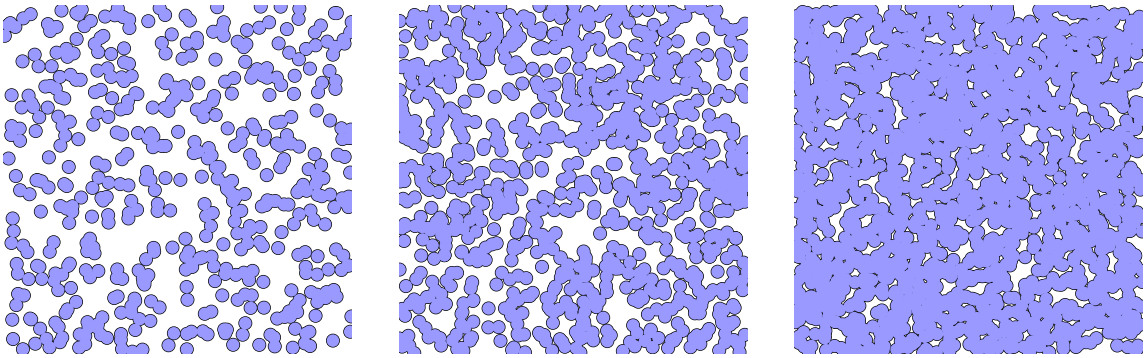


FIGURE 10.4: Trois réalisations du processus booléen de l'Exemple 10.2.3 pour des intensités croissantes du processus de Poisson sous-jacent.

Exemple 10.2.3 (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 $D_r(\mathbf{x}) = \left\{\mathbf{y} \in \mathbb{R}^2 \,:\, \|\mathbf{y}-\mathbf{x}\|_2 \leq r\right\}$ *; c.f. Fig. 10.4. (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 9.*

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,

$$\begin{aligned}\mathbb{P}(\mathbf{x} \notin \overline{\Pi}) &= \mathbb{P}(\textit{Aucun point de } \Pi \textit{ ne se trouve à distance au plus } r \textit{ de } \mathbf{x})\\ &= \mathbb{P}\left(\Pi \cap D_r(\mathbf{x}) = \varnothing\right)\\ &= \mathbb{P}\left(N\big(D_r(\mathbf{x})\big) = 0\right)\\ &= \exp(-\lambda\pi r^2).\end{aligned}$$

Par conséquent, la fraction de A *couverte par les disques est donnée par*

$$\frac{\mathbb{E}(|A \cap \overline{\Pi}|)}{|A|} = 1 - e^{-\lambda\pi r^2}.$$

10.2.7 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.

Il est aisé d'exprimer la loi des temps de renouvellement T_k à partir de celle des temps d'attente.

Lemme 10.2.1. *$f_{T_1}(t) = f(t)$, 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 10.2.2. $\mathbb{P}(N_t = k) = F_{T_k}(t) - F_{T_{k+1}}(t)$.

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 10.2.7. *La fonction de renouvellement est définie par $m(t) = \mathbb{E}(N_t)$.*

Lemme 10.2.3. $m(t) = \sum_{k=1}^{\infty} F_{T_k}(t)$.

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 10.2.4. *La fonction de renouvellement satisfait l'équation de renouvellement,*

$$m(t) = F(t) + \int_0^t m(t-s)f(s)\,\mathrm{d}s, \qquad t \geq 0.$$

3. Mais tout ce que nous dirons ici s'étend à des lois quelconques.

Démonstration. En conditionnant sur X_1, on a

$$m(t) = \mathbb{E}(\mathbb{E}(N_t \,|\, X_1)).$$

À présent, $\mathbb{E}(N_t \,|\, X_1 = x) = 0$ si $t < x$. D'un autre côté,

$$\mathbb{E}(N_t \,|\, X_1 = x) = 1 + \mathbb{E}(N_{t-x}), \qquad \text{si } t \geq x.$$

On en déduit que

$$m(t) = \int_0^\infty \mathbb{E}(N_t \,|\, X_1 = x) f(x)\,\mathrm{d}x = \int_0^t \big(1 + m(t-x)\big) f(x)\,\mathrm{d}x.$$

□

Remarque 10.2.2. *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 10.2.3. *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 10.2.5. *$\frac{1}{t} N_t \overset{\text{p.s.}}{\to} \frac{1}{\mu}$, lorsque $t \to \infty$.*

Démonstration. Puisque $T_{N_t} \leq t < T_{N_t+1}$, on a, lorsque $N_t > 0$,

$$\frac{T_{N_t}}{N_t} \leq \frac{t}{N_t} \leq \frac{T_{N_t+1}}{N_t+1}\big(1 + \frac{1}{N_t}\big).$$

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 10.2.6. *Supposons que* $0 < \sigma^2 = \mathrm{Var}(X_1) < \infty$. *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

$$\mathbb{P}\big(\frac{N_t - (t/\mu)}{\sqrt{t\sigma^2/\mu^3}} \geq x\big) = \mathbb{P}\big(N_t \geq (t/\mu) + x\sqrt{t\sigma^2/\mu^3}\big) = \mathbb{P}\big(T_{a(t)} \leq t\big),$$

où $a(t) = \lceil (t/\mu) + x\sqrt{t\sigma^2/\mu^3} \rceil$. À présent,

$$\mathbb{P}\big(T_{a(t)} \leq t\big) = \mathbb{P}\big(\frac{T_{a(t)} - \mu a(t)}{\sigma\sqrt{a(t)}} \leq \frac{t - \mu a(t)}{\sigma\sqrt{a(t)}}\big).$$

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,

$$\lim_{t\to\infty} \mathbb{P}\big(\frac{N_t - (t/\mu)}{\sqrt{t\sigma^2/\mu^3}} \geq x\big) = \Phi(-x).$$

□

Remarque 10.2.4. *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 11

Éléments de théorie de l'information

Ce chapitre est consacré à un bref aperçu d'un autre domaine dans lequel les probabilités jouent un rôle prépondérant : la théorie de l'information. Née d'un article classique de Claude Shannon [1] en 1948, cette théorie porte sur les systèmes d'information, les systèmes de communication et leur efficacité. Si elle s'est initialement intéressée à établir les limites fondamentales de la compression de données et de leur transmission sans perte, son domaine s'est depuis très fortement élargi, et ses applications pratiques sont très nombreuses.

Dans ce chapitre nous ne ferons que survoler ce sujet, en introduisant quelques concepts fondamentaux, et en démontrant des versions simples de deux résultats célèbres (les théorèmes de Shannon).

11.1 Sources, codages et entropie

On s'intéresse au problème de transmettre une information d'une source vers un destinataire à travers un canal. La source peut être de nature très variée : voix, images, texte, etc. Le canal peut être une ligne téléphonique, une fibre optique, une pellicule photographique, un CD, etc. En général, le canal peut être bruité, produisant des erreurs lors de la transmission de l'information. Afin de limiter la quantité d'information à transmettre (taille d'une photographie numérique, par exemple), ou afin d'introduire une redondance permettant de réparer le message à l'arrivée (dans le cas où celui-ci aurait été endommagé lors de sa transmission), on recourt à un codage de l'information. Ce dernier peut également être utile afin d'adapter le format de la source au canal utilisé, par exemple en transformant un signal analogique en un signal digital.

11.1.1 Codes binaires

On désignera par $\mathfrak{A}$ l'alphabet avec lequel sont construit les messages, et on notera $\{0,1\}^\star = \bigcup_{k\geq 1}\{0,1\}^k$ l'ensemble des mots binaires de longueur finie arbitraire. Commen-

1. Claude Elwood Shannon (1916, Gaylord – 2001, Medford), ingénieur électricien et mathématicien américain.

çons par quelques définitions.

