l'essentiel en théorie des probabilités

Jean Jacod
Philip Protter

CASSINI

L'ESSENTIEL EN THÉORIE DES PROBABILITÉS

Enseignement des mathématiques

1. J.-Y. Ouvrard, *Probabilités I*
2. J. Hubbard, B. West, *Équations différentielles et systèmes dynamiques*
3. M. Cottrell, V. Genon-Catalot, Ch. Duhamel, Th. Meyre, *Exercices de probabilités*
4. F. Rouvière, *Petit guide de calcul différentiel à l'usage de la licence et de l'agrégation*
5. J.-Y. Ouvrard, *Probabilités II*
6. G. Zémor, *Cours de cryptographie*
7. A. Szpirglas, *Exercices d'algèbre*
8. B. Perrin-Riou, *Algèbre, arithmétique et Maple*
9. V. I. Arnold, *Leçons sur les équations aux dérivées partielles* (en préparation)

10-14. S. Francinou, H. Gianella, S. Nicolas, *Exercices de mathématiques — Oraux X-ENS*

15. H. Krivine, *Exercices de mathématiques pour physiciens*
16. J. Jacod, Ph. Protter, *L'essentiel en théorie des probabilités*
17. M. Willem, *Analyse fonctionnelle élémentaire*
18. É. Amar, É. Matheron, *Analyse complexe*
19. B. Randé, *Problèmes corrigés. Concours 2002 et 2003 (MP)* (en préparation)

JEAN JACOD   PHILIP PROTTER

L'essentiel en théorie des probabilités

CASSINI

Jean Jacod est professeur à l'université Paris VI-Pierre et Marie Curie. Il est l'auteur de *Calcul stochastique et problèmes de martingales* (Springer, 1979, Lecture notes in mathematics 714), de *Limit theorems for stochastic processes*, en coll. avec Albert N. Shiryaev (Springer, 1987, Grundlehren der mathematischen Wissenschaften 288) et de *Malliavin calculus for processes with jumps*, en coll. avec Klaus Bichteler et Jean-Bernard Gravereaux (Gordon and Breach, 1987).

Philip Protter est professeur à Cornell University (Ithaca, NY, États-Unis). Il est l'auteur de *Stochastic integration and differential equations : a new approach* (Springer, 1990, Applications of mathematics 21) et de *Calculus with analytic geometry*, en coll. avec M.H. Protter (Jones and Bartlett, Boston, 1998).

ISBN 2-84225-050-8

Édition anglaise : *Probability Essentials*, 2nd ed., Springer, 2003, ISBN 3-540-43871-8

à

Diane et Sylvie

et à

Rachel, Margot,

Olivier, Martin, Serge, Thomas et Vincent

Préface

Nous présentons ici le contenu d'un cours de probabilité destiné aux étudiants de licence ou de master de mathématiques ou mathématiques appliquées : un tel cours peut raisonnablement être constitué des chapitres 1 à 23. Les derniers chapitres 24 à 28, en revanche, sont sans doute trop difficiles pour un tel cours, mais peuvent servir d'introduction à la théorie des processus stochastiques.

L'origine de ce livre se trouve dans les notes d'un cours enseigné à l'université Paris 6 par le premier auteur, puis reprises et largement modifiées par le second dans un cours enseigné d'abord à Pérouse en 1997, puis à l'université de Purdue. Nous tenons à remercier les étudiants de tous ces cours pour leurs commentaires. Nous remercions également notre éditeur Catriona Byrne pour l'édition anglaise, ainsi que Nick Bingham pour ses remarques et suggestions constructives, ainsi qu'un referee anonyme qui à contribué à rendre le texte plus lisible, et enfin Judy Mitchell pour la réalisation matérielle.

La version française est une traduction assez fidèle, maïs améliorée (du moins l'espérons-nous) sur certains points. La plupart de ces améliorations ont été suggérées par André Bellaïche que nous remercions chaleureusement. La bibliographie a été élargie de façon à citer les ouvrages de référence en français, mais les références en langue anglaise ont été conservées.

Jean Jacod, Paris
Philip Protter, Ithaca

Table des matières

1. Introduction - Phénomènes aléatoires 1
2. Axiomes des probabilités 7
3. Probabilités conditionnelles et indépendance 15
4. Probabilités sur un espace fini ou dénombrable 23
5. Variables aléatoires sur un espace fini ou dénombrable 29
6. Construction d'une mesure de probabilité 37
7. Probabilités sur $\mathbb{R}$ et fonctions de répartition 41
8. Variables aléatoires 51
9. Intégration par rapport à une mesure de probabilité 55
10. Variables aléatoires indépendantes 71
11. Lois de probabilité sur $\mathbb{R}$ 83
12. Probabilités sur $\mathbb{R}^n$ 93
13. Fonctions caractéristiques 109
14. Propriétés des fonctions caractéristiques 117
15. Sommes de variables aléatoires indépendantes 123
16. Variables aléatoires gaussiennes 131
17. Convergence des variables aléatoires 147
18. Convergence en loi 157
19. Convergence en loi et fonctions caractéristiques 173
20. La loi des grands nombres 179
21. Le théorème-limite central 187
22. L^2 et les espaces de Hilbert 195
23. Espérance conditionnelle 203
24. Martingales 217
25. Surmartingales et sous-martingales 225
26. Les inégalités de martingales 229
27. Les théorèmes de convergence de martingales 235
28. Le théorème de Radon-Nikodym 249
Bibliographie 255
Index des notations 257
Index 259

Chapitre 1

Introduction - Phénomènes aléatoires

Chacun est maintenant familier avec le concept de probabilité : on nous indique (en Amérique du Nord au moins) chaque jour la probabilité pour qu'il pleuve le lendemain ; on évalue la probabilité de gagner au loto ou de survivre à un accident d'avion. Les compagnies d'assurances calculent les probabilités qu'un homme ou une femme ayant atteint 20 ans soit encore vivant(e) à 80 ans. Les supermarchés calculent le nombre de caisses à ouvrir en fonction de la probabilité du nombre de clients. Les banques estiment la probabilité qu'un prêt ne puisse pas être honoré. En médecine on évalue les probabilités de succès de divers traitements. Un succès récent de la théorie des probabilités concerne la gestion de portefeuille par les compagnies boursières, en utilisant des modèles d'évolution aléatoire des cours de la bourse très sophistiqués. On pourrait poursuivre cette liste presque indéfiniment : les probabilités sont omniprésentes dans la société moderne, et dans l'ensemble des sciences.

La théorie des probabilités est un sujet relativement ancien. Les premières publications sur les jeux de hasard remontent à J. Cardan (1501-1576) avec son livre *De Ludo Aleae* [5], ou à Kepler (1571-1630) et Galilée (1564-1642). Cependant les historiens des sciences s'accordent à dater le véritable début de la théorie aux travaux de Pascal (1623-1662) et Fermat (1601-1665). Ces derniers ont échangé une correspondance fournie, résolvant certains « paradoxes » posés par le Chevalier de Méré pour des jeux de cartes ou de dés. Plus tard le mathématicien hollandais Christian Huygens (1629-1695) écrivit un livre [16] ayant eu une grande influence et développant les idées de Pascal et de Fermat. Finalement en 1685 Jacques Bernoulli (1654-1705) proposa de nombreux problèmes probabilistes (dans le « Journal des Sçavans" ; voir aussi [3]) nécessitant le développement d'une théorie élaborée. Après les travaux de J. Bernoulli et de son contemporain A. De Moivre (1667-1754) [9], de nombreux mathématiciens renommés de cette époque travaillèrent sur des questions de probabilité, notamment Daniel Bernoulli (1700-1782), Euler (1707-1803), Gauss (1777-1855), et Laplace (1749-1827) : on peut se référer à [24] pour une histoire documentée des probabilités à cette époque. Au cours du vingtième siècle, le mérite de la reconnaissance des rapports entre les probabilités et la « théorie de la mesure » de Borel et Lebesgue revient à Kolmogorov (1903-1987). Après le travail fondamental de Kolmogorov, Paul Lévy a introduit nombre d'idées modernes sur les « processus aléatoires » et sur les fonctions caractéristiques et les théorèmes-limite.

Nous concevons la théorie des probabilités comme un modèle mathématique pour le « hasard », ou les « événements aléatoires ». L'idée consiste à partir d'un petit nombre de principes de base décrivant comment un phénomène aléatoire se comporte ; ces principes doivent être assez simples pour qu'on puisse légitimement considérer qu'ils décrivent de manière adéquate la réalité, ou la « nature ». Ils doivent aussi permettre l'élaboration d'une théorie mathématique suffisamment complexe pour résoudre des problèmes plus compliqués. C'est là le but de ce livre.

Décrivons maintenant l'approche adoptée. En premier lieu nous décrivons les principaux éléments des probabilités discrètes de façon à fonder les idées essentielles de la théorie. Ces idées sont ensuite étendues au cas des probabilités sur l'ensemble des réels : tout ceci occupe les chapitres 2 à 7. La notion de variable aléatoire est traitée de manière analogue, d'abord dans le cas discret (fini ou dénombrable), puis dans le cas général. Puis l'espérance mathématique des variables aléatoires est introduite et les rapports entre l'espérance et l'intégrale de Lebesgue clarifiés, dans un cadre « abstrait », puis dans les cas particuliers de $\mathbb{R}$ (où $\mathbb{R}$ est l'ensemble des réels) et de $\mathbb{R}^n$, ce qui nous amène au chapitre 12.

Les chapitres 13 à 21 sont consacrés aux théorèmes-limite, en commençant par les outils analytiques nécessaires et notamment les fonctions caractéristiques (ou transformées de Fourier) et par une étude assez complète des variables aléatoires gaussiennes (normales) multidimensionnelles.

Nous introduisons ensuite l'espérance conditionnelle, en utilisant la méthode des projections dans l'espace de Hilbert des variables de carré intégrable, ce qui nous amène à consacrer le chapitre 22 à un exposé rapide des propriétés de base des espaces de Hilbert. Nous prolongeons l'espérance conditionnelle à toutes les variables intégrables ou positives, dans le chapitre 23. Enfin, dans les chapitres 24 à 28 un aperçu de la théorie des martingales est esquissé, avec une application au théorème de Radon-Nikodym. Ces cinq derniers chapitres sont sans doute « hors sujet » dans un livre consacré aux bases des probabilités ; nous les avons néanmoins inclus car la plupart des applications actuelles des probabilités utilisent la théorie des martingales, et aussi parce que cette théorie est une introduction naturelle au sujet des processus stochastiques.

Quelques exercices proposent la démonstration de propriétés énoncées mais non démontrées dans le corps du texte ; beaucoup sont des applications simples, permettant au lecteur de vérifier si les concepts sont bien assimilés ; enfin, certains proposent des extensions (en général assez simples) des résultats du texte lui-même. Les exercices flanqués d'un astérisque sont à notre avis les plus difficiles. Nous ne prétendons pour ces exercices à aucune originalité : la plupart se trouvent déjà dans d'autres livres ; un nombre appréciable d'entre eux est tiré du livre [14] de Allan Gut. Signalons aussi que la présentation de la théorie des martingales suit d'assez près celle du livre [1] de Richard Bass.

Aucune connaissance préalable des probabilités n'est requise, mais le lecteur est supposé maîtriser l'analyse élémentaire (séries, intégrale de Riemann, intégrales multiples) et un minimum d'algèbre linéaire.

Nous terminons cette introduction par une présentation générale – non mathématique – de la problématique et des principaux concepts des probabilités.

Les phénomènes aléatoires

Les phénomènes aléatoires sont des phénomènes dont on ne peut prévoir le résultat à l'avance, mais qui par « répétition » présentent un certain caractère de régularité. Un exemple typique est constitué par le jeu de pile ou face : on ne peut prévoir *a priori* le résultat d'un tirage particulier, mais si on fait un grand nombre de tirages on obtiendra une moyenne d'à peu près 50% de « pile », si la pièce n'est pas truquée.

La théorie des probabilités vise à fournir un modèle mathématique pour décrire ces phénomènes. Cette théorie contient trois ingrédients essentiels :

a) L'espace d'états : C'est l'ensemble, noté habituellement Ω, de tous les résultats possibles de l'expérience (aléatoire) qu'on réalise.

Exemples.

1) Un tirage à pile ou face : $\Omega = \{\mathrm{p},\mathrm{f}\}$.
2) Deux tirages successifs à pile ou face : $\Omega = \{\mathrm{pp},\mathrm{pf},\mathrm{fp},\mathrm{ff}\}$.
3) Tirage de deux dés : $\Omega = \{(i,j) : 1 \leqslant i \leqslant 6, 1 \leqslant j \leqslant 6\}$.
4) Mesure d'une longueur L, avec une erreur de mesure : $\Omega = \mathbb{R}_+$; $\omega \in \Omega$ désigne le résultat de la mesure, et $\omega - \mathrm{L}$ est l'erreur de mesure.
5) Durée de vie d'une ampoule électrique : $\Omega = \mathbb{R}_+$.

b) Les événements : Un événement est une propriété dont on peut dire si elle est vraie ou non, une fois l'expérience réalisée : lors du tirage de deux dés, par exemple, « le second dé donne 4 ou 5 », ou « le résultat du second dé est inférieur à celui du premier » sont des événements. En termes mathématiques, un événement est représenté par l'ensemble des résultats pour lesquels il est réalisé, c'est-à-dire par une partie de Ω. Si A et B sont deux événements,

- l'événement **contraire** de A est représenté par le complémentaire A^c ;
- l'événement « A **ou** B » est représenté par la réunion $\mathrm{A} \cup \mathrm{B}$;
- l'événement « A **et** B » est représenté par l'intersection $\mathrm{A} \cap \mathrm{B}$;
- l'événement **certain** est Ω ;
- l'événement **impossible** est l'ensemble vide $\emptyset$;
- un **événement élémentaire** est un « singleton », i.e. une partie $\{\omega\}$ ne contenant qu'un seul point ω de Ω.

On note $\mathcal{A}$ l'ensemble de tous les événements. Souvent (mais pas toujours : on verra pourquoi plus loin) $\mathcal{A}$ est l'ensemble, noté $\mathcal{P}(\Omega)$ ou 2^Ω, de toutes les parties de Ω. En tous cas, $\mathcal{A}$ doit être « stable » par les opérations logiques décrites ci-dessus : si $\mathrm{A}, \mathrm{B} \in \mathcal{A}$, alors on doit avoir $\mathrm{A}^c \in \mathcal{A}$, $\mathrm{A} \cap \mathrm{B} \in \mathcal{A}$, $\mathrm{A} \cup \mathrm{B} \in \mathcal{A}$, et aussi $\Omega \in \mathcal{A}$ et $\emptyset \in \mathcal{A}$.

c) La probabilité : À chaque événement A on associe un nombre, noté P(A) et appelé « probabilité de A ». Ce nombre mesure le degré de vraisemblance qu'on accorde *a priori* à A, avant la réalisation de l'expérience. Il est choisi entre 0 et 1, et il est d'autant plus près de 1 que l'événement est jugé plus vraisemblable.

Pour avoir une idée des propriétés de ces nombres, on peut imaginer la probabilité comme limite de « fréquences » : répétons n fois la même expérience ; les n résultats obtenus peuvent bien sûr être différents (penser à n jets successifs d'un même dé, par exemple). Notons $f_n(\mathrm{A})$ la fréquence de réalisation de l'événement A (i.e. le nombre de fois où il est réalisé, divisé par n). Alors, « intuitivement » on a :

$$\mathrm{P}(\mathrm{A}) \;=\; \text{limite de } \; f_n(\mathrm{A}) \; \text{ quand } \; n \uparrow +\infty. \tag{1.1}$$

(on donnera un sens précis à cette « limite » plus tard). Des propriétés évidentes des fréquences, on déduit immédiatement que :

1. $0 \leqslant \mathrm{P}(\mathrm{A}) \leqslant 1$, –
2. $\mathrm{P}(\Omega) = 1$, –

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

Un *espace de probabilité*, ou *espace probabilisé*, est donc un triplet $(\Omega, \mathcal{A}, P)$, constitué de l'espace Ω, de l'ensemble des événements $\mathcal{A}$, et de la famille des $P(A)$ pour $A \in \mathcal{A}$: on peut ainsi considérer P comme une application de $\mathcal{A}$ dans $[0, 1]$, qui vérifie au moins les propriétés (2) et (3) ci-dessus (plus une propriété supplémentaire, plus délicate à appréhender, et qui sera expliquée au chapitre suivant).

Une quatrième notion, importante également, quoique moins fondamentale, est celle de :

d) Variable aléatoire : Il s'agit là d'une grandeur qui dépend du résultat de l'expérience. En termes mathématiques, c'est une application de Ω dans un espace E, souvent $E = \mathbb{R}$ ou $E = \mathbb{R}^d$. **Attention :** cette terminologie, consacrée par l'usage, est assez malencontreuse : une « variable » aléatoire n'est pas une variable (au sens de l'analyse), mais une fonction ! c'est en fait une terminologie apparentée à la notion de variable en physique ou en sciences humaines, domaines où on désigne volontiers par « variable » la valeur prise par telle ou telle fonction de l'état du système.

Soit X une telle variable aléatoire, qui applique Ω dans E. On peut alors « transporter » la structure probabiliste sur l'espace d'arrivée E, en posant

$$P_X(B) = P(X^{-1}(B)) \quad \text{pour} \quad B \subset E,$$

où $X^{-1}(B)$ désigne l'image réciproque de B par X, c'est-à-dire l'ensemble des $\omega \in \Omega$ tels que $X(\omega) \in B$. L'événement $X^{-1}(B)$ correspond donc exactement à la propriété « $X \in B$ », et on écrit en général $P(X \in B)$ au lieu de $P(X^{-1}(B))$. Cette formule définit une nouvelle probabilité, notée P_X, cette fois-ci sur l'espace E (au lieu de Ω). Cette probabilité P_X s'appelle la **loi de la variable** X.

Exemple (tirage de deux dés). On a vu que $\Omega = \{(i, j) : 1 \leqslant i \leqslant 6, 1 \leqslant j \leqslant 6\}$, et il est naturel de prendre ici $\mathcal{A} = 2^{\Omega}$, et

$$P(A) = \frac{\text{card}(A)}{36} \qquad \text{si} \;\; A \subset \Omega,$$

où card(A) désigne le nombre de points contenus dans A. On vérifie aisément les propriétés (1), (2), (3) ci-dessus, et on a $P(\{\omega\}) = \frac{1}{36}$ pour chaque singleton. L'application $X : \Omega \to \mathbb{N}$ (où $\mathbb{N}$ est l'ensemble des entiers positifs ou nul) définie par $X(i, j) = i + j$ est la variable aléatoire « somme des deux dés », de loi

$$P_X(B) = \frac{\text{nombre de couple } (i, j) \text{ tels que } i + j \in B}{36}$$

(par exemple $P_X(\{2\}) = P_X(\{12\}) = \frac{1}{36}$, $P_X(\{3\}) = \frac{2}{36}$, etc.).