- Un code binaire est une application $c : \mathfrak{A} \to \{0,1\}^\star$, associant à chaque symbole $a \in \mathfrak{A}$ de l'alphabet un mot binaire $c(a)$ de longueur finie, appelé son mot de code.
- Un code binaire non-singulier est un code binaire injectif. En d'autres termes, des symboles différents reçoivent des mots de code différents, et ainsi un mot de code donné est décodable de façon unique.
- Un code préfixe est un code c non-singulier tel qu'aucun mot de code $c(a)$ ne soit le préfixe d'un autre mot de code $c(a')$, où $a, a' \in \mathfrak{A}$, $a' \neq a$.
- Soit $\mathfrak{A}^\star = \bigcup_{L \geq 1} \mathfrak{A}^L$. L'extension $c^\star : \mathfrak{A}^\star \to \{0,1\}^\star$ d'un code binaire c est l'application associant à un message de n symboles $(a_1, \ldots, a_n) \in \mathfrak{A}^n$ le mot de code

$$c^\star(a_1, \ldots, a_n) = c(a_1) \odot \cdots \odot c(a_n),$$

  où $x \odot y$ est la concaténation des mots x et y.
- Un code binaire uniquement décodable est un code c dont l'extension $c^\star$ est non-singulière. Observez que cette propriété est toujours satisfaite lorsque c est un code préfixe.

Un code préfixe possède la propriété désirable d'être instantanément décodable : lors de la réception du message codé, celui-ci peut-être décodé au fur et à mesure de la réception des bits, sans avoir à connaître la suite du message (dès que la suite de bits correspondant à un mot de code apparaît, le symbole correspondant du message initial est retrouvé). On vérifie aisément que cette propriété est équivalente à celle d'être un code préfixe.

Dans la suite, nous nous restreindrons toujours à des codes préfixes.

Exemple 11.1.1. *Un exemple de code bien connu, mais désuet, est le code morse, dans lequel les lettres de l'alphabet sont encodées selon une succession de signaux courts (« · ») ou longs (« − »). Dans ce code, les lettres les plus utilisées en anglais sont représentées par des mots de code courts, par exemple le mot de code associé à la lettre « E » est « · », celui de la lettre « A » est « · − », alors que celui de la lettre « Q » est donné par « − − · − ». Il ne s'agit pas d'un code préfixe, le mot de code de la lettre « E » étant le préfixe de celui de la lettre « A ». Afin de pouvoir séparer les différentes lettres, et ainsi décoder le message, il est nécessaire d'utiliser la longueur des silences séparant deux signaux successifs (par exemple, deux lettres sont séparées par un silence correspondant à trois « · », les mots par des silences correspondant cinq « · »).*

Exemple 11.1.2. *Supposons qu'on ait à transmettre par fax une page contenant du texte manuscrit. Si le texte n'est pas trop dense, une fois la page numérisée, l'essentiel, disons* 99%*, des pixels de l'image obtenue seront blancs (correspondant à la feuille), et une faible fraction noirs (correspondant au texte).*

On peut modéliser une telle situation en supposant que chaque pixel est soit blanc, soit noir, indépendamment avec probabilités $0,99$ *et* $0,01$ *respectivement.*

On considère l'algorithme de codage suivant : on décompose l'image en paquets de 10 *pixels consécutifs. Si tous les pixels sont blancs, on envoie un* 0*, sinon on envoie un* 1 *suivi d'une chaîne de* 10 *bits correspondant aux valeurs des* 10 *pixels. On a donc* $\mathfrak{A} = \{0,1\}^{10}$*,*

$c(0000000000) = 0$ *et, pour* $x \in \mathfrak{A}, x \neq 0000000000$, $c(x) = 1 \odot x$. *Il s'agit clairement d'un code uniquement décodable.*

La longueur moyenne de la chaîne transmise pour un bloc est donc donnée par

$$0,99^{10} \cdot 1 + (1 - 0,99^{10}) \cdot 11 \simeq 1,96.$$

Ce codage réduit donc la taille du message original à moins de 20% *de sa taille initiale.*

Est-il possible de faire mieux, et si oui : quel est le taux de compression maximal possible ?

11.1.2 Longueur de code, entropie

Afin de simplifier l'exposition au maximum, nous nous restreindrons à la discussion de sources discrètes sans mémoire. Nous supposerons l'alphabet $\mathfrak{A}$ fini. Chaque symbole $a \in \mathfrak{A}$ composant le message à coder est tiré au hasard, indépendamment des autres, avec une probabilité $p(a) > 0$. Les probabilités associées aux différents symboles modélisent leur fréquence d'apparition ; bien entendu, on peut considérer des modèles plus réalistes, dans lesquels la fréquence d'apparition d'un symbole dépend des lettres précédentes (par exemple, selon un schéma markovien), mais nous ne le ferons pas ici. Nous nous restreindrons également au cas particulièrement important où le message est encodé sous forme binaire. Une source discrète sans mémoire est donc la donnée d'un espace probabilisé fini $A = (\mathfrak{A}, \mathbb{P})$.

Nous aurons besoin des définitions suivantes.

- La longueur $\ell(a)$ du mot de code $c(a)$ est égal à la longueur du mot $c(a)$, c'est-à-dire $\ell(a) = k$ si et seulement si $c(a) \in \{0,1\}^k$.
- La longueur de code $L[c]$ d'un code binaire c est la longueur moyenne de ses mots de code : $L[c] = \mathbb{E}(\ell) = \sum_{a \in \mathfrak{A}} p(a)\ell(a)$.
- L'information propre d'un symbole $a \in \mathfrak{A}$ est définie par $I(a) = -\log_2 p(a)$.
- L'entropie $H(\mathbb{P})$ d'une source $A = (\mathfrak{A}, \mathbb{P})$ est la valeur moyenne de l'information propre de ses symboles,

$$H(\mathbb{P}) = \mathbb{E}(I) = \sum_{a \in \mathfrak{A}} I(a)p(a) = -\sum_{a \in \mathfrak{A}} p(a) \log_2 p(a).$$

Exemple 11.1.3. *L'entropie d'une source composée d'un alphabet de* n *symboles tirés selon la loi uniforme est égale à* $\log_2 n$.

La signification de la longueur de code est la suivante : supposons que l'on code un message x de longueur $k \gg 1$ (ou une collection de messages de longueur totale k), alors le nombre $\ell(x)/k$ de bits par symbole nécessaire pour pouvoir coder ce message converge vers $L[c]$. En effet, la loi des grands nombres implique que si $x = (x_1, \ldots, x_k) \in \mathfrak{A}^k$ est un message de longueur k, alors

$$\frac{\ell(x)}{k} = \frac{1}{k} \sum_{i=1}^{k} \ell(x_i) \xrightarrow{\mathbb{P}} \mathbb{E}(\ell) = L[c].$$

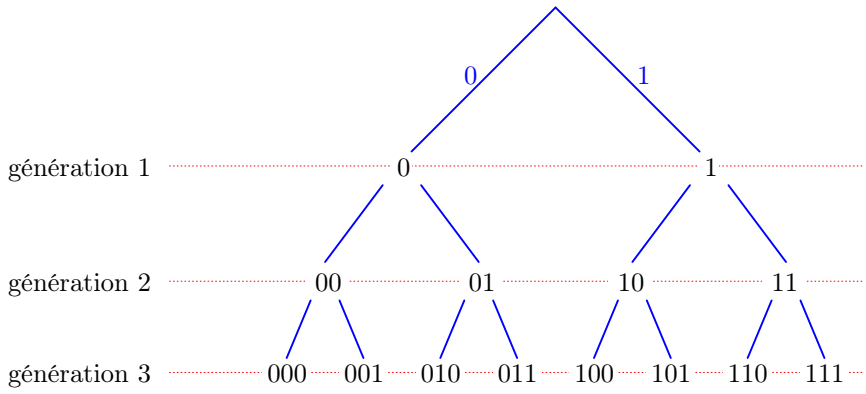


FIGURE 11.1: Le début de l'arbre binaire de la preuve du Théorème 11.1.1. Si un nœud du graphe est associé à un mot binaire $x \in \{0,1\}^n$, alors les mots associés à ses deux enfants sont $x \odot 0$ pour le descendant à gauche, et $x \odot 1$ pour le descendant à droite. Le mot associé à la racine est vide.

Théorème 11.1.1 (Inégalité de Kraft)**.** *Soit $\mathfrak{A}$ un alphabet. Pour tout code préfixe c, on a*

$$\sum_{a \in \mathfrak{A}} 2^{-\ell(a)} \leq 1. \tag{11.1}$$

Remarque 11.1.1. *En fait, on peut montrer que tout code uniquement décodable satisfait l'inégalité de Kraft (ce sera fait en exercices).*

Démonstration. La preuve est basée sur la représentation des éléments de $\{0,1\}^\star$ comme nœuds d'un arbre (voir la Fig. 11.1). Manifestement, à la génération $n \geq 1$ de l'arbre, les mots binaires associés aux nœuds correspondent (de gauche à droite) à la suite des représentations binaires des nombres $0, 1, \ldots, 2^n - 1$.

On ordonne les mots de code selon leur longueur : $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_n$. On note A_i l'ensemble des nœuds de génération ℓ_n qui sont des descendants du $i^{\text{ème}}$ mot de code (de longueur ℓ_i) ; voir la Fig. 11.2. Le code étant préfixe, on doit avoir $A_i \cap A_j = \varnothing$, si $i \neq j$. De plus, $|A_i| = 2^{\ell_n - \ell_i}$. Comme il y a en tout 2^{ℓ_n} nœuds à la génération ℓ_n, on en déduit que

$$2^{\ell_n} \geq |\bigcup_{i=1}^{n} A_i| = \sum_{i=1}^{n} |A_i| = \sum_{i=1}^{n} 2^{\ell_n - \ell_i},$$

et le résultat suit. □

Le théorème précédent admet une réciproque.

Théorème 11.1.2. *Soit $\{\ell(a)\}_{a \in \mathfrak{A}}$ une famille d'entiers positifs satisfaisant l'inégalité de Kraft (11.1). Alors, il existe un code préfixe c possédant les $\ell(a)$ comme longueurs des mots de code.*

Démonstration. On construit le code préfixe explicitement à l'aide de l'algorithme suivant :

– On ordonne les $\ell(a)$: $\ell_1 \leq \ell_2 \leq \ell_3 \leq \ldots$.

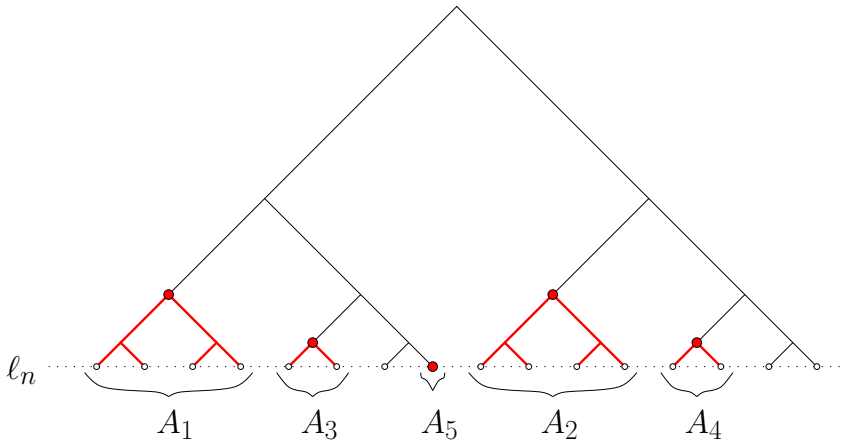


FIGURE 11.2: Dans cet exemple, on a $n = 5$ mots de code, de longueurs $\ell_1 = \ell_2 = 2$, $\ell_3 = \ell_4 = 3$, et $\ell_5 = 4$. Les nœuds correspondant aux mots de code sont indiqués en rouge, et leurs descendants sont marqués. Observez que l'ensemble des descendants de deux nœuds correspondant à des mots de code différents sont disjoints, car le code est préfixe.

- On choisit pour le premier mot de code c_1 le nœud de génération ℓ_1 le plus à gauche ; en l'occurence, $c_1 = 0 \cdots 0$ (composé de ℓ_1 « 0 »). On marque ce nœud, ainsi que tous ses descendants.
- On choisit pour c_{k+1} le mot correspondant au nœud non marqué de la génération ℓ_{k+1} se trouvant le plus à gauche. On marque le nœud, ainsi que tous ses descendants.

Cette construction ne peut échouer que si on doit attribuer le $k^{\text{ème}}$ mot de code alors que tous les nœuds de génération ℓ_k sont déjà marqués. Or, le nombre de nœuds de la génération ℓ_k qui ont été marqués lors des $k-1$ étapes précédentes de la construction est donné par

$$\sum_{j=1}^{k-1} 2^{\ell_k - \ell_j}.$$

Par conséquent, si les 2^{ℓ_k} nœuds de la génération ℓ_k étaient marqués, on aurait $\sum_{j=1}^{k-1} 2^{\ell_j} = 1$, et l'inégalité de Kraft serait violée. □

11.2 Taux optimal de compression

Dans cette section, nous allons démontrer le premier théorème de Shannon. Celui-ci établit la limite absolue du taux de compression pour des messages issus d'une source discrète sans mémoire $A = (\mathfrak{A}, \mathbb{P})$: le nombre de bits par symbole nécessaires pour coder de façon réversible un message est compris entre $H(\mathbb{P})$ et $H(\mathbb{P}) + 1$.

Théorème 11.2.1 (Premier théorème de Shannon)**.** *1. Soit $A = (\mathfrak{A}, \mathbb{P})$ une source discrète sans mémoire, et c un code associé satisfaisant l'inégalité de Kraft. Alors,*

$$L[c] \geq H(\mathbb{P}),$$

avec égalité si et seulement si $\ell(a) = I(a)$, pour tout $a \in \mathfrak{A}$.

2. Pour toute source discrète sans mémoire $A = (\mathfrak{A}, \mathbb{P})$, il existe un code préfixe c tel que

$$L[c] < H(\mathbb{P}) + 1.$$

Démonstration. On montre tout d'abord la première affirmation.

$$\begin{aligned} L[c] - H[\mathbb{P}] = \mathbb{E}\big(\ell(a) + \log_2(p(a))\big) &= -\mathbb{E}\big(\log_2(2^{-\ell(a)}/p(a))\big) \\ &\geq -\log_2\big(\mathbb{E}(2^{-\ell(a)}/p(a))\big) = -\log_2\big(\sum_{a\in\mathfrak{A}} 2^{-\ell(a)}\big) \geq 0, \end{aligned}$$

où la première inégalité suit d'une application de l'inégalité de Jensen, et la seconde de l'inégalité de Kraft. L'égalité n'a lieu que si les deux inégalités ci-dessus sont saturées. Or, l'inégalité de Jensen est saturée si et seulement si $2^{-\ell(a)}/p(a)$ est constante, c'est-à-dire si $p(a) = \lambda 2^{-\ell(a)}$ pour un $\lambda > 0$. L'inégalité de Kraft est elle saturée si et seulement si $\sum_{a\in\mathfrak{A}} 2^{-\ell(a)} = 1$. On en conclut que $\lambda = 1$, ce qui prouve la première affirmation.