Nous allons formaliser la notion d'espace de probabilité au chapitre 2, tandis que les variables aléatoires seront introduites rigoureusement dans les chapitres 5 et 8.

Chapitre 2

Axiomes des probabilités

Nous commençons par présenter ci-dessous les propriétés minimales qui nous sont nécessaires pour définir une probabilité. Nous espérons que le lecteur se convaincra que les deux axiomes dans la définition 2.3 ci-dessous sont raisonnables, notamment au vu de l'approche par les fréquences évoquée lors de l'introduction de la formule (1.1), et la théorie toute entière découlera de ces deux axiomes simples. Toutefois, avant de les présenter nous devons introduire le concept de tribu.

Soit Ω un espace « abstrait », c'est-à-dire sans structure particulière. Rappelons que $\mathcal{P}(\Omega)$ ou 2^Ω désigne l'ensemble de tous les sous-ensembles de Ω, y compris le sous-ensemble vide noté $\emptyset$. Étant donnée une partie $\mathcal{A}$ de 2^Ω, on considère les propriétés suivantes :

(1) $\emptyset \in \mathcal{A}$ et $\Omega \in \mathcal{A}$;

(2) $\mathcal{A}$ est *stable par complémentation* : si $\mathrm{A} \in \mathcal{A}$ alors $\mathrm{A}^c \in \mathcal{A}$, où A^c désigne le complémentaire de A (sous-entendu : dans Ω) ;

(3) $\mathcal{A}$ est *stable pour les réunions finies et les intersections finies* : si les A_i sont tous dans $\mathcal{A}$, alors $\bigcup_{i=1}^n \mathrm{A}_i$ et $\bigcap_{i=1}^n \mathrm{A}_i$ sont aussi dans $\mathcal{A}$ (il suffit pour cela que $\mathcal{A}$ soit stable par réunion et intersection de *deux* éléments quelconques) ;

(4) $\mathcal{A}$ est stable pour les réunions dénombrables et les intersections dénombrables : si les A_i sont tous dans $\mathcal{A}$, alors $\bigcup_{i=1}^\infty \mathrm{A}_i$ et $\bigcap_{i=1}^\infty \mathrm{A}_i$ sont aussi dans $\mathcal{A}$:

Définition 2.1. *$\mathcal{A}$ est une* algèbre *si elle vérifie* (1), (2) *et* (3) *ci-dessus. C'est une* tribu *(ou une σ-algèbre) si elle vérifie* (1), (2) *et* (4) *ci-dessus.*

Noter que si on a (2), alors (1) peut être remplacé soit par (1′) : $\emptyset \in \mathcal{A}$, soit par (1″) : $\Omega \in \mathcal{A}$. Noter aussi que (1)+(4) implique (3), de sorte que toute tribu est une algèbre (mais une algèbre peut ne pas être une tribu : voir l'exercice 17).

Définition 2.2. *Si $\mathcal{C} \subset 2^\Omega$, la* tribu engendrée par *$\mathcal{C}$ et notée $\sigma(\mathcal{C})$ est la plus petite tribu contenant $\mathcal{C}$.* (Elle existe toujours ; en effet l'intersection d'une famille quelconque de tribus est encore une tribu : voir l'exercice 2 ; on applique alors ce résultat à la famille de toutes les tribus contenant $\mathcal{C}$, famille qui contient au moins la tribu 2^Ω.)

Exemples.

(i) $\mathcal{A} = \{\emptyset, \Omega\}$ (la tribu « triviale »).

(ii) Si $\mathrm{A} \subset \Omega$, alors $\sigma(\{\mathrm{A}\}) = \{\emptyset, \mathrm{A}, \mathrm{A}^c, \Omega\}$.

(iii) Si $\Omega = \mathbb{R}$ (ou plus généralement si Ω est un espace topologique, un cas que nous traiterons au chapitre 8), la *tribu borélienne*, ou de Borel, est la tribu $\mathcal{B}$ engendrée par la classe des ouverts (ou de manière équivalente, par la classe des fermés) ; une partie de $\mathbb{R}$ qui est dans la tribu borélienne s'appelle un *borélien*.

Théorème 2.1. *La tribu borélienne $\mathcal{B}$ de $\mathbb{R}$ est engendrée par les intervalles de la forme $]-\infty, a]$, où $a \in \mathbb{Q}$* ($\mathbb{Q}$ = ensemble des rationnels).

Preuve. Soit $\mathcal{C}$ la classe de tous les intervalles ouverts. Tout ouvert A de $\mathbb{R}$ est réunion dénombrable d'intervalles ouverts : en effet pour tout $x \in \mathrm{A}$ on note $\mathrm{I}_x =]a_x, b_x[$ le plus grand intervalle ouvert contenant x et inclus dans A (on prend $b_x = \sup(y : y > x, [x, y[\subset \mathrm{A})$ et $a_x = \inf(y : y < x,]y, x] \subset \mathrm{A}))$; alors A est la réunion des I_x lorsque x décrit l'ensemble des rationnels contenus dans A. Cette propriété, jointe à l'axiome (4), entraine que $\sigma(\mathcal{C}) = \mathcal{B}$, tribu borélienne de $\mathbb{R}$.

Soit $\mathcal{D}$ la classe des intervalles de la forme $]-\infty, a]$, où $a \in \mathbb{Q}$. Soit $]a, b[\in \mathcal{C}$; il existe des rationnels $(a_n)_{n \geqslant 1}$ décroissant vers a et des rationnels $(b_n)_{n \geqslant 1}$ croissant strictement vers b, de sorte que

$$
\begin{aligned}
]a, b[&= \bigcup_{n=1}^{\infty}]a_n, b_n] \\
&= \bigcup_{n=1}^{\infty} \left( ]-\infty, b_n] \cap]-\infty, a_n]^c \right).
\end{aligned}
$$

Par suite $\mathcal{C} \subset \sigma(\mathcal{D})$, donc $\sigma(\mathcal{C}) \subset \sigma(\mathcal{D})$. Par ailleurs chaque élément de $\mathcal{D}$ est un fermé, donc un borélien, ce qui entraîne $\sigma(\mathcal{D}) \subset \mathcal{B}$. Le résultat découle alors de :

$$\mathcal{B} = \sigma(\mathcal{C}) \subset \sigma(\mathcal{D}) \subset \mathcal{B}. \qquad \blacksquare$$

Sur l'espace d'états Ω, la classe des événements sera toujours une tribu $\mathcal{A}$: les axiomes (1), (2) et (3) correspondent aux opérations « logiques » décrites dans le chapitre 1. L'axiome (4) est nécessaire pour assurer la cohérence interne de la théorie (par exemple, pour donner un sens mathématique précis à la convergence (1.1)), et aussi pour assurer sa richesse mathématique (par exemple si une suite (A_n) d'événements

« converge » vers une limite A, en un sens précisé plus loin, alors on veut que A soit aussi un événement).

On peut maintenant définir la notion de probabilité :

Définition 2.3. *Une* mesure de probabilité, *ou simplement une* probabilité, *définie sur une tribu $\mathcal{A}$ d'un espace Ω, est une application* $\mathrm{P} : \mathcal{A} \to [0,1]$ *qui vérifie :*

1. $\mathrm{P}(\Omega) = 1$.
2. *Pour toute suite* $\mathrm{A}_1, \mathrm{A}_2, \mathrm{A}_3, \ldots$ *d'éléments de $\mathcal{A}$ qui sont deux à deux disjoints (i.e. $\mathrm{A}_n \cap \mathrm{A}_m = \emptyset$ si $n \neq m$), on a*

$$\mathrm{P}\left(\bigcup_{n=1}^{\infty} \mathrm{A}_n\right) = \sum_{n=1}^{\infty} \mathrm{P}(\mathrm{A}_n).$$

Le nombre $\mathrm{P}(\mathrm{A})$ s'appelle la *probabilité* de l'événement A. L'axiome 2 ci-dessus s'appelle *σ-additivité*; voir l'exercice 9 pour un affaiblissement de l'axiome 1.

Dans cette définition on pourrait imaginer une condition plus simple que 2, à savoir que

$$\mathrm{A}, \mathrm{B} \in \mathcal{A},\ \mathrm{A} \cap \mathrm{B} = \emptyset \quad \Longrightarrow \quad \mathrm{P}(\mathrm{A} \cup \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B}). \tag{2.1}$$

Cette condition s'appelle *l'additivité* et, par une récurrence élémentaire, on voit qu'elle entraîne que pour toute famille *finie* $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_m$ d'éléments deux à deux disjoints de $\mathcal{A}$, on a

$$\mathrm{P}\left(\bigcup_{n=1}^{m} \mathrm{A}_n\right) = \sum_{n=1}^{m} \mathrm{P}(\mathrm{A}_n).$$

Théorème 2.2. *Si* P *est une probabilité sur la tribu $\mathcal{A}$, alors :*

(i) *on a* $\mathrm{P}(\emptyset) = 0$ *;*

(ii) P *est additive.*

Preuve. Si dans l'axiome 2 on prend $\mathrm{A}_n = \emptyset$ pour tout n, on voit que le nombre $a = \mathrm{P}(\emptyset)$ est égal à $a + a + a + \cdots$; comme par ailleurs $0 \leqslant a \leqslant 1$, ce n'est possible que si $a = 0$, et on a (i). Pour (ii) il suffit d'appliquer l'axiome 2 avec $\mathrm{A}_1 = \mathrm{A}$, $\mathrm{A}_2 = \mathrm{B}$ et $\mathrm{A}_3 = \mathrm{A}_4 = \cdots = \emptyset$, plus le fait que $\mathrm{P}(\emptyset) = 0$, pour obtenir (2.1). ■

À l'inverse, la σ-additivité n'est pas impliquée par l'additivité. En fait l'additivité, malgré son caractère intuitif, n'est pas suffisante pour traiter mathématiquement les problèmes de la théorie, même dans des cas aussi simples qu'une suite de jets de dés, comme nous le verrons plus loin.

Le théorème ci-dessous exprime ce qu'il faut exactement ajouter à l'additivité pour obtenir la σ-additivité. Afin de rendre son énoncé plus clair, nous présentons d'abord deux conséquences immédiates de l'additivité et de l'axiome $P(\Omega) = 1$:

$$A \in \mathcal{A} \;\Rightarrow\; P(A^c) = 1 - P(A)\;;$$
$$A, B \in \mathcal{A},\; A \subset B \;\Rightarrow\; P(A) \leqslant P(B)$$

(pour la première, appliquer (2.1) à A et $B = A^c$, et utiliser $P(\Omega) = 1$; pour la seconde appliquer (2.1) à A et $C = A^c \cap B$, donc $A \cap C = \emptyset$ et $A \cup C = B$).

Ci-dessous on note $A_n \downarrow A$ et on dit que A_n *décroît* (ou « tend en décroissant ») vers A si $A_n \supset A_{n+1}$ pour tout n et si $\cap_n A_n = A$. On note $A_n \uparrow A$ et on dit que A_n *croît* (ou « tend en croissant ») vers A si $A_n \subset A_{n+1}$ pour tout n et si $\cup_n A_n = A$. Noter que si les A_n sont dans la tribu $\mathcal{A}$, il en est de même de A dans les deux cas ci-dessus.

Théorème 2.3. *Soit $\mathcal{A}$ une tribu de Ω. Supposons que $P : \mathcal{A} \to [0, 1]$ satisfasse l'axiome 1 et soit* additive. *Les assertions suivantes sont alors équivalentes :*

(i) *on a l'axiome 2 (σ-*additivité*) ;*

(ii) *si $A_n \in \mathcal{A}$ et $A_n \downarrow \emptyset$, alors $P(A_n) \downarrow 0$;*

(iii) *si $A_n \in \mathcal{A}$ et $A_n \downarrow A$, alors $P(A_n) \downarrow P(A)$;*

(iv) *si $A_n \in \mathcal{A}$ et $A_n \uparrow \Omega$, alors $P(A_n) \uparrow 1$;*

(v) *si $A_n \in \mathcal{A}$ et $A_n \uparrow A$, alors $P(A_n) \uparrow P(A)$.*

Preuve. Observons que $A_n \downarrow A$ implique $A_n^c \uparrow A^c$. Comme $P(A_n^c) = 1 - P(A_n)$ et $P(A^c) = 1 - P(A)$, on voit que (ii) et (iv) sont équivalents, ainsi que (iii) et (v). L'implication (v)$\Rightarrow$(iv) est évidente.

Supposons maintenant (iv). Soit $A_n \in \mathcal{A}$ avec $A_n \uparrow A$. Posons $B_n = A_n \cup A^c$. Alors B_n croît vers Ω, de sorte que $P(B_n) \uparrow 1$. Comme $A_n \subset A$, on a $A_n \cap A^c = \emptyset$, donc $P(A_n \cup A^c) = P(A_n) + P(A^c)$ et il vient

$$1 = \lim_{n\to\infty} P(B_n) = \lim_{n\to\infty} \{P(A_n) + P(A^c)\}.$$

Par suite $\lim_{n\to\infty} P(A_n) = 1 - P(A^c) = P(A)$, et on a (v).

Il reste à montrer que (i) et (v) sont équivalents. Supposons d'abord (v). Soit des $A_n \in \mathcal{A}$ deux à deux disjoints, et définissons $B_n =$

$\bigcup_{1 \leqslant p \leqslant n} A_p$ et $B = \bigcup_{n=1}^{\infty} A_n$. L'additivité implique $P(B_n) = \sum_{p=1}^{n} P(A_p)$, qui croît vers $\sum_{n=1}^{\infty} P(A_n)$ quand $n \to \infty$, et croît aussi vers $P(B)$ par (v), donc

$$P(B) = P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n),$$

et (i) est prouvé.

Finalement, supposons (i). Soit $A_n \in \mathcal{A}$, avec A_n croissant vers A. On construit une nouvelle suite d'éléments de $\mathcal{A}$ ainsi :

$$\begin{aligned} &B_1 = A_1, \\ &B_2 = A_2 \setminus A_1 = A_2 \cap (A_1^c), \\ &\vdots \\ &B_n = A_n \setminus A_{n-1}. \end{aligned}$$

On a $\bigcup_{n=1}^{\infty} B_n = A$ et les $(B_n)_{n \geqslant 1}$ sont deux à deux disjoints. Donc (i) entraîne

$$P(A) = \lim_{n \to \infty} \sum_{p=1}^{n} P(B_p).$$

On a aussi $\sum_{p=1}^{n} P(B_p) = P(A_n)$, donc $\lim_{n \to \infty} P(A_n) = P(A)$ et on a (v). ■

Si A est une partie de Ω, on définit sa *fonction indicatrice* par

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

On dit que A_n *converge* vers A (et on écrit $A_n \to A$) si $\lim_{n \to \infty} 1_{A_n}(\omega) = 1_A(\omega)$ pour tout $\omega \in \Omega$. Noter que si la suite A_n croît (resp. décroît) vers A, elle tend aussi vers A au sens ci-dessus.

On associe aussi à toute suite A_n de parties de Ω les ensembles suivants :

$$\left.\begin{aligned} \limsup_{n \to \infty} A_n &= \bigcap_{n=1}^{\infty} \bigcup_{m \geqslant n} A_m, \\ \liminf_{n \to \infty} A_n &= \bigcup_{n=1}^{\infty} \bigcap_{m \geqslant n} A_m. \end{aligned}\right\} \tag{2.2}$$

Par suite $A_n \to A$ si et seulement si $A = \limsup_{n\to\infty} A_n = \liminf_{n\to\infty} A_n$: en effet, l'indicatrice d'une réunion (resp. d'une intersection) d'ensembles est le sup (resp. l'inf) des indicatrices, donc

$$1_{\limsup_{n\to\infty} A_n} = \limsup_{n\to\infty} 1_{A_n}, \qquad 1_{\liminf_{n\to\infty} A_n} = \liminf_{n\to\infty} 1_{A_n},$$

et par ailleurs on sait qu'une suite de réels u_n converge vers une limite u si et seulement si $u = \limsup_{n\to\infty} u_n = \limsup_{n\to\infty} u_n$.

Noter que $\limsup_{n\to\infty} A_n$ est l'ensemble des ω qui appartiennent à une infinité d'ensembles A_n, tandis que $\liminf_{n\to\infty} A_n$ est l'ensemble des ω qui appartiennent à tous les A_n sauf au plus à un nombre fini d'entre eux.

Théorème 2.4. *Soit* P *une probabilité et* A_n *une suite d'éléments de* $\mathcal{A}$ *qui converge vers* A. *Alors* $A \in \mathcal{A}$ *et* $\lim_{n\to\infty} P(A_n) = P(A)$.

Preuve. $\mathcal{A}$ étant une tribu, on a $\limsup_{n\to\infty} A_n \in \mathcal{A}$ et $\liminf_{n\to\infty} A_n \in \mathcal{A}$ (voir l'exercice 4). Comme $A_n \to A$ on a donc aussi $A \in \mathcal{A}$.

Ensuite, soit $B_n = \bigcap_{m\geqslant n} A_m$ et $C_n = \bigcup_{m\geqslant n} A_m$. Alors B_n croît vers A et C_n décroît vers A, donc $\lim_{n\to\infty} P(B_n) = \lim_{n\to\infty} P(C_n) = P(A)$ d'après le théorème 2.3. Mais $B_n \subset A_n \subset C_n$, donc $P(B_n) \leqslant P(A_n) \leqslant P(C_n)$, de sorte que $\lim_{n\to\infty} P(A_n) = P(A)$. ■

Exercices

1. Soit Ω un ensemble fini. Montrer que la tribu 2^Ω de toutes les parties de Ω est également finie.
2. Soit $(\mathcal{G}_\alpha)_{\alpha\in A}$ une famille quelconque de tribus sur l'ensemble Ω. Montrer que $\mathcal{H} = \bigcap_{\alpha\in A} \mathcal{G}_\alpha$ est aussi une tribu.
3. Soit $(A_n)_{n\geqslant 1}$ une suite de parties de Ω. Montrer que
   a) $(\bigcup_{n=1}^{\infty} A_n)^c = \bigcap_{n=1}^{\infty} A_n^c$
   b) $(\bigcap_{n=1}^{\infty} A_n)^c = \bigcup_{n=1}^{\infty} A_n^c$.
4. Soit $\mathcal{A}$ une tribu et $(A_n)_{n\geqslant 1}$ une suite d'éléments de $\mathcal{A}$. Avec les notations (2.2), montrer que

$$\liminf_{n\to\infty} A_n \in \mathcal{A}; \qquad \limsup_{n\to\infty} A_n \in \mathcal{A}; \qquad \liminf_{n\to\infty} A_n \subset \limsup_{n\to\infty} A_n.$$

5. Soit $(\mathrm{A}_n)_{n\geqslant 1}$ une suite de parties de Ω. Montrer que

$$\limsup_{n\to\infty} 1_{\mathrm{A}_n} - \liminf_{n\to\infty} 1_{\mathrm{A}_n} = 1_{\{\limsup_n \mathrm{A}_n \setminus \liminf_n \mathrm{A}_n\}}$$

(avec la notation $\mathrm{A}\setminus\mathrm{B} = \mathrm{A}\cap\mathrm{B}^c$ si $\mathrm{B}\subset\mathrm{A}$).