La seconde affirmation suit d'une construction explicite. Il est clair de l'argument ci-dessus que les codes les plus efficaces sont ceux qui vont satisfaire au mieux la relation $\ell(a) = I(a)$. Comme les longueurs doivent être entières, on choisit

$$\ell(a) = \lceil I(a) \rceil,$$

où, pour $x \in \mathbb{R}$, $\lceil x \rceil = \min\{n \in \mathbb{N} : n \geq x\}$. Ces longueurs satisfont l'inégalité de Kraft, car

$$\sum_{a\in\mathfrak{A}} 2^{-\ell(a)} = \sum_{a\in\mathfrak{A}} 2^{-\lceil I(a)\rceil} \leq \sum_{a\in\mathfrak{A}} 2^{-I(a)} = \sum_{a\in\mathfrak{A}} 2^{\log_2(p(a))} = \sum_{a\in\mathfrak{A}} p(a) = 1.$$

Il suit donc du Théorème 11.1.2 qu'il existe un code préfixe dont les longueurs des mots de code sont données par $\ell(a) = \lceil I(a) \rceil$. Ce code a par conséquent la longueur de code suivante :

$$L[c] = \mathbb{E}(\lceil I \rceil) < \mathbb{E}(I + 1) = H(\mathbb{P}) + 1.$$

□

Remarque 11.2.1. *On peut en fait virtuellement se débarrasser du bit supplémentaire de la borne supérieure. Il suffit pour cela de considérer des messages sur l'alphabet $\mathfrak{A}^k$, c'est-à-dire de coder les mots par blocs de longueurs k. Notons $A^k = (\mathfrak{A}^k, \mathbb{P}^k)$ la source correspondante. Il suit alors du Théorème précédent que le code optimal c_k satisfait*

$$H(\mathbb{P}^k) \leq L[c_k] < H(\mathbb{P}^k) + 1,$$

et donc

$$H(\mathbb{P}) \leq \frac{1}{k} L[c_k] < H(\mathbb{P}) + \frac{1}{k},$$

puisque $H(\mathbb{P}^k) = -\sum_{x\in\mathfrak{A}^k} p(x) \log_2 p(x) = -k \sum_{a\in\mathfrak{A}} p(a) \log_2 p(a) = kH(\mathbb{P})$. On peut donc s'approcher arbitrairement près du taux optimal de compression, en regroupant convenablement les messages.

Exemple 11.2.1. *On conclut de ce théorème, et de la remarque qui suit, que le taux de compression optimal de l'image à envoyer par fax de l'Exemple 11.1.2 est asymptotiquement de* $H(\mathbb{P}) = 0,99\log_2(0,99) + 0,01\log_2(0,01) \simeq 8\%$.

11.3 Transmission à travers un canal bruité

Nous allons à présent nous intéresser à la transmission d'un message dans un canal bruité. Le but est de coder ce message de façon à minimiser le risque de ne plus pouvoir le décoder à l'arrivée. Il faudra pour cela introduire de la redondance dans l'information transmise, et on cherche à minimiser la taille du code ainsi produit. Le second théorème de Shannon donne la taille optimale du code.

Le second théorème de Shannon s'applique à des canaux très généraux, mais par souci de simplicité, nous nous restreindrons à un type particulier de canal bruité : le canal binaire symétrique, défini comme suit. On suppose que l'on transmet un message binaire de longueur n. Le bruit est modélisé par une chaîne binaire aléatoire $Y = (Y_1, \ldots, Y_n)$, dont les n composantes sont i.i.d. et suivent une loi bernoulli(q), $0 \leq q \leq 1/2$; notons $\mathbb{Q}$ la loi de Y. Notons également, pour $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in \{0,1\}^n$, $x \oplus y$ la chaîne $z \in \{0,1\}^n$ avec $z_i = (x_i + y_i) \mod 2$. Ayant introduit un message $x \in \{0,1\}^n$ à l'entrée du canal, on obtient à la sortie le message aléatoire $x \oplus Y \in \{0,1\}^n$. En d'autres termes, chacun des bits du message x a une probabilité q d'être modifié, indépendamment des autres.

Le problème peut à présent être formulé précisément. On part d'une source générant un message $x \in \{0,1\}^k$. On code ce message à l'aide d'une application $c : \{0,1\}^k \to \{0,1\}^n$, avec $n > k$. On transmet ensuite le message codé $c(x)$ à travers le canal bruité, obtenant à l'arrivée une version bruitée $c(x) \oplus Y$, où Y est de loi $\mathbb{Q}$. On applique alors un algorithme de décodage $d : \{0,1\}^n \to \{0,1\}^k$. La question est de déterminer la probabilité que $d(c(x) \oplus Y) = x$.

On appelle taux d'un code binaire le rapport k/n ; le taux mesure la redondance introduite par le code.

On appelle distance de Hamming entre deux chaînes binaires x et y de mêmes longueurs la distance $d_{\mathrm{H}}(x,y) = \#\{i : x_i \neq y_i\}$. Pour $y \in \{0,1\}^m$ et $r \geq 0$, notons

$$B_{\mathrm{H}}(y,r) = \{x \in \{0,1\}^m : d_{\mathrm{H}}(x,y) \leq r\},$$

la boule de rayon r centrée en y (pour la distance de Hamming), et

$$\mathrm{Vol}(r,m) = |B_{\mathrm{H}}(y,r)| = \sum_{i=0}^{\lfloor r \rfloor} \binom{m}{i},$$

son volume (manifestement indépendant du choix de y).

Soit $H(q) = -q\log_2(q) - (1-q)\log_2(1-q)$; observez que $H(q) < 1$ lorsque $q < \frac{1}{2}$. Pour de telles valeurs de q, on a alors, lorsque $m \to \infty$.

$$\mathrm{Vol}(qm, m) = 2^{(H(q)+o(1))m}. \tag{11.2}$$

En effet, cela suit facilement des bornes $\binom{m}{\lfloor qm \rfloor} \leq \mathrm{Vol}(qm, m) \leq qm\binom{m}{\lfloor qm \rfloor}$, et de la formule de Stirling.

Le second théorème de Shannon affirme qu'il est possible de transmettre un message à travers un canal binaire symétrique de paramètre $q < \frac{1}{2}$ à n'importe quel taux inférieur à $1 - H(q)$, avec une probabilité de succès arbitrairement proche de 1. Nous verrons plus tard que cette borne est optimale.

Il est possible de comprendre heuristiquement pourquoi une telle borne apparaît. Soit $m \in \{0,1\}^k$ le message original. On le code en un message $c(m) \in \{0,1\}^n$, avec $n > k$. Lors de la transmission de $c(m)$ à travers le canal bruité de paramètre $q < 1/2$, une proportion approximativement q (par la loi des grands nombres appliquée au bruit Y) des bits de $c(m)$ vont être modifiés. Par conséquent, si $z \in \{0,1\}^n$ est le message reçu à la sortie du canal, on aura $d_{\mathrm{H}}(c(m), z) \approx qn$. Supposons que le codage ait la propriété que les boules $B_{\mathrm{H}}(c(m), qn)$ soient disjointes. Alors le message original m peut être retrouvé : ce sera l'unique message tel que $z \in B_{\mathrm{H}}(c(m), qn)$. Mais le volume total de ces boules est de l'ordre de $2^k 2^{H(q)n}$, par (11.2). L'existence d'un tel codage ne sera donc a priori possible que si le volume de l'espace des codes satisfait $2^n \gg 2^k 2^{H(q)n}$, c'est-à-dire si $k/n < 1 - H(q)$. Les preuves ci-dessous suivent de près cet argument heuristique.

Dans le théorème suivant, et encore davantage dans sa preuve, nous aurons besoin de considérer des espérances et des probabilités portant simultanément sur plusieurs variables aléatoires (indépendantes). Afin de rendre les notations aussi explicites que possible, nous indicerons les symboles $\mathbb{E}$ et $\mathbb{P}$ par les variables aléatoires correspondantes. Par exemple, l'espérance par rapport à deux variables aléatoires X et Y (indépendantes, et dont les lois sont déterminées par le contexte), s'écrira $\mathbb{E}_{X,Y}$.

Théorème 11.3.1 (Deuxième théorème de Shannon)**.** *Soit $q < \frac{1}{2}$, et $\epsilon, \delta > 0$. Il existe $n_0 < \infty$ tel que, pour tout $n \geq n_0$ et $k \leq (1 - H(q) - \epsilon)n$, il existe deux fonctions $c : \{0,1\}^k \to \{0,1\}^n$ et $d : \{0,1\}^n \to \{0,1\}^k$ telles que*

$$\mathbb{P}_{M,Y}\big(d(c(M) \oplus Y) = M\big) \geq 1 - \delta,$$

où M, de loi uniforme sur $\{0,1\}^k$, et Y, de loi $\mathbb{Q}$, sont indépendantes.

Démonstration. La preuve est non constructive. Soient $n > k$ comme dans l'énoncé. La fonction de codage $C : \{0,1\}^k \to \{0,1\}^n$ est choisie au hasard : à chaque $m \in \{0,1\}^k$, on associe un code $C(m)$ tiré au hasard uniformément sur $\{0,1\}^n$, indépendamment pour chaque m (le code obtenu peut ne pas être injectif).

La fonction de décodage (aléatoire, car dépendant du choix de C), $D : \{0,1\}^n \to \{0,1\}^k$ est définie de la façon suivante. Étant donné une chaîne $z \in \{0,1\}^n$, on cherche le message $m \in \{0,1\}^k$ minimisant $d_{\mathrm{H}}(z, C(m))$. On pose alors $D(z) = m$. (En cas de dégénérescence, on choisit le premier tel message m, selon l'ordre lexicographique)

Nous allons montrer que ces fonctions ont les bonnes propriétés, avec grande probabilité (par rapport au choix aléatoire de C).

On commence par considérer le problème de décoder un message particulier, mais quelconque, $m_0 \in \{0,1\}^k$. Soit y la réalisation du bruit dans le canal. Le message bruité est alors donné par $z = C(m_0) \oplus y$.

On fixe $\delta > 0$ (petit). Soit $r = (q+\delta)n$. Pour que le décodage échoue, il faut qu'au moins une des conditions suivantes soit vérifiée :

1. Trop d'erreurs de transmission : $z \notin B_{\mathrm{H}}(C(m_0), r)$.
2. Le mot reçu est trop proche d'autres mots : $\exists m' \neq m_0$ tel que $C(m') \in B_{\mathrm{H}}(z, r)$.

En effet, si ces deux conditions sont violées, alors m_0 est le seul message m tel que $C(m) \in B_{\mathrm{H}}(z, r)$, et on a bien $D(C(m_0) \oplus y) = m_0$.

Nous allons à présent montrer que les deux conditions ci-dessus ne sont vérifiées qu'avec faible probabilité (par rapport aux choix aléatoires de C et y).

Pour que la première ait lieu, il faut que la fraction de 1 dans y soit supérieure à $q+\delta$. Or, comme $\mathbb{E}(Y_i) = q$, il suit de la loi des grands nombres (ou d'une application des inégalités de Bienaymé-Tchebychev ou, mieux, de Chernoff) que cela a lieu avec probabilité

$$\mathbb{Q}\big(\sum_{i=1}^{n} Y_i \geq (q+\delta)n\big) \leq \mathbb{Q}\big(\big|\sum_{i=1}^{n}(Y_i - q)\big| \geq \delta n\big) \to 0,$$

lorsque $n \to \infty$, pour tout $\delta > 0$.

Analysons à présent le second événement. Fixons $z \in \{0,1\}^n$ et un message $m' \neq m_0$, et considérons l'événement $C(m') \in B_{\mathrm{H}}(z, r)$. La probabilité de cet événement, par rapport au choix aléatoire de la fonction C est donnée par

$$\mathbb{P}_C(C(m') \in B_{\mathrm{H}}(z, r)) = \mathrm{Vol}(r, n)/2^n = 2^{(H(q+\delta)-1+o(1))n}.$$

Cette probabilité est évidemment indépendante du choix de m'. Par conséquent,

$$\begin{aligned}
\mathbb{P}_C(\exists m' \neq m_0,\, C(m') \in B_{\mathrm{H}}(z, r)) &\leq 2^k\, 2^{(H(q+\delta)-1+o(1))n} \\
&= 2^{(k/n-1+H(q+\delta)+o(1))n} \\
&\leq 2^{(1-H(q)-\epsilon-1+H(q+\delta)+o(1))n} \\
&\leq 2^{-\frac{1}{2}\epsilon\, n},
\end{aligned}$$

en prenant δ suffisamment petit et n suffisamment grand. Cette probabilité est donc également négligeable.

Soit ρ la probabilité totale des deux événements ci-dessus (on a donc $\rho \to 0$, lorsque $n \to \infty$). Nous avons démontré que pour n'importe quel message $m_0 \in \{0,1\}^k$ donné, la probabilité d'avoir une erreur au décodage satisfait

$$\mathbb{P}_{Y,C}(D(C(m_0) \oplus Y) \neq m_0) \leq \rho.$$

On a alors, en moyennant sur le message initial M (tiré au hasard uniformément sur $\{0,1\}^k$),

$$\begin{aligned}
\mathbb{E}_{M,Y,C}(\mathbf{1}_{\{D(C(M)\oplus Y)\neq M\}}) &= \mathbb{E}_M(\mathbb{E}_{Y,C}(\mathbf{1}_{\{D(C(M)\oplus Y)\neq M\}})) \\
&= \mathbb{E}_M(\mathbb{P}_{Y,C}(D(C(M) \oplus Y) \neq M)) \leq \mathbb{E}_M(\rho) = \rho.
\end{aligned}$$

Mais cette inégalité implique qu'il doit exister une réalisation c de C (et donc d de D) telle que

$$\mathbb{E}_{M,Y}(\mathbf{1}_{\{d(c(M)\oplus Y)\neq M\}}) \leq \rho,$$

et donc

$$\mathbb{P}_{M,Y}(d(c(M)\oplus Y)\neq M) = \mathbb{E}_{M,Y}(\mathbf{1}_{\{d(c(M)\oplus Y)\neq M\}}) \leq \rho,$$

ce qui conclut la preuve. □

Le théorème suivant montre que la borne du théorème précédent est optimale.

Théorème 11.3.2. *Soit $q < \frac{1}{2}$, et $\epsilon, \delta > 0$. Il existe $n_0 < \infty$ tel que, pour tout $n \geq n_0$ et $k \geq (1 - H(q) + \epsilon)n$, et pour toutes fonctions $c : \{0,1\}^k \to \{0,1\}^n$ et $d : \{0,1\}^n \to \{0,1\}^k$, on a*

$$\mathbb{P}_{M,Y}\big(d(c(M)\oplus Y)\neq M\big) \geq 1 - \delta,$$

où M, de loi uniforme sur $\{0,1\}^k$, et Y, de loi $\mathbb{Q}$, sont indépendantes.

Ce résultat montre que si l'on essaie d'obtenir un taux supérieur à $1 - H(q) + \epsilon$, le message ne pourra pas être récupéré à l'arrivée, avec probabilité proche de 1, quelles que soient les procédures de codage/décodage choisies.

Démonstration. La preuve repose sur les deux observations suivantes. Soit $\epsilon' > 0$ et notons $A_{\mathrm{H}}(c(m), \epsilon') = B_{\mathrm{H}}(c(m), (q+\epsilon')n) \setminus B_{\mathrm{H}}(c(m), (q-\epsilon')n)$.