6. Soit $\mathcal{A}$ une tribu sur Ω et $\mathrm{B}\in\mathcal{A}$. Montrer que $\mathcal{F} = \{\mathrm{A}\cap\mathrm{B} : \mathrm{A}\in\mathcal{A}\}$ est une tribu sur l'ensemble B (appelée la « tribu trace » de $\mathcal{A}$ sur B). Est-ce encore vrai si B est une partie de Ω qui n'appartient pas à $\mathcal{A}$?

7. Soit f une application de Ω dans un autre espace E muni lui aussi d'une tribu $\mathcal{E}$. Soit $\mathcal{A} = \{\mathrm{A}\subset\Omega : \text{il existe } \mathrm{B}\in\mathcal{E} \text{ avec } \mathrm{A} = f^{-1}(\mathrm{B})\}$. Montrer que $\mathcal{A}$ est une tribu de Ω.

8. Soit $f : \mathbb{R}\to\mathbb{R}$ une fonction continue, et soit $\mathcal{A} = \{\mathrm{A}\subset\mathbb{R} : \text{il existe } \mathrm{B}\in\mathcal{B} \text{ avec } \mathrm{A} = f^{-1}(\mathrm{B})\}$, où $\mathcal{B}$ est la tribu borélienne de l'espace d'arrivée $\mathbb{R}$. Montrer que $\mathcal{A}\subset\mathcal{B}$, tribu borélienne de l'espace de départ $\mathbb{R}$.

Pour les problèmes 9 à 16 nous supposons donnés un espace Ω, une tribu $\mathcal{A}$ de cet espace, et une probabilité P *sur cette tribu ; les parties* A, B, A_i, *etc. sont toutes supposées appartenir à $\mathcal{A}$.*

9. Si $\mathrm{A}\cap\mathrm{B} = \emptyset$, montrer que $\mathrm{P}(\mathrm{A}\cup\mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B})$.

10. Montrer que $\mathrm{P}(\mathrm{A}\cup\mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B}) - \mathrm{P}(\mathrm{A}\cap\mathrm{B})$.

11. Montrer que $\mathrm{P}(\mathrm{A}) = 1 - \mathrm{P}(\mathrm{A}^c)$.

12. Montrer que $\mathrm{P}(\mathrm{A}\cap\mathrm{B}^c) = \mathrm{P}(\mathrm{A}) - \mathrm{P}(\mathrm{A}\cap\mathrm{B})$.

13. (Identité de Poincaré.) Montrer que

$$\mathrm{P}\left(\bigcup_{i=1}^n \mathrm{A}_i\right) = \sum_i \mathrm{P}(\mathrm{A}_i) - \sum_{i<j}\mathrm{P}(\mathrm{A}_i\cap\mathrm{A}_j) + \sum_{i<j<k}\mathrm{P}(\mathrm{A}_i\cap\mathrm{A}_j\cap\mathrm{A}_k) - \cdots + (-1)^{n+1}\mathrm{P}(\mathrm{A}_1\cap\mathrm{A}_2\cap\cdots\cap\mathrm{A}_n)$$

où, par exemple, $\sum_{i<j}$ signifie la somme pour toutes les paires d'indices (i,j) avec $1\leqslant i<j\leqslant n$.

14. Montrer que si $\mathrm{P}(\mathrm{A}) = \frac{3}{4}$ et $\mathrm{P}(\mathrm{B}) = \frac{1}{3}$, on a nécessairement $\frac{1}{12}\leqslant \mathrm{P}(\mathrm{A}\cap\mathrm{B})\leqslant\frac{1}{3}$.

15. (Sous-additivité.) Montrer que

$$\mathrm{P}\left(\bigcup_{i=1}^n \mathrm{A}_i\right) \leqslant \sum_{i=1}^n \mathrm{P}(\mathrm{A}_i),$$

pour chaque n, et aussi que

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

16. (Inégalités de Bonferroni.) Montrer que

$$\text{a) } P\left(\bigcup_{i=1}^{n} A_i\right) \geqslant \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j).$$

$$\text{b) } P\left(\bigcup_{i=1}^{n} A_i\right) \leqslant \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k).$$

17. Soit Ω un ensemble infini (dénombrable ou non) et $\mathcal{A}$ l'ensemble des parties de Ω qui sont finies, ou de complémentaire fini. Montrer que $\mathcal{A}$ est une algèbre, mais pas une tribu.

Chapitre 3

Probabilités conditionnelles et indépendance

Soit A et B deux événements d'un espace probabilisé ; rappelons « l'approche par les fréquences » : la fréquence $f_n(\mathrm{A})$ de réalisation de l'événement A lorsqu'on répète n fois la même expérience converge (en un sens à préciser !) vers la probabilité $\mathrm{P}(\mathrm{A})$, qui quantifie « les chances de voir A réalisé ».

Supposons maintenant qu'on sache que l'événement B est réalisé. Les chances de voir A réalisé vont changer et être quantifiées par un nouveau nombre $\mathrm{P}(\mathrm{A}\,|\,\mathrm{B})$, la « probabilité de A sachant B » : on peut à nouveau considérer la fréquence de A lorsque l'expérience est répétée n fois, sauf qu'il faut calculer cette fréquence sur l'ensemble des expériences où B se réalise. Il est ainsi naturel de considérer le nombre de fois où A *et* B sont réalisés, c'est-à-dire $nf_n(\mathrm{A}\cap\mathrm{B})$; pour obtenir une fréquence, il convient de diviser ce nombre par le nombre d'occurrences de B, i.e. $nf_n(\mathrm{B})$, de sorte qu'en définitive on doit avoir

$$\mathrm{P}(\mathrm{A}\,|\,\mathrm{B}) \approx \frac{nf_n(\mathrm{A}\cap\mathrm{B})}{nf_n(\mathrm{B})} = \frac{f_n(\mathrm{A}\cap\mathrm{B})}{f_n(\mathrm{B})}.$$

La définition 3.2 ci-après est tirée de cette relation en « prenant la limite en n ».

En particulier, il se peut que le fait de savoir que B est réalisé ne donne aucune information sur la réalisation de A : on dit alors que A est « indépendant » de B, et cela se traduit par le fait que $\mathrm{P}(\mathrm{A}\,|\,\mathrm{B}) = \mathrm{P}(\mathrm{A})$. Compte tenu de la définition 3.2, cela se traduit par

$$\frac{\mathrm{P}(\mathrm{A}\cap\mathrm{B})}{\mathrm{P}(\mathrm{B})} = \mathrm{P}(\mathrm{A}), \qquad \text{donc} \qquad \mathrm{P}(\mathrm{A}\cap\mathrm{B}) = \mathrm{P}(\mathrm{A})\mathrm{P}(\mathrm{B}).$$

Nous allons maintenant introduire les définitions proprement dites, en commençant par l'indépendance.

Définition 3.1. (a) *Deux événements* A *et* B *sont dits* indépendants *s'ils vérifient* $\mathrm{P}(\mathrm{A}\cap\mathrm{B}) = \mathrm{P}(\mathrm{A})\mathrm{P}(\mathrm{B})$.

(b) *Soit une famille (finie ou infinie) d'événements* $(A_i)_{i \in I}$. *Ces événements sont dits* indépendants, *ou parfois « mutuellement indépendants »*, *si pour toute partie finie* J *de* I *on a*

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

Attention : si les $(A_i)_{i \in I}$ sont indépendants, ils sont aussi *deux à deux indépendants*, ce qui signifie que A_i et A_j sont indépendants pour tous i, j avec $i \neq j$, mais la réciproque est fausse.

Théorème 3.1. *Si* A *et* B *sont indépendants, il en est de même de* A *et* B^c, *de* A^c *et* B, *et de* A^c *et* B^c.

Preuve. Pour A et B^c, on a

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)(1 - P(B)) \\ &= P(A)P(B^c). \end{aligned}$$

Les autres propriétés se montrent de manière analogue. ■

Exemples.

1. On jette 3 fois une pièce. Si A_i est l'événement « le i-ième jet donne pile » il est habituel de choisir la probabilité de sorte que les événements A_1, A_2 et A_3 soient indépendants.
2. On choisit une carte au hasard parmi 52 cartes. A = {la carte est un cœur}, et B = {la carte est une dame}. Un modèle naturel pour cette expérience consiste à affecter la probabilité $\frac{1}{52}$ au choix de chacune des cartes. Par additivité, $P(A) = \frac{13}{52}$ et $P(B) = \frac{4}{52}$ et $P(A \cap B) = \frac{1}{52}$, donc A et B sont indépendants.
3. Soit $\Omega = \{1, 2, 3, 4\}$ et $\mathcal{A} = 2^\Omega$. Soit $P(i) = \frac{1}{4}$ pour $i = 1, 2, 3, 4$. Soit enfin $A = \{1, 2\}$, $B = \{1, 3\}$, $C = \{2, 3\}$. Alors A, B, C sont deux à deux indépendants, mais pas (mutuellement) indépendants.

Définition 3.2. *Soit* A, B *deux événements, avec* $P(B) > 0$. *La probabilité conditionnelle de* A *sachant* B *est* $P(A \mid B) = P(A \cap B)/P(B)$.

Théorème 3.2. *Supposons que* $P(B) > 0$.

(a) A *et* B *sont indépendants si et seulement si* $P(A \mid B) = P(A)$.

(b) *L'application* $A \to P(A \mid B)$ *de* $\mathcal{A}$ *dans* $[0, 1]$ *définit une nouvelle mesure de probabilité sur* $\mathcal{A}$, *appelée la « probabilité conditionnelle sachant* B *».*

Preuve. (a) résulte d'un calcul immédiat. Pour (b), posons $Q(A) = P(A \mid B)$, avec B fixé. On doit montrer que Q vérifie (1) et (2) de la définition 2.3. Mais (1) provient de

$$Q(\Omega) = P(\Omega \mid B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$$

Quant à (2), on observe que si $(A_n)_{n\geqslant 1}$ est une suite d'éléments de $\mathcal{A}$ qui sont deux à deux disjoints, alors

$$Q\left(\bigcup_{n=1}^{\infty} A_n\right) = P\left(\bigcup_{n=1}^{\infty} A_n \,\middle|\, B\right) = \frac{P\left(\left(\bigcup_{n=1}^{\infty} A_n\right) \cap B\right)}{P(B)} = \frac{P\left(\bigcup_{n=1}^{\infty} (A_n \cap B)\right)}{P(B)}$$

et les $(A_n \cap B)_{n\geqslant 1}$ sont également deux à deux disjoints ; donc

$$= \sum_{n=1}^{\infty} \frac{P(A_n \cap B)}{P(B)} = \sum_{n=1}^{\infty} P(A_n \mid B) = \sum_{n=1}^{\infty} Q(A_n). \qquad \blacksquare$$

Les trois résultats suivants, quoique élémentaires, sont extrêmement utiles, et en particulier le second d'entre eux.

Théorème 3.3. (Théorème des probabilités composées.) *Si* $A_1, \ldots, A_n \in \mathcal{A}$ *et si* $P(A_1 \cap \cdots \cap A_{n-1}) > 0$, *alors*

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \ldots P(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

Preuve. La preuve se fait par récurrence. Pour $n = 2$, le résultat est simplement la définition 3.2. Supposons le résultat correct pour $n-1$ événements. Soit $B = A_1 \cap \cdots \cap A_{n-1}$. La définition 3.2 donne $P(B \cap A_n) = P(A_n \mid B)P(B)$; puis on remplace $P(B)$ par sa valeur, donnée par l'hypothèse de récurrence :

$$P(B) = P(A_1)P(A_2 \mid A_1) \ldots P(A_{n-1} \mid A_1 \cap \cdots \cap A_{n-2}),$$

et le résultat s'ensuit. $\blacksquare$

Une famille $(E_i)_{i\in I}$ de parties de Ω est appelée une *partition* si les E_i sont deux à deux disjointes et si $\bigcup_{i\in I} E_i = \Omega$.

Théorème 3.4. (Formule des probabilités totales.) *Soit* (E_n) *une partition* finie ou dénombrable *de* Ω, *constituée d'éléments de la tribu* $\mathcal{A}$. *Pour tout* $A \in \mathcal{A}$ *on a*

$$P(A) = \sum_n P(A \mid E_n)P(E_n).$$

Preuve. Observons que

$$\mathrm{A} = \mathrm{A} \cap \Omega = \mathrm{A} \cap \Big(\bigcup_n \mathrm{E}_n\Big) = \bigcup_n (\mathrm{A} \cap \mathrm{E}_n).$$

Les E_n étant deux à deux disjoints, il en est de même des $(\mathrm{A} \cap \mathrm{E}_n)_{n \geq 1}$, donc

$$\mathrm{P(A)} = \mathrm{P}\Big(\bigcup_n (\mathrm{A} \cap \mathrm{E}_n)\Big) = \sum_n \mathrm{P}(\mathrm{A} \cap \mathrm{E}_n) = \sum_n \mathrm{P}(\mathrm{A} \mid \mathrm{E}_n)\mathrm{P}(\mathrm{E}_n). \qquad \blacksquare$$

Exemple. On dispose de deux urnes, la première contenant 1 boule blanche et 3 noires, la seconde contenant 2 boules blanches et 2 noires. L'expérience consiste à choisir d'abord « au hasard » une urne, puis à tirer « au hasard » une boule dans l'urne choisie. On cherche la probabilité de l'événement A = « on tire une boule noire ».

L'espace Ω est $\{(i,j) : i = 1,2; j = 1,2,3,4\}$, où i est le numéro de l'urne et j le numéro de la boule dans l'urne. On considère la partition $(\mathrm{E}_1, \mathrm{E}_2)$ de Ω définie par $\mathrm{E}_i = \{(i,j) : j = 1,2,3,4\}$. Le mode opératoire conduit naturellement à supposer que $\mathrm{P}(\mathrm{E}_i) = 1/2$ pour $i = 1,2$, et aussi que $\mathrm{P}(\{(i,j)\} \mid \mathrm{E}_{i'})$ vaut $1/4$ si $i = i'$ et 0 sinon. Si on convient que dans chaque urne les boules noires ont les premiers numéros, on a $\mathrm{A} = \{(1,1),(1,2),(1,3),(2,1),(2,2)\}$. Comme la probabilité conditionnelle est aussi une mesure de probabilité (théorème 3.2(b)), donc est additive, on a $\mathrm{P}(\mathrm{A} \mid \mathrm{E}_1) = 3/4$ et $\mathrm{P}(\mathrm{A} \mid \mathrm{E}_2) = 1/2$. D'après la formule des probabilités totales il vient alors

$$\mathrm{P(A)} = \mathrm{P}(\mathrm{A} \mid \mathrm{E}_1)\mathrm{P}(\mathrm{E}_1) + \mathrm{P}(\mathrm{A} \mid \mathrm{E}_2)\mathrm{P}(\mathrm{E}_2) = \frac{3}{4} \times \frac{1}{2} + \frac{2}{4} \times \frac{1}{2} = \frac{5}{8}.$$

Théorème 3.5. (Théorème de Bayes, ou de « probabilité des causes ».) *Soit* (E_n) *une partition* finie ou dénombrable *de* Ω, *constituée d'éléments de la tribu* $\mathcal{A}$. *Pour tout* $\mathrm{A} \in \mathcal{A}$ *tel que* $\mathrm{P(A)} > 0$ *on a*

$$\mathrm{P}(\mathrm{E}_n \mid \mathrm{A}) = \frac{\mathrm{P}(\mathrm{A} \mid \mathrm{E}_n)\mathrm{P}(\mathrm{E}_n)}{\sum_m \mathrm{P}(\mathrm{A} \mid \mathrm{E}_m)\mathrm{P}(\mathrm{E}_m)}.$$

Preuve. Au vu du théorème 3.4 le dénominateur vaut

$$\sum_m \mathrm{P}(\mathrm{A} \mid \mathrm{E}_m)\mathrm{P}(\mathrm{E}_m) = \mathrm{P(A)}.$$

La formule devient donc

$$\frac{\mathrm{P}(\mathrm{A} \mid \mathrm{E}_n)\mathrm{P}(\mathrm{E}_n)}{\mathrm{P(A)}} = \frac{\mathrm{P}(\mathrm{A} \cap \mathrm{E}_n)}{\mathrm{P(A)}} = \mathrm{P}(\mathrm{E}_n \mid \mathrm{A}). \qquad \blacksquare$$

Ce théorème a des conséquences profondes en probabilités et en statistique : voir par exemple l'exercice 6.

Exercices

Dans tous les exercices l'espace de probabilité est fixé, et A, B, A_n, etc. sont des événements.

1. Montrer que si $A \cap B = \emptyset$, alors A et B *ne sont pas* indépendants, sauf si $P(A) = 0$ ou $P(B) = 0$.
2. Si $P(C) > 0$, montrer que $P(A \cup B \mid C) = P(A \mid C) + P(B \mid C) - P(A \cap B \mid C)$.
3. Si $P(C) > 0$ et si les A_i sont deux à deux disjoints, montrer que
$$P\left(\bigcup_{i=1}^{n} A_i \mid C\right) = \sum_{i=1}^{n} P(A_i \mid C).$$
4. Si $P(B) > 0$, montrer que $P(A \cap B) = P(A \mid B)P(B)$.
5. Si $0 < P(B) < 1$, montrer que
$$P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c).$$
6. Chaque don de sang est soumis à un test du SIDA. On suppose que ce test a une efficacité de 99% (= probabilité que le test soit positif pour une personne atteinte du SIDA), et une probabilité de fausse alarme de 5% (= probabilité que le test soit positif pour une personne non atteinte). Enfin, on suppose la fréquence de séropositivité est 1/10000. Quelle est la probabilité pour qu'une personne obtenant un test positif soit atteinte du SIDA ?
7. Soit $A_n \to A$ et $B_n \to B$ (voir avant le théorème 2.4). Supposons aussi que $P(B) > 0$ et $P(B_n) > 0$ pour tout n. Montrer que
   a) $\lim_{n\to\infty} P(A_n \mid B) = P(A \mid B)$;
   b) $\lim_{n\to\infty} P(A \mid B_n) = P(A \mid B)$;
   c) $\lim_{n\to\infty} P(A_n \mid B_n) = P(A \mid B)$.
8. On modélise le jet d'une pièce avec deux résultats possibles, p = Pile et f = Face, chacun avec la probabilité 1/2. On jette deux fois cette pièce, de sorte que l'ensemble des résultats possibles Ω contient les quatre points : pp, pf, fp, ff. On suppose que les deux jets sont indépendants.

a) Quelle est la probabilité pour obtenir deux fois Face, sachant que le premier jet donne Face ? [Rép. : $\frac{1}{2}$].

b) Quelle est la probabilité pour obtenir deux fois Face, sachant que l'un des deux jets au moins donne Face ? [Rép. : $\frac{1}{3}$].