– Par la loi des grands nombres (ou une application des inégalités de Bienaymé-Tchebychev ou Chernoff), on a que, pour tout $\epsilon' > 0$,

$$\mathbb{Q}\big(\big|\sum_{i=1}^{n} Y_i - nq\big| \geq \epsilon' n\big) \to 0, \qquad n \to \infty.$$

Par conséquent, uniformément en $m \in \{0,1\}^k$ et en la fonction de codage c,

$$\mathbb{Q}\big(c(m)\oplus Y \in A_{\mathrm{H}}(c(m), \epsilon')\big) \to 1, \qquad n \to \infty, \tag{11.3}$$

pour tout $\epsilon' > 0$. Ceci signifie que le nombre d'erreurs de transmissions ne peut être (avec grande probabilité) ni trop grand, ni trop petit (il doit être d'ordre qn).

– Pour tout $\epsilon' > 0$ suffisamment petit, $m \in \{0,1\}^k$, c, et $z \in A_{\mathrm{H}}(c(m), \epsilon')$,

$$\mathbb{Q}(c(m)\oplus Y = z) \leq 2^{-(H(q-\epsilon')-o(1))n}. \tag{11.4}$$

En effet, $R = d_{\mathrm{H}}(z, c(m)) \in [(q-\epsilon')n, (q+\epsilon')n]$. Soit $N_R = \binom{n}{R}$ le nombre de vecteurs binaires avec R « 1 » et $n - R$ « 0 ». Toutes les réalisations de Y avec le même nombre d'erreurs étant équiprobables, la probabilité ci-dessus ne peut être qu'au plus de $1/N_R$. Mais la formule de Stirling implique que, pour ϵ' suffisamment petit, $N_R \geq 2^{(H(q-\epsilon')-o(1))n}$.

Voyons à présent comment utiliser ces observations pour montrer que tout décodage a probabilité proche de 1 d'échouer.

Soit $c : \{0,1\}^k \to \{0,1\}^n$ et $d : \{0,1\}^n \to \{0,1\}^k$ des applications de codage et décodage arbitraires, et soit $S_m = \{z \in \{0,1\}^n \,:\, d(z) = m\}$. Notons ρ la probabilité que le décodage soit un succès,

$$\begin{aligned}\rho &= \mathbb{P}_{M,Y}(d(c(M) \oplus Y) = M) = \mathbb{P}_{M,Y}(c(M) \oplus Y \in S_M) \\ &= \sum_{m \in \{0,1\}^k} \mathbb{P}(M = m)\, \mathbb{Q}(c(m) \oplus Y \in S_m).\end{aligned}$$

M étant distribué uniformément sur $\{0,1\}^k$, on a $\mathbb{P}(M = m) = 2^{-k}$. Il reste à estimer le second facteur. On a, en écrivant $S_m = (S_m \cap A_{\mathrm{H}}(c(m), \epsilon')) \cup (S_m \setminus A_{\mathrm{H}}(c(m), \epsilon'))$,

$$\begin{aligned}\mathbb{Q}(c(m) \oplus Y \in S_m) &\le \mathbb{Q}(c(m) \oplus Y \in S_m \cap A_{\mathrm{H}}(c(m), \epsilon')) + \mathbb{Q}(c(m) \oplus Y \notin A_{\mathrm{H}}(c(m), \epsilon')) \\ &\le |S_m| \sup_{z \in A_{\mathrm{H}}(c(m), \epsilon')} \mathbb{Q}(c(m) \oplus Y = z) + \mathbb{Q}(c(m) \oplus Y \notin A_{\mathrm{H}}(c(m), \epsilon')).\end{aligned}$$

Observons à présent que $\sum_{m \in \{0,1\}^k} |S_m| = 2^n$. Comme $n - k \le H(q)n - \epsilon n$, il suit donc de (11.3) et (11.4) que

$$\begin{aligned}\rho &\le 2^{-(H(q-\epsilon')-o(1))n}\, 2^{-k} \sum_{m \in \{0,1\}^k} |S_m| + \sup_{m \in \{0,1\}^k} \mathbb{Q}(c(m) \oplus Y \notin A_{\mathrm{H}}(c(m), \epsilon')) \\ &\to 2^{-(H(q-\epsilon')-H(q)+\epsilon-o(1))n} + o(1),\end{aligned}$$

ce qui tend vers 0 lorsque $n \to \infty$, pour tout $\epsilon > 0$, pourvu que ϵ' soit choisi assez petit. □

Chapitre 12

La méthode probabiliste

Dans ce chapitre, nous allons voir quelques exemples d'application de méthodes et intuition d'origine probabiliste dans divers domaines de mathématiques où elles pourraient être inattendues. Cette méthode probabiliste, initiée par P. Erdős est devenue une technique centrale en combinatoire. Nous nous contenterons ici d'en donner quelques illustrations simples.

12.1 Combinatoire : le théorème d'Erdős-Ko-Rado

Une famille d'ensembles $\mathcal{F}$ est dite intersectante si $A \cap B \neq \varnothing$, pour tout $A, B \in \mathcal{F}$. Soit $n \geq 2k$ et $\mathcal{F}$ une famille intersectante de sous-ensembles de k éléments d'un ensemble de n éléments, disons $\{0, \ldots, n-1\}$.

Théorème 12.1.1 (Erdős-Ko[1]-Rado[2])**.** $|\mathcal{F}| \leq \binom{n-1}{k-1}$.

Remarque 12.1.1. *La borne est facilement saturée : il suffit de considérer la famille des sous-ensembles à k éléments contenant un élément donné.*

Démonstration. La preuve repose sur le résultat suivant.

Lemme 12.1.1. *Pour $0 \leq s \leq n-1$, on pose $A_s = \{s, s+1, \ldots, s+k-1\}$, l'addition étant modulo n. Alors $\mathcal{F}$ ne peut contenir plus de k ensembles A_s.*

Preuve du lemme. Supposons que $A_\ell \in \mathcal{F}$. À part A_ℓ lui-même, exactement $2k-2$ des ensembles A_s intersectent A_ℓ. Ceux-ci peuvent être répartis en $k-1$ paires d'ensembles disjoints. Puisque $\mathcal{F}$ ne peut contenir qu'au plus un membre de chacune de ces paires, le lemme est démontré. □

Revenons à la preuve du théorème. On tire au hasard une permutation σ de $\{0, \ldots, n-1\}$ et un élément $i \in \{0, \ldots, n-1\}$, tous deux de façon uniforme, et indépendamment l'un

1. Ke Zhao ou Chao Ko (1910, Taizhou – 2002, ? ? ?), mathématicien chinois.
2. Richard Rado (1906, Berlin – 1989, Henley-on-Thames), mathématicien allemand.

de l'autre. Soit $A = \{\sigma(i), \sigma(i+1), \ldots, \sigma(i+k-1)\}$ (la somme étant toujours prise modulo n). Il est clair (et facilement démontré) que la loi de A est uniforme sur l'ensemble des sous-ensembles à k éléments de $\{0, \ldots, n-1\}$. En particulier, $\mathbb{P}(A \in \mathcal{F}) = |\mathcal{F}|/\binom{n}{k}$. D'un autre côté,

$$\mathbb{P}(A \in \mathcal{F}) = \sum_{\sigma} \mathbb{P}(A \in \mathcal{F} \,|\, \sigma)\mathbb{P}(\sigma).$$

Conditionnellement à σ, A suit la loi uniforme sur les n sous-ensembles de k éléments consécutifs de l'ensemble ordonné $\{\sigma(0), \sigma(1), \ldots, \sigma(n-1)\}$ (consécutifs au sens du Lemme). Par conséquent, le lemme (appliqué à l'ensemble $\{\sigma(0), \ldots, \sigma(n-1)\}$) implique que $\mathbb{P}(A \in \mathcal{F} \,|\, \sigma) \leq k/n$, pour chaque permutation σ, et donc $\mathbb{P}(A \in \mathcal{F}) \leq k/n$. On doit donc avoir

$$\frac{|\mathcal{F}|}{\binom{n}{k}} \leq \frac{k}{n},$$

et le théorème est démontré. □

12.2 Théorie des nombres : facteurs premiers

Soit $n \in \mathbb{N}^*$, et $\nu(n)$ le nombre de nombres premiers p divisant n (sans multiplicité). Le résultat suivant a été démontré tout d'abord par Hardy [3] et Râmânujan [4] en 1920, par un argument plutôt complexe. La preuve ci-dessous est due à Paul Turán [5] (1934) et a joué un rôle clé dans le développement des méthodes probabilistes en théorie des nombres.

Théorème 12.2.1. *Soit $\epsilon > 0$. Lorsque $N \to \infty$, on a*

$$\frac{1}{N} \# \left\{ n \in \{1, \ldots, N\} \,:\, |\nu(n) - \log\log N| > (\log\log N)^{1/2+\epsilon} \right\} = o(1).$$

Démonstration. Soit $\pi(n)$ le nombre de nombres premiers inférieurs ou égaux à n. Le résultat suivant, dû à Mertens [6], est classique ; sa preuve, simple, peut être trouvée dans la plupart des livres de théorie analytique des nombres.

Théorème 12.2.2.

$$\sum_{p \leq N, \text{ premier}} \frac{1}{p} = \log\log N + O(1).$$

Passons à présent à la preuve du Théorème 12.2.1. On tire n au hasard uniformément dans $\{1, \ldots, N\}$. Pour $p \in \{1, \ldots, N\}$, on définit les variables aléatoires

$$X_p(n) = \begin{cases} 1 & \text{si } p|n, \\ 0 & \text{sinon,} \end{cases}$$

3. Godfrey Harold Hardy (1877, Cranleigh – 1947, Cambridge), mathématicien britannique.
4. Srinivâsa Aiyangâr Râmânujan (1887, Erode – 1920, Kumbakonam), mathématicien indien.
5. Paul, ou Pál, Turán (1910, Budapest – 1976, Budapest), mathématicien hongrois.
6. Franz Mertens (1840, Środa Wielkopolska – 1927, Vienne), mathématicien allemand.

et $X = \sum_{p \leq N, \text{premier}} X_p$. Clairement $X(n) = \nu(n)$. On a

$$\begin{aligned}\mathbb{E}(X_p) &= \sum_{1 \leq n \leq N} \frac{1}{N} X_p(n) = \frac{1}{N} \#\{kp : 1 \leq kp \leq N\} = \frac{\lfloor N/p \rfloor}{N} \\ &= \frac{1}{p} + O(\frac{1}{N}),\end{aligned}$$

et donc

$$\mathbb{E}(X) = \sum_{p \leq N, \text{premier}} \big(\frac{1}{p} + O(\frac{1}{N})\big) = \log\log N + O(1), \tag{12.1}$$

où la dernière identité suit du Théorème 12.2.2. Bornons à présent la variance de X. D'une part, $\mathbb{E}(X)^2 = (\log\log N)^2 + O(\log\log N)$. D'autre part, puisque pour deux nombres premiers distincts p, q, on a que $p|n$ et $q|n$ si et seulement si $pq|n$, et donc $X_p X_q = X_{pq}$, il suit que

$$\begin{aligned}\mathbb{E}(X^2) = \mathbb{E}(X) + \mathbb{E}\Big(\sum_{\substack{p \neq q \leq N \\ \text{premiers}}} X_p X_q\Big) &= \mathbb{E}(X) + \sum_{\substack{p \neq q \leq N \\ \text{premiers}}} \mathbb{E}(X_{pq}) \\ &\leq \mathbb{E}(X) + \sum_{\substack{p \neq q \leq N \\ \text{premiers}}} \frac{1}{pq} \leq (\log\log N)^2 + O(\log\log N),\end{aligned}$$

la dernière identité suivant de (12.1) et du Théorème 12.2.2. Par conséquent, $\mathrm{Var}(X) = O(\log\log N)$. Il suffit à présent d'appliquer l'inégalité de Tchebytchev (Théorème 5.2.2) :

$$\begin{aligned}\mathbb{P}(|\nu(n) - \log\log N| > (\log\log N)^{1/2+\epsilon}) &= \mathbb{P}(|X - \mathbb{E}(X)| > (\log\log N)^{1/2+\epsilon}(1 + o(1))) \\ &\leq \frac{\mathrm{Var}(X)}{(\log\log N)^{1+2\epsilon}}(1 + o(1)) = O(\log\log N)^{-2\epsilon}.\end{aligned}$$

□

La méthode utilisée dans la preuve ci-dessus, consistant à montrer qu'une variable aléatoire est proche de son espérance lorsque $\mathrm{Var}(X) \ll \mathbb{E}(X)^2$ est appelée méthode du second moment.

Le résultat précédent prend la forme d'une loi des grands nombres. On peut en fait aller beaucoup plus loin et montrer le résultat classique suivant, qui correspond au Théorème central limite. Nous ne le démontrerons pas ici, car la preuve est plus difficile, mais nous contenterons de remarquer que ce qui permet d'appliquer des approches probabilistes (et c'était déjà le cas dans la preuve précédente) est le fait que les variables aléatoires X_p sont presque indépendantes, lorsque N est grand.

Théorème 12.2.3 (Erdős-Kac[7]). *Soit $\lambda \in \mathbb{R}$ fixé. Alors,*

$$\lim_{N \to \infty} \frac{1}{N} \#\Big\{n \in \{1, \ldots, N\} : \nu(n) \geq \log\log N + \lambda\sqrt{\log\log N}\Big\} = \int_\lambda^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \mathrm{d}t.$$

7. Mark, ou Marek, Kac (1914, Krzemieniec – 1984, Californie), mathématicien américain d'origine polonaise.

12.3 Théorie des graphes : nombre chromatique

Soit $G = (V, E)$ un graphe et $k \in \mathbb{N}^\star$. Un k-coloriage de G est une application $c : V \to \{1, \dots, k\}$ telle que $c(i) \neq c(j)$, pour tout $\{i, j\} \in E$. En d'autres mots, on colorie chaque sommet du graphe de sorte à ce que deux sommets voisins (c'est-à-dire liés par une arête) soient de couleurs différentes. On appelle nombre chromatique du graphe G, noté $\chi(G)$, la plus petite valeur de k pour laquelle un k-coloriage de G existe. On appelle cycle de longueur k dans G une famille ordonnée de k sommets distincts $v_1, \dots, v_k \in V$ telle que $\{v_i, v_{i+1}\} \in E$, $i = 1, \dots, k-1$, $\{v_k, v_1\} \in E$, et chaque arête est utilisée au plus une fois ; on ne distinguera pas deux cycles correspondant à la même famille de sommets et d'arêtes, mais avec un point de départ différent ou un sens de parcours différent (il y a donc $2k$ « versions » d'un même cycle). On appelle maille du graphe G, notée $g(G)$, la longueur du plus petit de ses cycles ; la maille est infinie si G est sans cycle. Un ensemble de sommets de G est stable si aucune paire de sommets de l'ensemble n'est reliée par une arête. On note $\alpha(G)$ la taille du plus grand ensemble stable de G.

On pourrait penser qu'une condition suffisante pour pouvoir colorier un graphe avec un petit nombre de couleurs est que sa maille soit assez grande. Il se trouve que c'est faux : le théorème suivant montre qu'il existe des graphes dont la maille et le nombre chromatique sont arbitrairement grands.

Théorème 12.3.1. *Pour tout $k, \ell > 0$, il existe un graphe G tel que $\chi(G) > k$ et $g(G) > \ell$.*

Démonstration. Soit $\epsilon < 1/(2\ell)$, $n \in \mathbb{N}^\star$ et $p = n^{\epsilon - 1}$. On considère un graphe aléatoire dont l'ensemble des sommets est $V_n = \{1, \dots, n\}$, et avec une arête entre i et j avec probabilité p, indépendamment pour chaque couple $i \neq j \in V_n$.