9. Si A, B, C sont indépendants et $\mathrm{P}(\mathrm{A} \cap \mathrm{B}) \neq 0$, montrer que $\mathrm{P}(\mathrm{C} \mid \mathrm{A} \cap \mathrm{B}) = \mathrm{P}(\mathrm{C})$.

10. Une urne contient r boules rouges et n boules noires. Une boule est choisie au hasard (chaque boule a la même probabilité d'être tirée), et une seconde boule est ensuite choisie au hasard parmi les $r + n - 1$ boules restantes. Trouver la probabilité pour que

a) Les deux boules soient rouges $\left[\text{Rép. : } \frac{r(r-1)}{(r+n)(r+n-1)}\right]$.

b) La première boule soit rouge et la seconde noire $\left[\text{Rép. : } \frac{rn}{(r+n)(r+n-1)}\right]$.

11. (Urne de Pólya.) Une urne contient r boules rouges et n boules noires. Une boule est choisie au hasard, on note sa couleur, et on la remet avec d boules supplémentaires de la même couleur. Puis on recommence la même procédure aussi souvent que nécessaire. Trouver la probabilité pour que

a) La seconde boule tirée soit noire $\left[\text{Rép. : } \frac{n}{n+r}\right]$.

b) La première boule est noire, sachant que la seconde est noire $\left[\text{Rép. : } \frac{n+d}{n+r+d}\right]$.

12. Dans la situation de l'exercice 11, on note B_m l'événement selon lequel la m-ième boule tirée est noire. Montrer que $\mathrm{P}(\mathrm{B}_m) = \mathrm{P}(\mathrm{B}_1)$.

13. Toujours dans la situation de l'exercice 11, trouver la probabilité pour que la première boule soit noire, sachant que les $m-1$ suivantes sont noires ; trouver la limite de cette probabilité lorsque $m \to \infty$ $\left[\text{Rép. : } \frac{n+md}{n+r+md}\text{; la limite est } 1\right]$.

14. Une compagnie d'assurance assure un nombre égal de conducteurs masculins et féminins. Tous les conducteurs (masculins) ont, chaque année, la probabilité α d'avoir un accident, et ceci indépendamment des autres années, et des autres conducteurs. Il en est de même des conductrices, avec un probabilité d'accident égale à β. La compagnie sélectionne un/une conducteur/trice au hasard.

a) Quelle est la probabilité pour que la personne sélectionnée ait un accident cette année ? $\left[\text{Rép. : } \frac{\alpha+\beta}{2}\right]$.

b) Quelle est la probabilité pour que la personne sélectionnée ait un accident deux années consécutives ? $\left[\text{Rép. : } \frac{\alpha^2+\beta^2}{2}\right]$.

15. Dans la situation de l'exercice 14, soit A_i l'événement « la personne sélectionnée a un accident l'année numéro i ». Montrer que $P(A_2 \mid A_1) \geqslant P(A_1)$. $\left[\text{Rép. : } P(A_2 \mid A_1) - P(A_1) = \frac{(\alpha-\beta)^2}{2(\alpha+\beta)}\right]$.

16. Toujours dans la même situation, trouver la probabilité pour que, une année donnée, une personne sélectionnée au hasard parmi celles ayant eu un accident soit une conductrice. $\left[\text{Rép. : } \frac{\beta}{\alpha+\beta}\right]$.

17. Si $A_1, A_2, \ldots, A_n$ sont des événements indépendants, montrer que la probabilité pour qu'aucun des A_i ne soit réalisé est au plus égale à $\exp\left(-\sum_{i=1}^{n} P(A_i)\right)$.

18. Si $P(A) > 0$, montrer que $P(A \cap B \mid A \cup B) \leqslant P(A \cap B \mid A)$.

Chapitre 4

Probabilités sur un espace fini ou dénombrable

Dans tout le chapitre 4 nous supposons que l'espace d'états Ω est fini ou dénombrable, et nous choisissons pour tribu la classe $\mathcal{A} = 2^\Omega$ de toutes les parties de Ω.

Théorème 4.1. (a) *Une probabilité sur l'ensemble fini ou dénombrable Ω est caractérisée par ses valeurs sur les singletons : $p_\omega = \mathrm{P}(\{\omega\})$, $\omega \in \Omega$.*

(b) *Si $(p_\omega)_{\omega\in\Omega}$ est une famille de réels indexée par l'ensemble fini ou dénombrable Ω, il existe une probabilité P (nécessairement unique par (a)) sur Ω telle que $\mathrm{P}(\{\omega\}) = p_\omega$ pour tout $\omega \in \Omega$ si et seulement si $p_\omega \geqslant 0$ et $\sum_{\omega\in\Omega} p_\omega = 1$.*

Lorsque Ω est infini dénombrable, $\sum_\omega p_\omega$ est la somme d'une infinité de termes, qui *a priori* ne sont pas ordonnés (bien qu'on puisse énumérer les points de Ω, cette énumération est en fait arbitraire) ; on n'a donc pas à proprement parler une série, mais une « famille sommable ». Nous donnons en appendice à ce chapitre un résumé des propriétés des familles sommables qui nous sont utiles dans ce chapitre et dans le suivant.

Preuve. Soit $\mathrm{A} \in \mathcal{A}$; on a $\mathrm{A} = \bigcup_{\omega\in\mathrm{A}}\{\omega\}$, réunion finie ou dénombrable de singletons deux à deux disjoints. Si P est une probabilité sur Ω, la σ-additivité implique

$$\mathrm{P}(\mathrm{A}) = \mathrm{P}\left(\bigcup_{\omega\in\mathrm{A}}\{\omega\}\right) = \sum_{\omega\in\mathrm{A}} \mathrm{P}(\{\omega\}) = \sum_{\omega\in\mathrm{A}} p_\omega.$$

On a donc (a).

Pour la condition nécessaire de (b), on remarque que si $\mathrm{P}(\{\omega\}) = p_\omega$, alors par définition $p_\omega \geqslant 0$, et aussi

$$1 = \mathrm{P}(\Omega) = \mathrm{P}\left(\bigcup_{\omega\in\Omega}\{\omega\}\right) = \sum_{\omega\in\Omega} \mathrm{P}(\omega) = \sum_{\omega\in\Omega} p_\omega.$$

Supposons inversement que les p_ω vérifient $p_\omega \geqslant 0$ et $\sum_\omega p_\omega = 1$. On pose $\mathrm{P}(\mathrm{A}) = \sum_{\omega\in\mathrm{A}} p_\omega$, avec la convention qu'une somme « vide » est nulle. On a donc $P(\Omega) = \sum_\omega p_\omega = 1$. Pour la σ-additivité, elle est évidente lorsque Ω est fini, et lorsque Ω est dénombrable elle provient de ce qu'on peut « sommer par paquets » pour obtenir que $\sum_{i\in I}\sum_{\omega\in A_i} p_\omega = \sum_{\omega\in\bigcup_{i\in I} A_i} p_\omega$ si les A_i sont deux à deux disjoints. ■

Supposons d'abord Ω fini. N'importe quelle famille de réels positifs indicée par Ω et de somme 1 constitue un exemple de probabilité sur Ω, mais parmi tous ces exemples le suivant est particulièrement important :

Définition 4.1. *On dit que la probabilité* P *sur l'espace fini* Ω *est* uniforme *si* $p_\omega = P(\{\omega\})$ *ne dépend pas de* ω.

Dans ce cas il est immédiat que

$$P(A) = \frac{\operatorname{card}(A)}{\operatorname{card}(\Omega)},$$

où card(A) désigne le « cardinal », ou nombre de points, de A. Dans ce cas le calcul des probabilités se ramène à des dénombrements : on est dans le cas de la *combinatoire*. Sur un espace fini donné Ω il existe une et une seule probabilité uniforme.

Nous allons maintenant donner deux exemples de probabilités très importantes pour les applications, puis examiner les rapports entre ces deux exemples.

a) La loi hypergéométrique. Une urne contient N boules blanches et M boules noires. On tire n boules (sans remettre les boules tirées dans l'urne, donc $n \leqslant N + M$). Parmi les boules tirées, il y en a X blanches et $n - X$ noires. On cherche la probabilité pour que $X = x$, où x est un entier (arbitraire) fixé.

Il s'agit d'une épreuve aléatoire, dans la mesure où on ne connaît pas *a priori* le résultat. Comme il s'agit d'un tirage sans remise, on peut supposer qu'on tire *simultanément* les n boules. Ainsi, il est naturel de considérer qu'un résultat est une partie à n éléments de l'ensemble $\{1, 2, \ldots, N + M\}$ des $N + M$ boules (qu'on peut supposer numérotées de 1 à $N + M$). Donc Ω est l'ensemble de toutes les parties à n éléments, et $\operatorname{card}(\Omega) = C^n_{N+M} = \frac{(N+M)!}{n!(N+M-n)!}$ (rappelons que la factorielle d'un entier p est $p\,! = 1.2\ldots(p-1).p$).

Ensuite, il est également naturel de considérer que tous les tirages possibles sont équiprobables, donc P est la probabilité uniforme sur Ω. La quantité X est une « variable aléatoire » car si on connaît le tirage ω, on connaît aussi le nombre $X(\omega)$ de boules blanches qu'il contient. L'ensemble $X^{-1}(\{x\})$, noté aussi $\{X = x\}$, contient $C^x_N C^{n-x}_M$ éléments si

$x \leqslant \mathrm{N}$ et $n - x \leqslant \mathrm{M}$, et est vide sinon. Donc

$$\mathrm{P}(\mathrm{X} = x) \;=\; \begin{cases} \dfrac{\mathrm{C}_{\mathrm{N}}^{x}\,\mathrm{C}_{\mathrm{M}}^{n-x}}{\mathrm{C}_{\mathrm{N+M}}^{n}} & \text{si } 0 \leqslant x \leqslant \mathrm{N} \text{ et } 0 \leqslant n - x \leqslant \mathrm{M} \\ 0 & \text{sinon.} \end{cases}$$

On a ainsi obtenu, lorsque x varie, la *loi* de X. Cette loi, appelée loi hypergéométrique, intervient naturellement dans la théorie des sondages : il y a $\mathrm{N} + \mathrm{M}$ électeurs, dont N ont l'opinion « blanche » et M l'opinion « noire », et on sonde avec un échantillon de taille n (cf. l'exercice 3 pour une généralisation à plus de deux « opinions »).

b) La loi binomiale. De la même urne que ci-dessus, on tire n boules, avec remise après chaque tirage (donc n peut être aussi grand qu'on veut). On cherche encore la probabilité $\mathrm{P}(\mathrm{X} = x)$, lorsque x est un entier entre 0 et n.

Ici, l'espace d'états naturel est le produit cartésien $\Omega = \{1, 2, \dots, \mathrm{N} + \mathrm{M}\}^n$, avec encore la probabilité uniforme. On a donc $\mathrm{card}(\Omega) = (\mathrm{N}+\mathrm{M})^n$, et un calcul simple montre que le nombre d'éléments $\mathrm{card}(\mathrm{X} = x)$ vaut $\mathrm{C}_n^x\ \mathrm{N}^x \mathrm{M}^{n-x}$. Donc

$$\mathrm{P}(\mathrm{X} = x) \;=\; \mathrm{C}_n^x \left(\frac{\mathrm{N}}{\mathrm{N} + \mathrm{M}}\right)^x \left(\frac{\mathrm{M}}{\mathrm{N} + \mathrm{M}}\right)^{n-x} \quad \text{pour} \quad x = 0, 1, \dots, n.$$

On écrit en général le résultat ainsi, en posant $p = \frac{\mathrm{N}}{\mathrm{N+M}}$:

$$\mathrm{P}(\mathrm{X} = x) \;=\; \mathrm{C}_n^x\, p^x (1-p)^{n-x} \quad \text{pour} \quad x = 0, 1, \dots, n.$$

Cette formule donne la *loi binomiale de taille n et de paramètre p*. A priori p est quelconque dans $[0,1]$ (dans l'exemple ci-dessus, p est bien sûr rationnel). On note $\mathrm{B}(p, n)$ cette loi.

c) La loi binomiale comme limite de lois hypergéométriques. Dans la situation a) ci-dessus, on suppose que n est fixé et que N et M tendent vers $+\infty$, de telle sorte que $\frac{\mathrm{N}}{\mathrm{N+M}}$ tende vers une limite p (nécessairement dans $[0,1]$). En développant les combinaisons dans (11), on voit facilement que

$$\mathrm{P}(\mathrm{X} = x) \;\to\; \mathrm{C}_n^x\, p^x (1-p)^{n-x} \qquad \text{pour} \quad x = 0, 1, \dots, n,$$

donc les lois hypergéométriques « convergent » vers la loi $\mathrm{B}(p, n)$ (en comparant à b) ci-dessus, on pourra vérifier que le résultat est intuitivement évident).

Exemples avec Ω dénombrable.

1. *La loi de Poisson* de paramètre $\lambda > 0$ est la probabilité P sur $\mathbb{N}$ définie par
$$p_n = e^{-\lambda}\frac{\lambda^n}{n!}, \qquad n = 0, 1, 2, 3, \ldots$$
2. *La loi géométrique* de paramètre $\alpha \in [0, 1[$ est la probabilité P sur $\mathbb{N}$ définie par
$$p_n = (1-\alpha)\alpha^n, \qquad n = 0, 1, 2, 3, \ldots$$

Noter que dans le modèle binomial, lorsque n est grand, le nombre $C_n^j p^j(1-p)^{n-j}$ peut se révéler impossible à calculer numériquement. Si p n'est proche ni de 0 ni de 1, il existe une approximation dite « normale » qui sera obtenue ultérieurement, quand on parlera du théorème-limite central. Lorsque p est « petit » il existe une autre approximation que nous décrivons ci-dessous.

Pour rendre compte du fait que p est « petit » et n est « grand », on suppose que p dépend de n et on le note p_n : on fait tendre n vers l'infini, en supposant que $\lim_{n\to\infty} np_n = \lambda$. On peut alors montrer (cf. exercice 1) que (pour j fixé)
$$\lim_{n\to\infty} C_n^j(p_n)^j(1-p_n)^{n-j} = e^{-\lambda}\frac{\lambda^j}{j!},$$
et ainsi on approche la loi binomiale par une loi de Poisson.

Appendice : Quelques résultats utiles sur les séries

Dans cet appendice nous faisons un résumé, essentiellement sans démonstrations, des résultats sur les séries et familles sommables qui sont d'usage constant dans l'étude des probabilités sur un espace dénombrable.

Auparavant, signalons que nous serons amenés très souvent à faire des opérations faisant intervenir $+\infty$ (qu'on écrit plus simplement ∞) ou $-\infty$. Pour que ces opérations aient un sens précis, on fera *toujours* les conventions suivantes :

$$+\infty+\infty = +\infty,\ -\infty-\infty = -\infty,\ a+\infty = +\infty,\ a-\infty = -\infty \text{ si } a \in \mathbb{R}, \tag{4.1}$$
$$0\times\infty = 0,\ a \in]0,\infty] \Rightarrow a\times\infty = +\infty,\ a \in [-\infty, 0[\Rightarrow a\times\infty = -\infty. \tag{4.2}$$

Soit u_n une suite numérique, et $S_n = u_1 + \cdots + u_n$ la « somme partielle » à l'ordre n.

S1 : La série $\sum_n u_n$ est dite *convergente* si S_n converge vers une limite *finie* S, notée aussi $S = \sum_n u_n$ (c'est la « somme » de la série).

S2 : La série $\sum_n u_n$ est dite *absolument convergente* si la série $\sum_n |u_n|$ converge.

S3 : Si $u_n \geq 0$ pour tout n, la suite S_n est croissante, donc elle tend toujours vers une limite $S \in [0, \infty]$. On écrit encore $S = \sum_n u_n$, bien que la série converge au sens de (S1) si et seulement si $S < \infty$. Avec les conventions (1.1) ceci s'applique même si les u_n sont à valeurs dans $[0, \infty]$.

En général la convergence d'une série dépend de l'ordre dans lequel on considère les termes. Il existe deux cas importants, signalés ci-dessous, où l'ordre des termes n'a pas d'importance ; au lieu de « série » on dit alors plutôt qu'on a une « famille sommable » :

S4 : Lorsque les u_n sont des réels de signe quelconque et lorsque la série est absolument convergente, on peut modifier de manière arbitraire l'ordre des termes sans changer la propriété d'être absolument convergente, ni la somme de la série.

S5 : Si $u_n \in [0, \infty]$ pour tout n, la somme $\sum_n u_n$ (finie ou infinie : cf. (S3) ci-dessus) ne change pas si on change l'ordre de sommation.

Dans ces deux situations, une propriété supplémentaire mérite d'être signalée :

S6 : Si $u_n \in [0, \infty]$, ou si la série est absolument convergente, on peut « sommer par paquets » : soit $(A_i)_{i \in I}$ une partition de $\mathbb{N}^*$, avec $I = \{1, 2, \ldots, N\}$ pour un entier N, ou $I = \mathbb{N}^*$. Pour chaque $i \in I$ on pose $v_i = \sum_{n \in A_i} u_n$: si A_i est fini, c'est une somme ordinaire ; sinon, v_i est elle-même la somme d'une série. On a alors $\sum_n u_n = \sum_{i \in I} v_i$ (cette dernière somme est de nouveau la somme d'une série si $I = \mathbb{N}^*$).

Exercices

1. (Approximation poissonnienne de la loi binomiale, suite.) On considère la loi binomiale $B(p, n)$, et on pose $\lambda = pn$. Montrer que

$$P(\{k\}) = \frac{\lambda^k}{k!}\left(1 - \frac{\lambda}{n}\right)^n \left\{\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-k+1}{n}\right)\right\}\left(1 - \frac{\lambda}{n}\right)^{-k}.$$

On fait tendre n vers l'infini, et on fait dépendre p de n, de sorte que λ reste constant. Montrer que pour k fixé on a

$$P(\{k\}) \to \frac{\lambda^k}{k!} e^{-\lambda}.$$

[Remarque : pour que cette approximation soit bonne il faut que n soit grand et p petit, de sorte que $\lambda = np$ soit de taille modérée, par exemple $\lambda \leqslant 20$.]

2. (Approximation poissonnienne de la loi binomiale, suite.) On se place dans la situation de l'exercice 1. Si $p_k = \mathrm{P}(\{k\})$, montrer que les $q_k = p_{n-k}$ sont les probabilités des singletons pour une loi binomiale $\mathrm{B}(1-p, n)$. En déduire une approximation de la loi binomiale par la loi de Poisson lorsque n est grand et p est proche de 1.

3. On se place dans le cadre de la « loi hypergéométrique », sauf qu'il y a m couleurs et N_i boules de couleur i. On note $\mathrm{N} = \mathrm{N}_1 + \cdots + \mathrm{N}_m$, et X_i le nombre de boules de couleur i parmi les n boules tirées. On a bien sûr $\mathrm{X}_1 + \cdots + \mathrm{X}_m = n$. Montrer que

$$\mathrm{P}(\mathrm{X}_1{=}x_1, \ldots, \mathrm{X}_m{=}x_m) = \begin{cases} \dfrac{\mathrm{C}_{\mathrm{N}_1}^{x_1} \ldots \mathrm{C}_{\mathrm{N}_m}^{x_m}}{\mathrm{C}_{\mathrm{N}}^{n}} & \text{si } x_1 + \cdots + x_m = n \\ 0 & \text{sinon.} \end{cases}$$

Chapitre 5

Variables aléatoires sur un espace fini ou dénombrable

Dans ce chapitre nous supposons encore l'espace d'états Ω fini ou dénombrable, et nous choisissons pour tribu la classe $\mathcal{A} = 2^{\Omega}$ de toutes les parties de Ω. Une *variable aléatoire* X est définie comme une application de Ω dans un ensemble E *a priori* arbitraire. Une variable aléatoire représente une quantité inconnue (d'où la terminologie « variable ») qui dépend de l'issue de l'expérience : ce n'est donc pas une « variable » au sens algébrique du terme, comme x dans la relation $x^2 - 9 = 0$. Avant de réaliser l'expérience, ce qui revient à choisir un point ω dans Ω, on connaît les valeurs que X peut éventuellement prendre, mais on ne sait pas encore quelle valeur elle va effectivement prendre lorsqu'on réalise l'expérience.

Noter que l'espace d'arrivée E n'est pas nécessairement dénombrable, c'est par exemple $\mathbb{R}$ si l'expérience consiste à choisir une personne dans une salle et si $X(\omega)$ représente la taille de la personne ω ; mais l'image E' de Ω par X (i.e., tous les points i de E pour lesquels il existe au moins un $\omega \in \Omega$ avec $X(\omega) = i$) est, quant à elle, nécessairement finie ou dénombrable.

On peut alors définir la *loi* de X, appelée aussi la *distribution* de X, par la formule

$$P_X(A) = P(\{\omega : X(\omega) \in A\}) = P(X^{-1}(A)) = P(X \in A).$$

Le fait que cette formule définisse une mesure de probabilité sur E' (muni de la tribu de toutes ses parties $2^{E'}$) est évident. Comme E' est au plus dénombrable, cette probabilité est complètement déterminée par les nombre suivants :

$$p_j^X = P(X = j) = \sum_{\{\omega : X(\omega) = j\}} p_\omega,$$

et la famille $(p_j^X : j \in E')$ est aussi appelée parfois la loi de la variable X. On a bien sûr $P_X(A) = \sum_{j \in A} p_j^X$. Si P_X est une loi munie d'un nom, par exemple une loi de Poisson, on dit aussi que X est une variable de Poisson.

Définition 5.1. *Soit X est une variable aléatoire à valeurs dans* $\mathrm{E} = \mathbb{R}$, *définie sur un espace fini ou dénombrable* Ω. *On appelle* espérance, *ou* espérance mathématique, *de X le nombre*

$$\mathrm{E}(\mathrm{X}) = \sum_{\omega} \mathrm{X}(\omega) p_{\omega},$$

lorsqu'il est bien défini : c'est le cas si Ω *est fini ; c'est aussi le cas, lorsque* Ω *est infini dénombrable, si la série ci-dessus est absolument convergente, ou si* $\mathrm{X}(\omega) \geqslant 0$ *pour tout* ω (dans ce dernier cas on peut avoir $\mathrm{E}(\mathrm{X}) = +\infty$).

La motivation de cette définition vient de l'approche par les fréquences : si on répète n fois l'expérience et si on note $\mathrm{X}_1, \ldots, \mathrm{X}_n$ les valeurs successives obtenues pour X, un calcul simple montre que la *moyenne empirique* $\frac{1}{n}(\mathrm{X}_1 + \cdots + \mathrm{X}_n)$ vaut $\sum_{\omega \in \Omega} \mathrm{X}(\omega) f_n(\{\omega\})$, et donc « converge » vers la quantité E(X) définie ci-dessus (au moins si Ω est fini).

Notons $\mathcal{L}^1$ l'espace de toutes les variables aléatoires réelles qui ont une espérance finie. Il est alors immédiat que :

(i) $\mathcal{L}^1$ est un espace vectoriel, et la fonctionnelle espérance E est linéaire.

(ii) La fonctionnelle espérance E est positive : si $\mathrm{X} \in \mathcal{L}^1$ et $\mathrm{X} \geqslant 0$, alors $\mathrm{E}(\mathrm{X}) \geqslant 0$. Plus généralement si $\mathrm{X}, \mathrm{Y} \in \mathcal{L}^1$ et $\mathrm{X} \leqslant \mathrm{Y}$ alors $\mathrm{E}(\mathrm{X}) \leqslant \mathrm{E}(\mathrm{Y})$.

(iii) $\mathcal{L}^1$ contient toutes les variables aléatoires bornées. Si $\mathrm{X} \equiv a$, alors $\mathrm{E}(\mathrm{X}) = a$.

(iv) Si $\mathrm{X} \in \mathcal{L}^1$, son espérance dépend seulement de sa loi, et si E′ est l'image de Ω par X, on a

$$\mathrm{E}(\mathrm{X}) = \sum_{j \in \mathrm{E}'} j \mathrm{P}(\mathrm{X} = j). \tag{5.1}$$

(Dans la formule de définition de E(X), appliquer la « sommation par paquets », propriété S6 du chapitre précédent, avec les « paquets » $\mathrm{A}_j = \{\omega : \mathrm{X}(\omega) = j\}$, en remarquant que $\sum_{\omega \in \mathrm{A}_j} p_\omega = \mathrm{P}(\mathrm{X} = j)$.)

(v) Si $\mathrm{X} = 1_{\mathrm{A}}$ est l'indicatrice d'un événement A, on a $\mathrm{E}(\mathrm{X}) = \mathrm{P}(\mathrm{A})$.

Théorème 5.1. *Soit $h : \mathbb{R} \to [0, \infty[$ une fonction positive et X une variable aléatoire réelle. On a alors*

$$P(\{\omega : h(X(\omega)) \geqslant a\}) \leqslant \frac{E(h(X))}{a}$$

pour tout $a > 0$.

Preuve. Comme X est une variable aléatoire, il en est de même de $Y = h(X)$; soit

$$A = Y^{-1}([a, \infty[) = \{\omega : h(X(\omega)) \geqslant a\} = \{h(X) \geqslant a\}.$$

On a $h(X) \geqslant a1_A$, donc

$$E(h(X)) \geqslant E(a1_A) = aE(1_A) = aP(A)$$

et le résultat suit. ■

Corollaire 5.1. (Inégalité de Markov.) *On a*

$$P(|X| \geqslant a) \leqslant \frac{E(|X|)}{a}.$$

Preuve. Prendre $h(x) = |x|$ dans le théorème 5.1. ■

Pour bien comprendre la définition suivante, on remarquera que si une X est une variable aléatoire réelle on a $|X| \leqslant 1 + X^2$, de sorte que si X^2 est dans $\mathcal{L}^1$ alors X est aussi dans $\mathcal{L}^1$ (cela sera un résultat général pour les variables réelles, voir le théorème 9.3).

Définition 5.2. *Soit X une variable aléatoire réelle telle que X^2 soit dans $\mathcal{L}^1$. On appelle* variance *de X le nombre*

$$\sigma^2 = \sigma_X^2 = E\big((X - E(X))^2\big).$$

L'écart-type *de X, soit σ_X, est la racine carrée positive de la variance.* La motivation essentielle pour introduire l'écart-type est le fait qu'il est comparable à la variable X elle-même, au sens de la « dimension » ; si X est par exemple une longueur, alors σ_X^2 représente une surface, tandis que σ_X représente à nouveau une longueur.

Si $\mu = E(X)$ représente l'espérance de X, ou *moyenne* comme on l'appelle souvent, alors $E(|X - E(X)|) = E(|X - \mu|)$ représente la moyenne de l'écart entre X et sa moyenne ; cette quantité mesure donc comment X diffère de sa moyenne. La variance est la moyenne du carré de l'écart entre

X et sa moyenne : par rapport à $E(|X-\mu|)$ cela diminue l'importance des « petits » écarts et augmente l'importance des grands. Cependant la variance est en général (beaucoup) plus facile à calculer que $E(|X-\mu|)$ (voir par exemple l'exercice 11.) Ainsi, la variance peut également être conçue comme une mesure de la variabilité de la variable aléatoire X.

Corollaire 5.2. (Inégalités de Bienaymé-Chebyshev.) *Si* X^2 *est dans* $\mathcal{L}^1$, *on a*

$$\text{(a) } P(|X| \geqslant a) \leqslant \frac{E(X^2)}{a^2} \qquad \textit{pour } a > 0,$$

$$\text{(b) } P(|X - E(X)| \geqslant a) \leqslant \frac{\sigma_X^2}{a^2} \qquad \textit{pour } a > 0.$$

Preuve. Pour (a), on applique le théorème 5.1 avec $h(x) = x^2$, ce qui donne

$$P(|X| \geqslant a) = P(h(X) \geqslant a^2) \leqslant \frac{E(X^2)}{a^2}.$$

Pour (b), on applique (a) à la variable $Y = |X - E(X)|$. ■

Exemples.

1) X est *de Poisson* de paramètre λ. Si $A \subset \mathbb{N}$, on a

$$P(X \in A) = \sum_{j \in A} P(X = j) = \sum_{j \in A} \frac{\lambda^j}{j!} e^{-\lambda}.$$

L'espérance de X est

$$\begin{aligned} E(X) &= \sum_{j=0}^{\infty} j P(X = j) = \sum_{j=0}^{\infty} j \frac{\lambda^j}{j!} e^{-\lambda} \\ &= \lambda \sum_{j=1}^{\infty} \frac{\lambda^{j-1}}{(j-1)!} e^{-\lambda} = \lambda e^{\lambda} e^{-\lambda} = \lambda. \end{aligned}$$

2) X est *de Bernoulli*, i.e. elle prend seulement les deux valeurs 1 et 0 avec les probabilités respectives p et $q = 1 - p$ (avec $p \in [0, 1]$). Alors

$$E(X) = 1.P(X = 1) + 0.P(X = 0) = 1.p + 0.q = p.$$

3) X est *binomiale* $B(p, n)$, i.e. X peut prendre les valeurs $0, 1, 2, \ldots, n$, et

$$P(X = k) = C_n^k p^k (1 - p)^{n-k}.$$

Supposons qu'on réalise n expériences successives pour une variable de Bernoulli (avec le même paramètre p), en notant Y_i le résultat de la i-ième expérience. La somme $X = Y_1 + \cdots + Y_n$ est précisément une variable binomiale $B(p,n)$ (cf. l'exemple 3 du chapitre 4). Donc

$$E(X) = E\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} E(Y_i) = \sum_{i=1}^{n} p = np.$$

Noter qu'on aurait aussi pu calculer $E(X)$ algébriquement en utilisant la définition :

$$E(X) = \sum_{i=0}^{n} iP(X = i) = \sum_{i=1}^{\infty} iC_n^i p^i (1-p)^{n-i},$$

mais ce calcul n'est pas immédiat.

4) Supposons encore qu'on réalise des expériences de Bernoulli avec le même paramètre p, mais au lieu de fixer le nombre n d'expériences *a priori*, on s'arrête dès que le nombre de succès (le nombre de fois où Y_i vaut 1) atteint un entier fixé $r \geqslant 1$.
Soit d'abord X le nombre d'échecs avant le premier succès (i.e. $r = 1$). Dans ce cas X suit une *loi géométrique* de paramètre $1 - p$:

$$P(X = k) = (1-p)^k p, \qquad k = 0, 1, 2, 3, \dots$$

et on a

$$E(X) = \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=0}^{\infty} kp(1-p)^k = p(1-p)\frac{1}{p^2} = \frac{1-p}{p}.$$

5) Dans la même situation que ci-dessus, X est maintenant le nombre d'échecs avant le r-ième succès, avec $r \geqslant 1$. Cette variable X suit une *loi de Pascal*, ou *loi binomiale négative*, de paramètres r et p :

$$P(X = j) = C_{j+r-1}^{r-1} p^r (1-p)^j, \qquad j = 0, 1, 2, \dots.$$

Noter que dans ce cas $X = \sum_{i=1}^{r} Z_i$, où les Z_i sont des variables géométriques de paramètre p. Par suite

$$E(X) = \sum_{i=1}^{r} E(Z_i) = \frac{r(1-p)}{p}.$$

6) Une loi fréquemment rencontrée en sciences sociales est la *loi de Pareto*, ou *loi Zêta*. Une variable aléatoire X suit cette loi si elle prend ses valeurs dans $\mathbb{N}^*$ et vérifie

$$P(X = j) = c\frac{1}{j^{\alpha+1}}, \qquad j = 1, 2, 3, \dots$$

pour un paramètre fixé $\alpha > 0$. La constante c est telle que $c\sum_{j=1}^{\infty} \frac{1}{j^{\alpha+1}} = 1$. La fonction

$$\zeta(s) = \sum_{k=1}^{\infty} \frac{1}{k^s}, \qquad s > 1,$$

est appelée *fonction Zêta de Riemann*, et elle est tabulée. On a donc $c = \frac{1}{\zeta(\alpha+1)}$, et on peut écrire

$$\mathrm{P}(\mathrm{X} = j) = \frac{1}{\zeta(\alpha+1)} \frac{1}{j^{\alpha+1}}.$$

On peut facilement calculer l'espérance :

$$\begin{aligned} \mathrm{E}(\mathrm{X}) &= \sum_{j=1}^{\infty} j\mathrm{P}(\mathrm{X} = j) = \frac{1}{\zeta(\alpha+1)} \sum_{j=1}^{\infty} \frac{j}{j^{\alpha+1}} \\ &= \frac{1}{\zeta(\alpha+1)} \sum_{j=1}^{\infty} \frac{1}{j^{\alpha}} = \frac{\zeta(\alpha)}{\zeta(\alpha+1)}. \end{aligned}$$

Noter que cette espérance est infinie si $0 < \alpha \leqslant 1$, et finie si $\alpha > 1$.

7) On dit que la variable X est *uniforme* sur $\{1, 2, \ldots, n\}$ si sa loi est la loi uniforme sur cet ensemble fini, c'est-à-dire si

$$\mathrm{P}(\mathrm{X} = j) = \frac{1}{n}, \qquad j = 1, 2, \ldots, n.$$

Comme $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$, on voit que

$$\mathrm{E}(\mathrm{X}) = \sum_{j=1}^{n} j\mathrm{P}(\mathrm{X} = j) = \sum_{j=1}^{n} j\frac{1}{n} = \frac{1}{n}\sum_{j=1}^{n} j = \frac{n(n+1)}{n \cdot 2} = \frac{n+1}{2}.$$

Exercices

1. Si $g : [0, \infty[\to [0, \infty[$ est strictement croissante, montrer que

$$\mathrm{P}(|\mathrm{X}| \geqslant a) \leqslant \frac{\mathrm{E}(g(|\mathrm{X}|))}{g(a)} \qquad \text{pour } a > 0.$$

2. Si $h : \mathbb{R} \to [0, \alpha]$ montrer que pour $0 \leqslant a < \alpha$ on a

$$P(h(X) \geqslant a) \geqslant \frac{E(h(X)) - a}{\alpha - a}.$$

3. Montrer que $\sigma_X^2 = E(X^2) - E(X)^2$ si les deux espérances existent et sont finies.
4. Montrer que $E(X)^2 \leqslant E(X^2)$ si les deux espérances existent.
5. Montrer que $\sigma_X^2 = E(X(X-1)) + \mu_X - \mu_X^2$, si $\mu_X = E(X)$ et si toutes les espérances existent et sont finies.
6. Soit X binomiale $B(p, n)$. Quelle valeur de j maximise l'expression $P(X = j)$? (*Indication :* Calculer $\frac{P(X=k)}{P(X=k-1)}$.) [Rép. : $[(n+1)p]$, où $[x]$ est la « partie entière » de x.]
7. Avec X est comme ci-dessus, calculer la probabilité pour que X soit pair. [Rép. : $\frac{1}{2}(1 + (1-2p)^n)$.]
8. Soit X_n binomiale $B(p_n, n)$, avec $\lambda = np_n$ constant. Soit $A_n = \{X_n \geqslant 1\}$, et soit Y une variable de Poisson de paramètre (λ). Montrer que $\lim_{n\to\infty} P(X_n = j \mid A_n) = P(Y = j \mid Y \geqslant 1)$.
9. Soit X une variable de Poisson de paramètre λ. Quelle valeur de j maximise $P(X = j)$? [Voir l'exercice 6. Rép. : $[\lambda]$.]
10. Avec X comme ci-dessus, pour $j > 0$ fixé, quelle valeur de λ maximise $P(X = j)$? [Rép. : j.]
11. Avec X comme ci-dessus, et si λ *est un entier*, montrer que $E(|X - \lambda|) = \dfrac{2\lambda^\lambda e^{-\lambda}}{(\lambda - 1)!}$ et que $\sigma_X^2 = \lambda$.

12.* Soit X binomiale $B(p, n)$. Montrer que pour tous $\lambda > 0$ et $\varepsilon > 0$ on a

$$P(X - np > n\varepsilon) \leqslant E\big(\exp(\lambda(X - np - n\varepsilon))\big).$$

13. Soit X_n binomiale $B(p, n)$. Montrer que pour tout $b > 0$ fixé on a $P(X_n \leqslant b) \to 0$ quand $n \to \infty$.
14. Soit X_n binomiale $B(p, n)$. Montrer que

$$P\left(\left|\frac{X_n}{n} - p\right| > a\right) \leqslant \frac{\sqrt{p(1-p)}}{a^2 n} \min\left\{\sqrt{p(1-p)}, a\sqrt{n}\right\},$$

et que $P(|X_n - np| \leqslant n\varepsilon) \to 1$ pour tout $\varepsilon > 0$ quand $n \to \infty$.

15.* Soit X_n binomiale $B(1/2, 2m)$ avec m entier, et soit

$$a(m,k) = \frac{4^m}{C_{2m}^m} P(X = m + k).$$

Montrer que $\lim_{m\to\infty}(a(m,k))^m = e^{-k^2}$.

16. Soit X géométrique. Montrer que si $i, j > 0$, on a $P(X > i + j \mid X > i) = P(X > j)$.

17. Soit X géométrique de paramètre p. Montrer que

$$E\left(\frac{1}{1+X}\right) = \log\big((1-p)^{\frac{p-1}{p}}\big).$$

18. On jette plusieurs fois de suite et indépendamment une pièce de monnaie, la probabilité de tomber sur « pile » étant p. Trouver :
    a) la probabilité de ne pas avoir de « face » au cours des n premiers jets ;
    b) la probabilité d'obtenir « face » pour la première fois au n-ième jet ;
    c) l'espérance du nombre de jets jusqu'à la première obtention de « face ». [Rép. : $\frac{1}{1-p}$.]