On commence par estimer le nombre X de cycles de longueur au plus ℓ. Le nombre de cycles potentiels de longueur i est donné par $\frac{1}{2i}(n)_i \leq n^i$ (on tire sans remise i éléments parmi n et on identifie les $2i$ cycles ne différant que par leur point de départ ou leur orientation). Chaque arête étant présente avec probabilité p, on a

$$\mathbb{E}(X) = \sum_{i=3}^{\ell} \frac{1}{2i}(n)_i \leq \sum_{i=3}^{\ell} n^i p^i = \sum_{i=3}^{\ell} n^{i/(2\ell)} \leq \ell n^{1/2},$$

puisque $i/(2\ell) \leq 1/2$ pour tout $3 \leq i \leq \ell$. On déduit donc du Théorème 5.2.2 que

$$\mathbb{P}\big(X \geq \frac{n}{2}\big) \leq \frac{\mathbb{E}(X)}{n/2} \leq 2\ell n^{-1/2}.$$

Nous allons à présent contrôler $\alpha(G)$. Soit $a = \lceil (3/p) \log n \rceil$. Observant que $\alpha(G) \geq a$ implique l'existence d'un ensemble de a sommets qui soit stable, on obtient

$$\mathbb{P}(\alpha(G) \geq a) \leq \binom{n}{a}(1-p)^{a(a-1)/2} \leq \big(ne^{-p(a-1)/2}\big)^a \leq \big(ne^{-3\log n/2}\big)^a = n^{-a/2},$$

puisque $\lceil x \rceil - 1 < x$, pour tout x. On choisit à présent n suffisamment grand pour que $\max(2\ell n^{-1/2}, n^{-a/2}) < 1/2$. On en déduit alors que

$$\mathbb{P}(X < \frac{n}{2}, \alpha(G) < a) > 0,$$

ce qui montre qu'il existe un graphe G_0 possédant moins de $n/2$ cycles de longueur au plus ℓ, et avec $\alpha(G_0) < 3n^{1-\epsilon}\log n$ (si ce n'était pas le cas, la probabilité ci-dessus serait nulle).

On enlève à présent un sommet de chacun des cycles de G_0 de longueur au plus ℓ, obtenant ainsi un graphe $G_0^\star$ possédant au moins $n/2$ sommets, une maille supérieure à ℓ, et $\alpha(G_0^\star) \leq \alpha(G_0)$. Puisque dans chaque coloriage de G, les sommets d'une même couleur forment un ensemble stable et sont donc de taille au plus $a-1$, on en déduit que

$$\chi(G_0^\star) \geq \frac{n/2}{a-1} \geq \frac{pn}{6\log n} = \frac{n^\epsilon}{6\log n}.$$

La conclusion suit en choisissant n suffisamment grand pour que le membre de droite soit strictement supérieur à k. □

12.4 Géométrie : triangles vides

Soit X un ensemble fini de points dans le plan en position générique (c'est-à-dire sans triplets de points alignés). Notons $f(X)$ le nombre de triangles vides déterminés par les triplets de points dans X, c'est-à-dire le nombre de triangle dont les sommets sont des points de X et dont l'intérieur ne contient aucun point de X. Plusieurs personnes se sont intéressées à estimer la valeur minimale que peut prendre $f(X)$ lorsque X contient n points. On introduit $f(n) = \min f(X)$, où le minimum est pris sur tous les ensembles génériques de n points dans le plan. Bárány et Füredi ont montré en 1987 que, lorsque n croît,

$$(1+o(1))n^2 \leq f(n) \leq (1+o(1))2n^2.$$

Nous allons démontrer la borne supérieure.

Théorème 12.4.1. *Soient*

$$I_k = \left\{(x,y) \in \mathbb{R}^2 \,:\, x = k, 0 \leq y \leq 1\right\}, \qquad 1 \leq k \leq n.$$

Pour chaque k, on choisit indépendamment un point p_k au hasard, uniformément sur I_k. Soit X l'ensemble constitué de ces n points. Alors $\mathbb{E}(f(X)) \leq 2n^2 + O(n\log n)$.

Observez que ceci démontre bien la borne supérieure. En effet, presque toute réalisation de l'ensemble de points aléatoire X est générique, et le fait que $\mathbb{E}(f(X)) \leq 2n^2 + O(n\log n)$ implique l'existence d'un ensemble de probabilité positive de telles configurations de points pour lesquelles $f(X) < 2n^2 + O(n\log n)$.

Démonstration. On commence par estimer la probabilité que le triangle déterminé par les trois points p_i, p_{i+a}, p_{i+k} soit vide, pour des i, a fixés et $k = a+b \geq 3$. Notons $A = (i,x)$, $B = (i+a,y)$ et $C = (i+k,z)$ les points p_i, p_{i+a}, p_{i+k}. Soit m la distance séparant B du point d'intersection des segments AC et I_{i+a}. La probabilité que le triangle ABC est vide

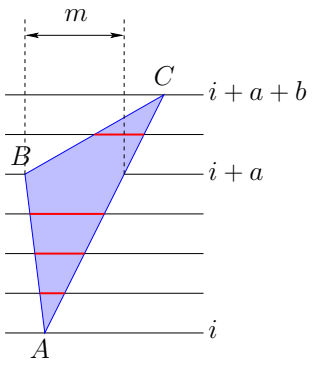


FIGURE 12.1: Pour que le triangle ABC soit vide, il faut qu'aucun point ne soit choisi sur les segments en rouge.

est donnée par (*cf.* 12.1)

$$\begin{aligned}
&\big(1-\frac{m}{a}\big)\big(1-2\frac{m}{a}\big)\cdots\big(1-(a-1)\frac{m}{a}\big)\big(1-(b-1)\frac{m}{b}\big)\cdots\big(1-\frac{m}{b}\big)\\
&\quad\le \exp\big(-\frac{m}{a}-2\frac{m}{a}-\cdots-(a-1)\frac{m}{a}-(b-1)\frac{m}{b}-\cdots-\frac{m}{b}\big)\\
&\quad= \exp\big(-\frac{a(a-1)}{2}\frac{m}{a}-\frac{b(b-1)}{2}\frac{m}{b}\big)\\
&\quad= \exp\big(-(k-2)\frac{m}{2}\big).
\end{aligned}$$

Il suit que la probabilité que le triangle ABC soit vide est bornée supérieurement par (on fixe A et C et on intègre sur m)

$$2\int_0^\infty \exp\big(-(k-2)\frac{m}{2}\big)\mathrm{d}m = \frac{4}{k-2},$$

uniformément en i, a, b tels que $a+b=k$. On peut à présent aisément borner l'espérance de $f(X)$. On observe tout d'abord que

$$f(X) = \sum_{k=2}^{n-1}\sum_{i=1}^{n-k}\sum_{a=1}^{k-1}\mathbf{1}_{\{ABC \text{ vide}\}} = (n-2) + \sum_{k=3}^{n-1}\sum_{i=1}^{n-k}\sum_{a=1}^{k-1}\mathbf{1}_{\{ABC \text{ vide}\}},$$

la seconde identité suivant du fait qu'un triangle dont les sommets se trouvent sur 3 lignes

consécutives ($k = 2$) est nécessairement vide. On obtient donc la borne

$$\begin{aligned}
\mathbb{E}(f(X)) &\leq n - 2 + \sum_{k=3}^{n-1}(n-k)(k-1)\mathbb{P}(ABC \text{ vide}) \\
&= n - 2 + \sum_{k=3}^{n-1}(n-k)\frac{4(k-1)}{k-2} \\
&= n - 2 + 4\sum_{k=3}^{n-1}(n-k)\frac{1}{k-2} + 4\sum_{k=3}^{n-1}(n-k) \\
&= 2n^2 + O(n \log n).
\end{aligned}$$

□

Index