19. Étant donnée une suite quelconque d'événements $(A_n)_{n\geq 1}$, montrer que

$$E\left(\sum_{n=1}^{\infty} 1_{A_n}\right) = \sum_{n=1}^{\infty} P(A_n),$$

(∞ étant une valeur possible pour chacun des membres de cette équation).

20. Si X prend ses valeurs dans $\mathbb{N}$, montrer que

$$E(X) = \sum_{n=0}^{\infty} P(X > n).$$

21. Soit X une variable de Poisson de paramètre λ. Montrer que pour $r = 2, 3, 4, \ldots$ on a

$$E\big(X(X-1)\ldots(X-r+1)\big) = \lambda^r.$$

22. Soit X une variable géométrique de paramètre p. Montrer que pour $r = 1, 2, 3, \ldots$, on a

$$E\big(X(X-1)\ldots(X-r+1)\big) = \frac{r!\,p^r}{(1-p)^r}.$$

Chapitre 6

Construction d'une mesure de probabilité

Dans ce chapitre Ω est un espace totalement quelconque, muni d'une tribu $\mathcal{A}$; le couple $(\Omega, \mathcal{A})$ est appelé un *espace mesurable.* Nous nous proposons de construire, ou de caractériser, une mesure de probabilité sur $\mathcal{A}$ (on dit aussi, sur $(\Omega, \mathcal{A})$) à partir de la connaissance des probabilités $\mathrm{P}(\mathrm{A})$ pour les événements A appartenant à une certaine sous-classe de $\mathcal{A}$, qu'on aimerait être « aussi petite que possible ». Quand Ω est fini ou dénombrable nous avons vu que c'était une chose simple à réaliser, à partir par exemple de la probabilité $p_\omega = \mathrm{P}(\{\omega\})$ des singletons. Mais dans le cas général la même méthode ne marche plus : une probabilité « typique » P vérifiera $\mathrm{P}(\{\omega\}) = 0$ pour tout ω et donc les nombres $\mathrm{P}(\{\omega\})$ ne caractérisent pas P.

Pour construire P il convient d'abord de préciser les ingrédients de base. On verra dans des cas « concrets », et notamment au chapitre suivant, qu'il est souvent aisé de construire une « probabilité » sur une algèbre qui engendre la tribu $\mathcal{A}$, et le problème qu'on se pose ci-dessous est l'extension de cette probabilité à la tribu elle-même.

Soit donc une algèbre $\mathcal{A}_0$ qui engendre la tribu $\mathcal{A}$, et une probabilité P sur $\mathcal{A}_0$: par là, on entend une fonction $\mathrm{P} : \mathcal{A}_0 \to [0, 1]$ qui satisfait :

1. $\mathrm{P}(\Omega) = 1$,
2. (σ-additivité.) Pour toute suite (A_n) d'éléments de $\mathcal{A}_0$ deux à deux disjoints et telle que $\bigcup_n \mathrm{A}_n \in \mathcal{A}_0$, on a $\mathrm{P}(\bigcup_n \mathrm{A}_n) = \sum_n \mathrm{P}(\mathrm{A}_n)$.

Théorème 6.1. *Toute probabilité* P *définie sur une algèbre* $\mathcal{A}_0$ *admet une unique extension (notée aussi* P*) à la tribu* $\mathcal{A} = \sigma(\mathcal{A}_0)$.

Il importe de souligner qu'on ne peut en général pas étendre la probabilité P à la tribu 2^Ω de toutes les parties de Ω : il n'existe pas de probabilité sur cette tribu qui coïncide avec P sur $\mathcal{A}_0$.

Nous allons démontrer ci-dessous l'unicité ; quant à l'existence, assez difficile, nous renvoyons à l'un des nombreux livres de théorie de la mesure, par exemple [19], [4] ou [26]. Nous devons d'abord établir un résultat très utile dans de nombreux contextes.

Définition 6.1. *Une classe $\mathcal{C}$ de parties de Ω est* stable par intersection *si pour tous $A, B \in \mathcal{C}$ on a $A \cap B \in \mathcal{C}$ (et alors, si $A_1, \dots, A_n \in \mathcal{C}$, on a $A_1 \cap A_2 \cap \dots \cap A_n \in \mathcal{C}$ pour n arbitraire* fini*).*

Une classe $\mathcal{C}$ est stable par limite croissante *si pour toute suite croissante $A_1 \subset A_2 \subset A_3 \subset \cdots \subset A_n \subset \cdots$ d'éléments de $\mathcal{C}$ on a $\bigcup_{n=1}^{\infty} A_n \in \mathcal{C}$.*

Une classe $\mathcal{C}$ est stable par différence *si pour tous $A, B \in \mathcal{C}$ avec $A \subset B$ on a*[1] $B \setminus A \in \mathcal{C}$.

Théorème 6.2. (Théorème des classes monotones.) *Soit $\mathcal{C}$ une classe de parties de Ω, stable par intersection et contenant Ω. Soit $\mathcal{D}$ la plus petite classe contenant $\mathcal{C}$ et stable par limite croissante et par différence. Alors $\mathcal{D} = \sigma(\mathcal{C})$.*

Preuve. Remarquons d'abord que l'intersection d'une famille quelconque de classes qui sont stables par limite croissante et par différence est une classe également stable par limite croissante et par différence. Ainsi, en prenant l'intersection de toutes les classes ayant ces propriétés et contenant $\mathcal{C}$ on voit qu'il existe une *plus petite* classe contenant $\mathcal{C}$ et stable par limite croissante et par différence.

Pour tout $B \subset \Omega$ notons $\mathcal{D}_B$ la classe de toutes les parties A vérifiant $A \in \mathcal{D}$ et $A \cap B \in \mathcal{D}$. Étant données les propriétés de $\mathcal{D}$, on voit facilement que $\mathcal{D}_B$ est stable par limite croissante et par différence.

Soit $B \in \mathcal{C}$; pour chaque $C \in \mathcal{C}$ on a $B \cap C \in \mathcal{C} \subset \mathcal{D}$ et $C \in \mathcal{D}$, donc $C \in \mathcal{D}_B$. Par suite $\mathcal{C} \subset \mathcal{D}_B \subset \mathcal{D}$. Donc $\mathcal{D} = \mathcal{D}_B$, grâce aux propriétés de $\mathcal{D}$ et de $\mathcal{D}_B$.

Maintenant, soit B dans $\mathcal{D}$. Pour chaque $C \in \mathcal{C}$ on a $B \in \mathcal{D}_C$, et à cause de ce qui précède il vient $B \cap C \in \mathcal{D}$, donc $C \in \mathcal{D}_B$, donc $\mathcal{C} \subset \mathcal{D}_B \subset \mathcal{D}$, donc $\mathcal{D} = \mathcal{D}_B$.

Comme $\mathcal{D} = \mathcal{D}_B$ pour tout $B \in \mathcal{D}$, on conclut que $\mathcal{D}$ est stable par intersection. De plus $\Omega \in \mathcal{D}$, et $\mathcal{D}$ est stable par différence, donc aussi par complémentation. Comme $\mathcal{D}$ est aussi stable par limite croissante, il s'ensuit que $\mathcal{D}$ est une tribu, et c'est clairement la plus petite tribu contenant $\mathcal{C}$. ∎

La preuve de l'unicité dans le théorème 6.1 est une conséquence immédiate du corollaire 6.1 ci-dessous, qui est lui-même une conséquence du théorème 6.2.

Corollaire 6.1. *Soit P et Q deux probabilités définies sur $\mathcal{A}$. Si P et Q coïncident sur une classe $\mathcal{C} \subset \mathcal{A}$ qui est stable par intersection, et si $\sigma(\mathcal{C}) = \mathcal{A}$, alors $P = Q$.*

1. $B \setminus A$ désigne $B \cap A^c$.

Preuve. $\Omega \in \mathcal{A}$ parce que $\mathcal{A}$ est une tribu, et comme $\mathrm{P}(\Omega) = \mathrm{Q}(\Omega) = 1$ on peut supposer sans perte de généralité que $\Omega \in \mathcal{C}$. Soit $\mathcal{B} = \{\mathrm{A} \in \mathcal{A} : \mathrm{P}(\mathrm{A}) = \mathrm{Q}(\mathrm{A})\}$. Par définition d'une mesure de probabilité et par le théorème 2.3, $\mathcal{B}$ est stable par différence et par limite croissante. De plus $\mathcal{B}$ contient $\mathcal{C}$ par hypothèse. Par suite, comme $\sigma(\mathcal{C}) = \mathcal{A}$, on a $\mathcal{B} = \mathcal{A}$ par application du théorème des classes monotones 6.2. ■

Il existe une version du théorème 6.2 pour les fonctions. Nous ne l'utiliserons pas dans ce livre, mais c'est un théorème qui est utile pour de nombreuses applications et nous allons l'énoncer sans démonstration, et en anticipant sur le chapitre 8 pour la notion de fonction mesurable : le lecteur pourra trouver une preuve dans [18] ou [22]. Si $\mathcal{M}$ une classe de fonctions de Ω dans $\mathbb{R}$, on note $\sigma(\mathcal{M})$ la plus petite des tribus de Ω qui rendent toutes les fonctions de $\mathcal{M}$ mesurables.

Théorème 6.3. (Théorème des classes monotones.) *Soit $\mathcal{M}$ une classe de fonctions bornées de Ω dans $\mathbb{R}$. Supposons que $\mathcal{M}$ soit stable par multiplication : $f, g \in \mathcal{M}$ impliquent $fg \in \mathcal{M}$. Soit $\mathcal{A} = \sigma(\mathcal{M})$ (la plus petite tribu rendant toutes les fonctions de $\mathcal{M}$ mesurables). Soit enfin $\mathcal{H}$ un espace vectoriel de fonctions, contenant $\mathcal{M}$. Si $\mathcal{H}$ contient les fonctions constantes et est stable par limite croissante (i.e. si les $(f_n)_{n\geqslant 1}$ sont des fonctions de $\mathcal{H}$ telles que $0 \leqslant f_1 \leqslant f_2 \leqslant f_3 \leqslant \cdots$, et si $f = \lim_{n\to\infty} f_n$ est bornée, alors $f \in \mathcal{H}$). Alors $\mathcal{H}$ contient toutes les fonctions bornées $\mathcal{A}$-mesurables.*

Si en général on ne peut pas étendre la probabilité P sur la tribu $\mathcal{A}$ à la tribu de toutes les parties de Ω, on peut néanmoins l'étendre à une tribu un peu plus grande que $\mathcal{A}$. C'est l'objet de la fin de ce chapitre.

Définition 6.2. *Soit* P *une probabilité sur $\mathcal{A}$. Un ensemble* négligeable *pour* P *est une partie* A *de* Ω *telle qu'il existe un élément* B *de $\mathcal{A}$ vérifiant* $\mathrm{A} \subset \mathrm{B}$ *et* $\mathrm{P}(\mathrm{B}) = 0$.

On dit qu'une propriété est vraie *presque sûrement* (en abrégé : p.s.) si elle est vraie en dehors d'un ensemble négligeable ; cette notion dépend bien sûr de la probabilité, de sorte qu'on dit parfois P-presque sûrement, ou P-p.s.

Les ensembles négligeables ne sont pas nécessairement dans $\mathcal{A}$. Néanmoins il est naturel de leur affecter une probabilité nulle, et on peut même étendre la probabilité à la tribu engendrée par $\mathcal{A}$ et les ensembles P-négligeables. C'est ce qu'on fait dans le théorème suivant, qui ne sera pas utilisé dans la suite.

Théorème 6.4. *Soit* P *une probabilité sur $\mathcal{A}$ et $\mathcal{N}$ la classe de tous les ensembles* P*-négligeables. Alors* $\mathcal{A}' = \{\mathrm{A} \cup \mathrm{N} : \mathrm{A} \in \mathcal{A}, \mathrm{N} \in \mathcal{N}\}$ *est une* *tribu, appelée* P-complétion *de* $\mathcal{A}$. *C'est la plus petite tribu contenant* $\mathcal{A}$ *et* $\mathcal{N}$, *et* P *s'étend de manière unique en une probabilité (notée aussi* P*) sur* $\mathcal{A}'$, *en posant* $\mathrm{P}(\mathrm{A} \cup \mathrm{N}) = \mathrm{P}(\mathrm{A})$ *pour* $\mathrm{A} \in \mathcal{A}$ *et* $\mathrm{N} \in \mathcal{N}$.

Preuve. L'unicité de l'extension, si elle existe, est triviale ; de même, comme $\emptyset$ appartient à $\mathcal{A}$ et à $\mathcal{N}$, le fait que $\mathcal{A}'$, si c'est une tribu, soit la plus petite tribu contenant $\mathcal{A}$ et $\mathcal{N}$ est trivial. Il nous suffit donc de montrer que $\mathcal{A}'$ est une tribu et que si on pose $\mathrm{Q}(\mathrm{B}) = \mathrm{P}(\mathrm{A})$ pour $\mathrm{B} = \mathrm{A} \cup \mathrm{N}$ (avec $\mathrm{A} \in \mathcal{A}$ et $\mathrm{N} \in \mathcal{N}$), alors $\mathrm{Q}(\mathrm{B})$ ne dépend pas de la décomposition $\mathrm{B} = \mathrm{A} \cup \mathrm{N}$ et Q est une probabilité sur $\mathcal{A}'$.

D'abord, $\Omega \in \mathcal{A} \subset \mathcal{A}'$ et on a déjà vu que $\emptyset \in \mathcal{A}'$. Ensuite comme $\mathcal{A}$ et $\mathcal{N}$ sont stables par les unions dénombrables (pour vérifier ceci pour $\mathcal{N}$, utiliser l'exercice 16 du chapitre 2), $\mathcal{A}'$ l'est également. Soit $\mathrm{B} = \mathrm{A} \cup \mathrm{N} \in \mathcal{A}'$, avec $\mathrm{A} \in \mathcal{A}$ et $\mathrm{N} \in \mathcal{N}$. Il existe $\mathrm{C} \in \mathcal{A}$ avec $\mathrm{P}(\mathrm{C}) = 0$ et $\mathrm{N} \subset \mathrm{C}$. Posons $\mathrm{A}' = \mathrm{A}^c \cap \mathrm{C}^c$ (qui est dans $\mathcal{A}$) et $\mathrm{N}' = \mathrm{N}^c \cap \mathrm{A}^c \cap \mathrm{C}$ (qui est contenu dans C et qui donc appartient à $\mathcal{N}$) ; comme $\mathrm{B}^c = \mathrm{A}' \cup \mathrm{N}'$, on a $\mathrm{B}^c \in \mathcal{A}'$, et $\mathcal{A}'$ est stable par complémentation : ainsi, $\mathcal{A}'$ est une tribu.

Supposons maintenant que $\mathrm{A}_1 \cup \mathrm{N}_1 = \mathrm{A}_2 \cup \mathrm{N}_2$ avec $\mathrm{A}_i \in \mathcal{A}$ et $\mathrm{N}_i \in \mathcal{N}$. La différence symétrique $\mathrm{A}_1 \Delta \mathrm{A}_2 = (\mathrm{A}_1 \cap \mathrm{A}_2^c) \cup (\mathrm{A}_1^c \cap \mathrm{A}_2)$ est contenue dans $\mathrm{N}_1 \cup \mathrm{N}_2$, lui-même contenu dans un élément C de $\mathcal{A}$ de probabilité nulle : par suite $\mathrm{P}(\mathrm{A}_1) = \mathrm{P}(\mathrm{A}_2)$, ce qui montre que Q est défini sans ambiguïté, et coïncide de manière évidente avec P sur $\mathcal{A}$. Enfin le fait que Q soit une probabilité sur $\mathcal{A}'$ est à peu près évident. ■

Chapitre 7

Probabilités sur $\mathbb{R}$ et fonctions de répartition

Dans ce chapitre nous traitons un cas particulier – concret – du problème exposé au chapitre précédent, à savoir la caractérisation d'une probabilité P à partir des P(A) pour une classe aussi petite que possible d'événements A. On suppose que $\Omega = \mathbb{R}$, et on munit cet espace de la tribu borélienne $\mathcal{B}$ (c'est-à-dire $\mathcal{B} = \sigma(\mathcal{O})$, où $\mathcal{O}$ désigne la classe de tous les ouverts de $\mathbb{R}$).

Définition 7.1. *Si* P *est une probabilité sur* $(\mathbb{R}, \mathcal{B})$, *sa* fonction de répartition *est la fonction*

$$\mathrm{F}(x) = \mathrm{P}(]-\infty, x]). \tag{7.1}$$

Théorème 7.1. *La fonction de répartition* F *caractérise la probabilité.*

Preuve. Nous voulons montrer que si Q est une autre probabilité sur $(\mathbb{R}, \mathcal{B})$, de fonction de répartition

$$\mathrm{G}(x) = \mathrm{Q}(]-\infty, x])$$

et si F = G, alors P = Q.

Soit d'abord $\mathcal{B}_0$ la classe de toutes les unions finies disjointes d'intervalles de la forme $]x, y]$, avec $-\infty \leqslant x \leqslant y \leqslant +\infty$ (avec la convention que $]x, \infty] =]x, \infty[$; noter aussi que $]x, y] = \emptyset$ si $x = y$). Il est facile de vérifier que $\mathcal{B}_0$ est une algèbre. De plus tout intervalle ouvert $]a, b[$ s'écrit $]a, b[= \bigcup_{n=\mathrm{N}}^{\infty}]a, b - \frac{1}{n}]$ pour un N assez grand, de sorte que $\sigma(\mathcal{B}_0)$ contient tous les intervalles ouverts. Comme tout ouvert de $\mathbb{R}$ est une réunion dénombrable d'intervalles ouverts et comme $\mathcal{B}$ est la tribu engendrée par les ouverts, on en déduit que $\sigma(\mathcal{B}_0) \supset \mathcal{B}$. Enfin $\bigcap_{n=1}^{\infty}]a, b + \frac{1}{n}[=]a, b]$, donc $\mathcal{B}_0 \subset \mathcal{B}$ et par suite $\mathcal{B} = \sigma(\mathcal{B}_0)$.

La relation (7.1) implique

$$\mathrm{P}(]x, y]) = \mathrm{F}(y) - \mathrm{F}(x),$$

et si $\mathrm{A} \in \mathcal{B}_0$ s'écrit

$$\mathrm{A} = \bigcup_{1 \leqslant i \leqslant n}]x_i, y_i], \quad \text{avec} \quad y_i < x_{i+1},$$

alors $P(A) = \sum_{1\leqslant i\leqslant n}\{F(y_i) - F(x_i)\}$. Comme $G = F$, $Q(A)$ est donné par la même expression : par suite $Q(A) = P(A)$ pour toute $A \in \mathcal{B}_0$, et le théorème 6.1 implique que $P = Q$ sur $\mathcal{B}$, i.e. P et Q sont la même probabilité. ■

La signification du théorème 7.1 est que la probabilité est entièrement connue si on connaît seulement F : on peut donc en principe retrouver $P(A)$ pour n'importe quel borélien A à partir de la fonction F (déterminer « explicitement » $P(A)$ en fonction de F est une tout autre chose !)

Il est donc important de savoir caractériser les fonctions F qui sont des fonctions de répartition : cela permet en principe de caractériser *toutes* les probabilités sur $\mathbb{R}$, et cela fait l'objet du théorème suivant. (Rappelons qu'une fonction F est *continue à droite* si $\lim_{y\downarrow x} F(y) = F(x)$, pour tout $x \in \mathbb{R}$.)

Théorème 7.2. *Une fonction F est la fonction de répartition d'une (unique) probabilité sur $(\mathbb{R}, \mathcal{B})$ si et seulement si elle vérifie les trois propriétés suivantes :*

(i) F *est croissante (au sens large) ;*

(ii) F *est continue à droite ;*

(iii) $\lim_{x\to-\infty} F(x) = 0$ *et* $\lim_{x\to+\infty} F(x) = 1$.

Preuve. Supposons que F soit la fonction de répartition de la probabilité P. Si $y > x$ on a $]-\infty, x] \subset]-\infty, y]$, donc $F(x) = P(]-\infty, x]) \leqslant P(]-\infty, y]) = F(y)$ et on a (i). Ensuite, soit x_n une suite de réels décroissant vers x. On a $\bigcap_{n=1}^{\infty}]-\infty, x_n] =]-\infty, x]$, et la suite d'intervalles $\{]-\infty, x_n], n \geqslant 1\}$ est décroissante. Donc $P(]-\infty, x]) = P(\bigcap_{n=1}^{\infty}]-\infty, x_n]) = \lim_{n\to\infty} P(]-\infty, x_n])$ d'après le théorème 2.3, et on a (ii). De la même manière le théorème 2.3 nous donne (iii).

Inversement, supposons qu'on ait (i), (ii) et (iii). En accord avec (iii), on pose $F(-\infty) = 0$ et $F(+\infty) = 1$. Introduisons la même algèbre $\mathcal{B}_0$ que dans la preuve du théorème 7.1. Définissons ensuite la fonction $P : \mathcal{B}_0 \to [0, 1]$ comme suit : $P(\emptyset) = 0$ et, pour $A = \bigcup_{1\leqslant i\leqslant n}]x_i, y_i]$ avec $-\infty \leqslant x_1 < \cdots < y_i < x_{i+1} < y_{i+1} < \cdots < y_n \leqslant \infty$,

$$P(A) = \sum_{1\leqslant i\leqslant n} \{F(y_i) - F(x_i)\}.$$

(Noter que la représentation de chaque $A \in \mathcal{B}_0$ ci-dessus est unique). La condition (iii) entraîne $P(\mathbb{R}) = 1$. Il nous reste à montrer que P est une probabilité sur $\mathcal{B}_0$: dans ce cas, le théorème 6.1 impliquera que P admet une unique extension à la tribu $\mathcal{B}$, dont la fonction de répartition est trivialement F, et le résultat sera démontré.

Il nous reste donc à montrer la σ-additivité de P sur $\mathcal{B}_0$: pour toute suite $(\mathrm{A}_n) \in \mathcal{B}_0$ d'ensembles deux à deux disjoints, tels que $\bigcup_{n=1}^{\infty} \mathrm{A}_n \in \mathcal{B}_0$, alors $\mathrm{P}(\bigcup_{n=1}^{\infty} \mathrm{A}_n) = \sum_{n=1}^{\infty} \mathrm{P}(\mathrm{A}_n)$. Mais l'additivité de P sur $\mathcal{B}_0$ (i.e. $\mathrm{P}(\mathrm{A} \cup \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B})$ si $\mathrm{A} \cap \mathrm{B} = \emptyset$ et $\mathrm{A}, \mathrm{B} \in \mathcal{B}_0$) est une conséquence immédiate de la définition de P, et le théorème 2.3 donne plusieurs conditions équivalentes pour qu'une mesure « additive » sur une tribu soit σ-additive. Il est immédiat de vérifier que ces conditions restent valables si on remplace la tribu par une algèbre : par suite, il nous suffit de montrer que la condition (ii) est vérifiée, autrement dit que si $\mathrm{A}_n \in \mathcal{B}_0$ avec A_n décroissant vers $\emptyset$, alors $\mathrm{P}(\mathrm{A}_n) \to 0$.

L'idée de la preuve consiste, pour chaque $\varepsilon > 0$ fixé, à construire une suite décroissante d'ensembles B_n de $\mathcal{B}_0$ tels que $\overline{\mathrm{B}}_n \subset \mathrm{A}_n$ et $\mathrm{P}(\mathrm{A}_n) - \mathrm{P}(\mathrm{B}_n) \leqslant \varepsilon$, et que les $\overline{\mathrm{B}}_n$ soient compacts : comme $\bigcap \mathrm{A}_n = \emptyset$, on a $\bigcap \overline{\mathrm{B}}_n = \emptyset$ et donc $\overline{\mathrm{B}}_n = \emptyset$ pour n assez grand.

Soit donc $\mathrm{A}_n \in \mathcal{B}_0$ comme ci-dessus. Chaque A_n s'écrit

$$\mathrm{A}_n = \bigcup_{1 \leqslant i \leqslant k_n}]x_i^n, y_i^n],$$

avec $y_i^n < x_{i+1}^n$. Soit $\varepsilon > 0$. Par (iii) il existe un nombre z tel que $\mathrm{F}(-z) \leqslant \varepsilon$ et $1 - \mathrm{F}(z) \leqslant \varepsilon$. Pour chaque n et i il existe $a_i^n \in]x_i^n, y_i^n]$ tel que $\mathrm{F}(a_i^n) - \mathrm{F}(x_i^n) \leqslant \frac{\varepsilon}{2^{i+n}}$, grâce à (ii). Posons alors

$$\mathrm{B}'_n = \bigcup_{1 \leqslant i \leqslant k_n} \{]a_i^n, y_i^n] \cap]-z, z]\}, \qquad \mathrm{B}_n = \bigcap_{m \leqslant n} \mathrm{B}'_m.$$

On remarque que $\mathrm{B}'_n \in \mathcal{B}_0$ et $\mathrm{B}'_n \subset \mathrm{A}_n$, donc $\mathrm{B}_n \in \mathcal{B}_0$ et $\mathrm{B}_n \subset \mathrm{A}_n$ également. De plus $\mathrm{A}_n \backslash \mathrm{B}_n \subset \bigcup_{m \leqslant n} (\mathrm{A}_m \backslash \mathrm{B}'_m)$, par suite

$$\begin{aligned}
\mathrm{P}(\mathrm{A}_n) - \mathrm{P}(\mathrm{B}_n) &\leqslant \mathrm{P}(]-z, z]^c) + \sum_{m=1}^{n} \mathrm{P}((\mathrm{A}_m \backslash \mathrm{B}'_m) \cap]-z, z]) \\
&\leqslant \mathrm{P}(]-z, z]^c) + \sum_{m=1}^{n} \sum_{i=1}^{k_m} \mathrm{P}(]x_i^m, a_i^m]) \\
&\leqslant \mathrm{F}(-z) + 1 - \mathrm{F}(z) + \sum_{m=1}^{n} \sum_{i=1}^{k_m} \{\mathrm{F}(a_i^m) - \mathrm{F}(x_i^m)\} \\
&\leqslant 3\varepsilon.
\end{aligned} \tag{7.2}$$

De plus $\overline{\mathrm{B}}_n \subset \mathrm{A}_n$ (où $\overline{\mathrm{B}}_n$ est la fermeture de B_n), donc la suite des $\overline{\mathrm{B}}_n$, qui est décroissante, est d'intersection vide par hypothèse. Enfin

$\overline{B}_n \subset [-z, z]$, donc chaque $\overline{B}_n$ est un ensemble compact. Une propriété des compacts, connue sous le nom de « propriété d'intersection finie »[1], est que si l'intersection d'une infinité de compacts est vide, il existe une sous-famille finie parmi ces compacts qui est déjà d'intersection vide : la suite de compacts $\overline{B}_n$ décroissant vers $\emptyset$, il s'ensuit qu'il existe un m tel que $\overline{B}_m = \emptyset$, donc *a fortiori* $B_n = \emptyset$ et $P(B_n) = 0$, pour tout $n \geqslant m$. Donc

$$P(A_n) = P(A_n) - P(B_n) \leqslant 3\varepsilon$$

par (7.2), si $n \geqslant m$. Comme ε est arbitraire, on en déduit que $P(A_n) \downarrow 0$.

(Noter que cette démonstration – plutôt longue – deviendrait presque triviale si la suite k_n ci-dessus était bornée ; mais, bien que les A_n décroissent vers $\emptyset$, la suite k_n tend typiquement vers l'infini !). ■

Proposition 7.1. *Soit* F *la fonction de répartition de la probabilité* P *sur* $\mathbb{R}$. *En notant* $F(x-)$ *la limite à gauche de* F *au point* x (qui existe puisque F est croissante), *pour tous* $x < y$ *on a*

(i) $P(]x, y]) = F(y) - F(x)$,

(ii) $P([x, y]) = F(y) - F(x-)$,

(iii) $P([x, y[) = F(y-) - F(x-)$,

(iv) $P(]x, y[) = F(y-) - F(x)$,

(v) $P(\{x\}) = F(x) - F(x-)$,

et en particulier $P(\{x\}) = 0$ *pour tout* x *si et seulement si la fonction* F *est continue.*

Preuve. (i) a déjà été montré. Pour (ii) on écrit que

$$P\big(\big]x - \frac{1}{n}, y\big]\big) = F(y) - F\big(x - \frac{1}{n}\big)$$

par (i). Le membre de droite converge quand $n \to \infty$ vers $F(y) - F(x-)$ par définition de la limite à gauche de F ; quant au membre de gauche, il converge vers $P([x, y])$ par application du théorème 2.3 et par le fait que la suite d'intervalles $]x - \frac{1}{n}, y]$ décroît vers $[x, y]$. Les assertions (iii), (iv) et (v) se montrent de manière analogue. ■

Exemples.

Considérons d'abord deux exemples généraux :

1. Voir par exemple [15, p. 81] ou [8].

1. Si f est positive et Riemann-intégrable et $\int_{-\infty}^{\infty} f(x)dx = 1$, la fonction $\mathrm{F}(x) = \int_{-\infty}^{x} f(y)dy$ est la fonction de répartition d'une probabilité P sur $\mathbb{R}$. La fonction f est appelée sa *densité.* (Il n'est *pas vrai* que toute probabilité sur $\mathbb{R}$ admet une densité : en effet s'il existe une densité la fonction F est continue, alors qu'il existe beaucoup de fonctions de répartition discontinues, ainsi que le montre l'exemple suivant.)
2. Soit $\alpha \in \mathbb{R}$. La « masse de Dirac » au point α est la probabilité sur $\mathbb{R}$ qui est donné par

$$\mathrm{P}(\mathrm{A}) = \begin{cases} 1 & \text{si } \alpha \in \mathrm{A}, \\ 0 & \text{sinon.} \end{cases}$$

Sa fonction de répartition est

$$\mathrm{F}(x) = \begin{cases} 0 & \text{si } x < \alpha, \\ 1 & \text{si } x \geqslant \alpha. \end{cases}$$

Dans les exemples 3 à 10 ci-dessous nous définissons la probabilité (ou « loi ») P sur $\mathbb{R}$ par sa densité f ; c'est-à-dire que nous définissons la fonction positive f d'intégrale 1, et la fonction de répartition F de P est $\mathrm{F}(x) = \int_{-\infty}^{x} f(u)du$. Dans chaque cas le lecteur vérifiera aisément que f est positive et d'intégrale 1. On commet un léger abus de langage (habituel) en identifiant f avec la loi P correspondante, puisqu'elle détermine en fait de manière unique cette loi.

3. $f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } a \leqslant x \leqslant b, \\ 0 & \text{sinon,} \end{cases}$ est appelée la *loi uniforme sur* $[a, b]$. Cette loi correspond à l'idée que « chaque point de $[a, b]$ est également vraisemblable », et elle est l'analogue « continu » de la loi uniforme sur un ensemble d'entiers $\{p, p+1, \ldots, q\}$; elle correspond à une densité constante sur l'intervalle considéré $[a, b]$.
4. $f(x) = \begin{cases} \beta e^{-\beta x} & \text{si } x \geqslant 0, \\ 0 & \text{si } x < 0, \end{cases}$ est appelée la *loi exponentielle* de paramètre $\beta > 0$. Cette loi est souvent utilisée pour modéliser la durée de vie d'un objet qui ne « vieillit » pas : sachant que cette durée de vie dépasse s, la probabilité conditionnelle pour qu'elle dépasse $s + t$ est la même que la probabilité *a priori* pour qu'elle dépasse t. C'est par exemple ce qui se passe (semble-t-il !) pour la

durée de vie des lampes au néon : si on croit en la validité d'un tel modèle, il est stupide de remplacer un tube au néon en état de marche par un tube neuf. Cette propriété de « non vieillissement » caractérise en fait la loi exponentielle : voir les exercices 21 et 22 du chapitre 9.

5. $f(x) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{si } x \geqslant 0, \\ 0 & \text{si } x < 0, \end{cases}$ est appelée la *loi gamma* de paramètres α et β (avec $0 < \alpha < \infty$ et $0 < \beta < \infty$; Γ désigne la fonction gamma[2]. Cette loi intervient dans diverses applications. Un exemple en est la fiabilité : si l'élément essentiel d'une machine a une durée de vie exponentielle de paramètre β, on augmente la fiabilité en stockant $n-1$ pièces de rechange. La durée de vie totale suit alors une loi gamma de paramètres (n, β). (voir l'exercice 17 du chapitre 15). Les lois gamma ont aussi des relations étroites avec la loi de Poisson (voir l'exercice 22 du chapitre 9) et avec les lois du chi-deux (voir l'exemple 6 du chapitre 15).

6. $f(x) = \begin{cases} \alpha\beta^\alpha x^{\alpha-1} e^{-(\beta x)^\alpha} & \text{si } x \geqslant 0, \\ 0 & \text{si } x < 0, \end{cases}$ est appelée la *loi de Weibull* de paramètres α et β (avec $0 < \alpha < \infty$ et $0 < \beta < \infty$). Cette loi apparaît comme une généralisation de la loi exponentielle, dans une direction différente de celle des lois gamma : voir par exemple l'exercice 23 du chapitre 9.

7. $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$ est appelée la *loi normale* de paramètres (μ, σ^2), avec $\mu \in \mathbb{R}$ et $0 < \sigma^2 < \infty$, ou *loi de Gauss*, et elle est notée $\mathrm{N}(\mu, \sigma^2)$. Cette famille de lois sera discutée en détail dans les chapitres 16 et 21 ; ce sont certainement les lois les plus importantes en probabilités ainsi qu'en statistique.

8. Soit $g_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$ la densité normale (cf. ci-dessus). Alors

$$f(x) = \begin{cases} \dfrac{1}{x} g_{\mu,\sigma^2}(\log x) & \text{si } x > 0, \\ 0 & \text{si } x \leqslant 0, \end{cases}$$

est appelée la *loi lognormale* de paramètres (μ, σ^2), (avec $\mu \in \mathbb{R}$ et $0 < \sigma^2 < \infty$). Cette famille de lois est utilisée pour modéliser des

---

2. La fonction gamma est définie par $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$, $\alpha > 0$; on a $\Gamma(\alpha) = (\alpha-1)!$ si $\alpha \in \mathbb{N}^*$, avec la convention $0! = 1$, et $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

quantités aléatoires *positives*. On les appelle aussi lois de Galton-McAlister, ou de Cobb-Douglas dans un contexte économique.

9. $f(x) = \dfrac{\beta}{2} e^{-\beta|x-\alpha|}$ est appelée la *loi exponentielle double* de paramètres α et β (avec $\alpha \in \mathbb{R}$ et $0 < \beta < \infty$), ou aussi *loi de Laplace*.

10. $f(x) = \dfrac{1}{\beta\pi} \dfrac{1}{1 + (x-\alpha)^2/\beta^2}$ est appelée la *loi de Cauchy* de paramètres α et β (avec $\alpha \in \mathbb{R}$ et $0 < \beta < \infty$). Cette famille de lois est souvent utilisée pour fournir des contre-exemples en théorie des probabilités, et a été introduite historiquement pour cette raison précise[3], du fait qu'elle présente des queues « épaisses ». Elle est utilisée en mécanique et en électricité.

Exercices des chapitres 6 et 7

1. Soit $(A_n)_{n\geqslant 1}$ une suite d'événements deux à deux disjoints sur un espace de probabilité $(\Omega, \mathcal{A}, P)$. Montrer que $\lim_{n\to\infty} P(A_n) = 0$.
2.* Soit $(A_\beta)_{\beta\in B}$ une famille d'événements deux à deux disjoints sur un espace de probabilité $(\Omega, \mathcal{A}, P)$. Montrer que si $P(A_\beta) > 0$ pour tout $\beta \in B$, alors l'ensemble d'indices B est fini ou dénombrable.
3. Montrer que le maximum de la densité de la loi gamma est atteint pour $x = \frac{\alpha-1}{\beta}$, lorsque $\alpha \geqslant 1$.
4. Montrer que le maximum de la densité de la loi de Weibull est atteint pour $x = \frac{1}{\beta}(\frac{\alpha-1}{\alpha})^{\frac{1}{\alpha}}$, lorsque $\alpha \geqslant 1$.
5. Montrer que le maximum de la densité de la loi normale est atteint pour $x = \mu$.
6. Montrer que le maximum de la densité de la loi lognormale est atteint pour $x = e^{\mu - \sigma^2}$.
7. Montrer que le maximum de la densité de la loi exponentielle double est atteint pour $x = \alpha$.
8. Montrer que les lois gamma et de Weibull contiennent comme cas particulier la loi exponentielle.

3. En fait Poisson utilisa ces lois dès 1824 pour exhiber un cas où le théorème-limite central (cf. chapitre 21) est faux. Plus tard ces lois ont joué un rôle central dans une polémique entre Cauchy et Bienaymé, d'où leur nom.

9. Montrer que les densités uniforme, normale, exponentielle double et Cauchy sont symétriques par rapport à un point qu'on déterminera.

10. Une loi est dite *unimodale* si elle admet une densité ayant un et un seul maximum (un « mode » est un maximum de la densité, ou pour certains auteurs un maximum local de la densité). Montrer que les lois normale, exponentielle, exponentielle double, de Cauchy, gamma, de Weibull et lognormale sont unimodales.

11. Soit P une probabilité sur $\mathbb{R}$ admettant une densité f. Montrer que $P(\{x\}) = 0$ pour tout x.

12. Soit P comme dans l'exercice 11. Montrer que si B est une partie finie ou dénombrable de $\mathbb{R}$, alors B est un borélien et $P(B) = 0$.

13. Soit P et B comme dans l'exercice 12. Soit A un borélien vérifiant $P(A) = \frac{1}{2}$. Montrer que $A \cup B$ est un borélien et que $P(A \cup B) = \frac{1}{2}$.

14. Soit $A_1, \ldots, A_n, \ldots$ une suite d'ensembles négligeables. Montrer que $B = \bigcup_{i=1}^{\infty} A_i$ est également négligeable.

15. Soit X une variable aléatoire définie sur un espace dénombrable. Supposons que $E(|X|) = 0$. Montrer que $X = 0$ sauf peut-être sur un ensemble négligeable. Est-il possible d'en conclure que $X = 0$ partout (i.e., $X(\omega) = 0$ pour tout ω) ? [Rép. : Non.]

16.* Soit F une fonction de répartition. Montrer que F peut avoir une infinité de points de discontinuité, mais que le nombre de ces points est au plus dénombrable.

17. Soit F la fonction de répartition donnée par

$$F(x) = \frac{1}{4} 1_{[0,\infty[}(x) + \frac{1}{2} 1_{[1,\infty[}(x) + \frac{1}{4} 1_{[2,\infty[}(x),$$

et P la probabilité admettant F pour fonction de répartition. Trouver la probabilité des événements suivants :

a) $A =]-\frac{1}{2}, \frac{1}{2}[$,
b) $B =]-\frac{1}{2}, \frac{3}{2}[$,
c) $C =]\frac{2}{3}, \frac{5}{2}[$,
d) $D = [0, 2[$,
e) $E =]3, \infty[$.

18. Soit la fonction F donnée par

$$\mathbf{F}(x) = \sum_{i=1}^{\infty} \frac{1}{2^i} 1_{[\frac{1}{i},\infty[}(x),$$

Montrer que c'est la fonction de répartition d'une probabilité P sur $\mathbb{R}$. Trouver la probabilité des événements suivants :

a) $\mathrm{A} = [1, \infty[$.

b) $\mathrm{B} = [\frac{1}{10}, \infty[$.

c) $\mathrm{C} = \{0\}$.

d) $\mathrm{D} =]0, \frac{1}{2}[$.

e) $\mathrm{E} =]-\infty, 0[$.

f) $\mathrm{G} =]0, \infty[$.

Chapitre 8

Variables aléatoires

Dans le chapitre 5 nous avons considéré des variables aléatoires définies sur un espace fini ou dénombrable Ω. Nous voulons maintenant considérer un espace Ω quelconque, éventuellement non dénombrable, cet espace étant muni d'une tribu $\mathcal{A}$ d'événements et d'une probabilité P (sur $\mathcal{A}$).

Comme nous l'avons vu au chapitre 1, une variable aléatoire est une application X de Ω dans un autre espace F. Nous cherchons en premier lieu à déterminer la probabilité pour que X appartienne à une partie donnée A de F : c'est-à-dire, la probabilité de la partie $\{X \in A\} = X^{-1}(A) = \{\omega : X(\omega) \in A\}$ (trois manières équivalentes d'écrire la même chose, la première étant la plus parlante). Pour cela, il faut évidemment que l'ensemble $\{X \in A\}$ soit dans la tribu $\mathcal{A}$, ce qui *a priori* n'est pas vrai pour une partie arbitraire A de F. Cela motive la définition suivante.

Définition 8.1. (a) *Soit* $(E, \mathcal{E})$ *et* $(F, \mathcal{F})$ *deux espaces mesurables. Une fonction* $X : E \to F$ *est appelée* mesurable *(relativement aux tribus* $\mathcal{E}$ *et* $\mathcal{F}$*) si* $X^{-1}(A) \in \mathcal{E}$ *pour tout* $A \in \mathcal{F}$. *(On écrit aussi* $X^{-1}(\mathcal{F}) \subset \mathcal{E}$ *: voir l'exercice 1 du chapitre 9.)*

(b) *Si* $(E, \mathcal{E}) = (\Omega, \mathcal{A})$, *une fonction mesurable* X *est appelée une* variable aléatoire *(en abrégé : v.a.)*

(c) *Quand* $F = \mathbb{R}$ *(resp.* $F = \mathbb{R}^d$*), on prend en général pour* $\mathcal{F}$ *la tribu borélienne* $\mathcal{B}$ *de* $\mathbb{R}$ *(resp.* $\mathcal{B}^d$ *de* $\mathbb{R}^d$*) : sauf mention contraire, ce sera* toujours *le cas dans la suite.*

Théorème 8.1. *Soit* $\mathcal{C}$ *une classe de parties de* F *telle que* $\sigma(\mathcal{C}) = \mathcal{F}$. *Pour qu'une application* X *de* E *dans* F *soit mesurable* (relativement aux tribus $\mathcal{E}$ et $\mathcal{F}$), *il faut et il suffit que* $X^{-1}(\mathcal{C}) \subset \mathcal{E}$, *c'est-à-dire* $X^{-1}(C) \in \mathcal{E}$ *pour tout* $C \in \mathcal{C}$.

Preuve. La condition nécessaire est évidente. Pour la condition suffisante, supposons que $X^{-1}(C) \in \mathcal{E}$ pour tout $C \in \mathcal{C}$. Posons $\mathcal{D} = \{A \in \mathcal{F} : X^{-1}(A) \in \mathcal{E}\}$. On a $\mathcal{C} \subset \mathcal{D}$ par hypothèse et, comme $X^{-1}(\bigcup_n A_n) = \bigcup_n X^{-1}(A_n)$, $X^{-1}(\bigcap_n A_n) = \bigcap_n X^{-1}(A_n)$ et $X^{-1}(A^c) = (X^{-1}(A))^c$, ou en d'autres termes X^{-1} « commute » avec les réunions dénombrables, les intersections dénombrables et le passage au complémentaire, on voit que $\mathcal{D}$ est aussi une tribu. Donc $\mathcal{D} \supset \sigma(\mathcal{C})$ et aussi $\mathcal{F} \supset \mathcal{D}$. Comme $\mathcal{F} = \sigma(\mathcal{C})$ on conclut que $\mathcal{F} = \mathcal{D}$, ce qui donne le résultat. ■

Un cas particulier du théorème précédent permet de caractériser de manière particulièrement simple les v.a. à valeurs réelles (i.e. quand $\mathrm{F} = \mathbb{R}$) :

Corollaire 8.1. *Soit* $(\mathrm{F}, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$ *et* X *et* X_n *des applications de* E, *espace muni de la tribu* $\mathcal{E}$, *dans* F.

(a) X *est mesurable si et seulement si* $\{\mathrm{X} \leqslant a\} = \{\omega : \mathrm{X}(\omega) \leqslant a\} = \mathrm{X}^{-1}(]-\infty, a]) \in \mathcal{E}$ *pour tout* $a \in \mathbb{R}$, *ou même si et seulement si* $\{\mathrm{X} \leqslant a\} \in \mathcal{E}$ *pour tout* $a \in \mathbb{Q}$.

(b) *Si les* X_n *sont toutes mesurables, les fonctions* $\sup_n \mathrm{X}_n$, $\inf_n \mathrm{X}_n$, $\limsup_{n\to\infty} \mathrm{X}_n$ *et* $\liminf_{n\to\infty} \mathrm{X}_n$ *sont toutes mesurables.*

(c) *Si les* X_n *sont toutes mesurables et si* X_n *converge simplement vers* X, *alors* X *est mesurable.*

Preuve. (a) D'après le théorème 2.1 on sait que $\mathcal{B} = \sigma(\mathcal{C}) = \sigma(\mathcal{C}')$, où $\mathcal{C} = \{]-\infty, a]; a \in \mathbb{R}\}$ et $\mathcal{C}' = \{]-\infty, a]; a \in \mathbb{Q}\}$. Il suffit alors d'appliquer le théorème 2.1.

(b) Comme X_n est mesurable, $\{\mathrm{X}_n \leqslant a\} \in \mathcal{E}$. Donc $\{\sup_n \mathrm{X}_n \leqslant a\} = \bigcap_n \{\mathrm{X}_n \leqslant a\} \in \mathcal{E}$ pour tout a, et $\sup_n \mathrm{X}_n$ est mesurable par (a). De manière analogue $\{\inf_n \mathrm{X}_n < a\} = (\bigcap_n \{\mathrm{X}_n \geqslant a\}^c)^c \in \mathcal{E}$ et donc $\inf_n \mathrm{X}_n$ est mesurable. On a $\limsup_{n\to\infty} \mathrm{X}_m = \inf_n \sup_{m\geqslant n} \mathrm{X}_m = \inf_n \mathrm{Y}_n$, où $\mathrm{Y}_n = \sup_{m\geqslant n} \mathrm{X}_m$. On vient de voir que chaque Y_n est mesurable et donc que $\inf_n \mathrm{Y}_n$ est aussi mesurable : donc $\limsup_{n\to\infty} \mathrm{X}_n$ est mesurable. De manière analogue $\liminf_{n\to\infty} \mathrm{X}_n = \sup_n \inf_{m\geqslant n} \mathrm{X}_m$ est mesurable.

(c) Si $\mathrm{X}_n \to \mathrm{X}$ simplement, on a $\mathrm{X} = \limsup_{n\to\infty} \mathrm{X}_n = \liminf_{n\to\infty} \mathrm{X}_n$, et la mesurabilité de X se déduit de (b). ■

Théorème 8.2. *Soit* X *mesurable de* $(\mathrm{E}, \mathcal{E})$ *dans* $(\mathrm{F}, \mathcal{F})$, *et soit* Y *mesurable de* $(\mathrm{F}, \mathcal{F})$ *dans* $(\mathrm{G}, \mathcal{G})$; *alors la composée* $\mathrm{Y} \circ \mathrm{X}$ *est mesurable de* $(\mathrm{E}, \mathcal{E})$ *dans* $(\mathrm{G}, \mathcal{G})$.

Preuve. Soit $\mathrm{A} \in \mathcal{G}$. On a $(\mathrm{Y} \circ \mathrm{X})^{-1}(\mathrm{A}) = \mathrm{X}^{-1}(\mathrm{Y}^{-1}(\mathrm{A}))$. Comme Y est mesurable, $\mathrm{B} = \mathrm{Y}^{-1}(\mathrm{A}) \in \mathcal{F}$. Comme X est mesurable, $\mathrm{X}^{-1}(\mathrm{B}) \in \mathcal{E}$. ■

Un *espace topologique* est un ensemble F muni d'une famille d'ouverts[1]. Cette famille d'ouverts est aussi appelée la *topologie* de F. La

1. Une « famille d'ouverts » est une famille de sous-ensembles qui est stable par réunion quelconque et par intersection finie, et qui contient l'ensemble F tout entier et l'ensemble vide $\emptyset$.

définition d'une *application continue* est alors la suivante : étant donnés deux espaces topologiques $(\mathrm{E},\mathcal{U})$ et $(\mathrm{F},\mathcal{V})$ (où $\mathcal{U}$ et $\mathcal{V}$ sont les familles d'ouverts de E et F respectivement), alors une application $f : \mathrm{E} \to \mathrm{F}$ est *continue* si $f^{-1}(\mathrm{A}) \in \mathcal{U}$ pour tout $\mathrm{A} \in \mathcal{V}$ (c'est-à-dire si $f^{-1}(\mathcal{V}) \subset \mathcal{U}$). La *tribu borélienne*, ou *de Borel*, d'un espace topologique $(\mathrm{E},\mathcal{U})$ est $\sigma(\mathcal{U})$: lorsque $\mathrm{E} = \mathbb{R}$, on retrouve la tribu $\mathcal{B}$ introduite avant le théorème 2.1 du chapitre 2. (La famille des ensembles ouverts n'est pas elle-même une tribu, car elle n'est pas stable par intersection dénombrable, ni par complémentation.)

Théorème 8.3. *Soit* $(\mathrm{E},\mathcal{U})$ *et* $(\mathrm{F},\mathcal{V})$ *deux espaces topologiques, et soit* $\mathcal{E}$ *et* $\mathcal{F}$ *leurs tribus boréliennes. Chaque application continue* X *de* E *dans* F *est alors mesurable* (on dit aussi que X est borélienne).

Preuve. Comme $\mathcal{F} = \sigma(\mathcal{V})$, d'après le théorème 6.4 il suffit de montrer que $\mathrm{X}^{-1}(\mathcal{V}) \subset \mathcal{E}$. Mais comme X est continue on a $\mathrm{X}^{-1}(\mathcal{V}) \subset \mathcal{U}$, d'où le résultat. ∎

Rappelons que si A est une partie de E, son *indicatrice* (ou « fonction indicatrice »), est définie par

$$1_{\mathrm{A}}(x) = \begin{cases} 1 & \text{si } x \in \mathrm{A}, \\ 0 & \text{si } x \notin \mathrm{A} \end{cases} .$$

Ainsi la fonction $x \mapsto 1_{\mathrm{A}}(x)$, écrite 1_{A}, « indique » si un point x appartient ou non à la partie A.

Théorème 8.4. *Soit* $(\mathrm{F},\mathcal{F}) = (\mathbb{R},\mathcal{B})$, *et* $(\mathrm{E},\mathcal{E})$ *un espace mesurable quelconque.*

a) *Une indicatrice* 1_{A} *sur* E *est mesurable si et seulement si* $\mathrm{A} \in \mathcal{E}$.

b) *Si* $\mathrm{X}_1,\dots,\mathrm{X}_n$ *sont des fonctions réelles mesurables sur* $(\mathrm{E},\mathcal{E})$ *et si* f *est une fonction borélienne réelle sur* $\mathbb{R}^n$, *alors* $f(\mathrm{X}_1,\dots,\mathrm{X}_n)$ *est mesurable.*

c) *Si* X, Y *sont des fonctions réelles mesurables sur* $(\mathrm{E},\mathcal{E})$, *il en est de même de* $\mathrm{X}+\mathrm{Y}$, XY, $\mathrm{X} \vee \mathrm{Y}$ (une manière concise d'écrire $\max(\mathrm{X},\mathrm{Y})$), $\mathrm{X} \wedge \mathrm{Y}$ (une manière concise d'écrire $\min(\mathrm{X},\mathrm{Y})$), *et* X/Y *(si* Y *ne s'annule pas).*

Preuve. (a) Si $\mathrm{B} \subset \mathbb{R}$, on a

$$(1_{\mathrm{A}})^{-1}(\mathrm{B}) = \begin{cases} \emptyset & \text{si } 0 \notin \mathrm{B},\ 1 \notin \mathrm{B} \\ \mathrm{A} & \text{si } 0 \notin \mathrm{B},\ 1 \in \mathrm{B} \\ \mathrm{A}^c & \text{si } 0 \in \mathrm{B},\ 1 \notin \mathrm{B} \\ \mathrm{E} & \text{si } 0 \in \mathrm{B},\ 1 \in \mathrm{B} \end{cases}$$

Le résultat est alors immédiat.

(b) La tribu borélienne $\mathcal{B}^n$ de $\mathbb{R}^n$ est engendrée par les ensembles $\prod_{i\leqslant n}]-\infty, a_i]$: ceci se montre exactement de la même manière que le théorème 2.1. Si $X : E \to \mathbb{R}^n$ désigne la fonction « vectorielle » $X = (X_1, \ldots, X_n)$, on a $X^{-1}(\prod_{i\leqslant n}]-\infty, a_i]) = \bigcap_{i\leqslant n}\{X_i \leqslant a_i\} \in \mathcal{E}$; donc X est une application mesurable de $(E, \mathcal{E})$ dans $(\mathbb{R}^n, \mathcal{B}^n)$. L'assertion désirée suit alors du théorème 8.2.

(c) Noter que la fonction $f_1 : \mathbb{R}^2 \mapsto \mathbb{R}$ donnée par $f_1(x, y) = x + y$ est continue, ainsi que les fonctions $f_2(x, y) = xy$, $f_3(x, y) = x \vee y = \max(x, y)$ et $f_4(x, y) = x \wedge y = \min(x, y)$. La fonction $f_5(x, y) = \frac{x}{y}$ est continue de $\mathbb{R} \times (\mathbb{R} \setminus \{0\})$ dans $\mathbb{R}$. Donc (c) découle de (b) et du théorème 8.3 (les fonctions continues sont boréliennes). ■

Soit maintenant X une v.a. sur un espace de probabilité $(\Omega, \mathcal{A}, P)$, à valeurs dans un espace mesurable $(E, \mathcal{E})$. La *loi* de X est alors définie par

$$P_X(B) = P(X^{-1}(B)) = (P \circ X^{-1})(B) = P(\{\omega : X(\omega) \in\}) = P(X \in B),$$

pour tout $B \in \mathcal{E}$. Les quatre manières d'écrire $P_X(B)$ du côté droit sont utilisées, mais la plus usuelle est

$$P_X(B) = P(X \in B),$$

ce qui évite de spécifier explicitement l'espace Ω : cet espace est en effet souvent difficile à caractériser mathématiquement, tandis qu'on s'intéresse en fait seulement à une ou plusieurs v.a. définies sur cet espace. Parfois on appelle aussi P_X *l'image de* P *par* X.

Comme X^{-1} commute avec les réunions et les intersections (finies, dénombrables ou infinies non dénombrables!), et comme $X^{-1}(E) = \Omega$, on a de manière immédiate :

Théorème 8.5. *La loi* P_X *de* X *est une mesure de probabilité sur* $(E, \mathcal{E})$.

Lorsque X est une v.a. réelle, sa loi P_X est une probabilité sur $\mathbb{R}$, dont on sait qu'elle est entièrement caractérisée par sa fonction de répartition :

$$F_X(x) = P_X(]-\infty, x]) = P(X \leqslant x).$$

La fonction F_X s'appelle la *fonction de répartition de la v.a.* X. Lorsque F_X admet une densité f_X (i.e. $F_X(x) = \int_{-\infty}^{x} f_X(y)dy$ pour tout $x \in \mathbb{R}$), on dit aussi que la fonction f_X est la *densité de la v.a.* X.

Chapitre 9

Intégration par rapport à une mesure de probabilité

Soit $(\Omega, \mathcal{A}, \mathrm{P})$ un espace de probabilité. Nous voulons définir *l'espérance*, ou ce qui est équivalent, « l'intégrale » d'une v.a. réelle sur cet espace. Nous avons déjà fait cela lorsque Ω est fini ou dénombrable, mais le cas général est plus délicat.

Commençons par des cas particuliers.

Définition 9.1. a) *Une v.a. réelle* X *est appelée* simple, *ou* étagée, *si elle ne prend qu'un nombre fini de valeurs, ce qui revient à dire qu'elle se met sous la forme*

$$\mathrm{X} = \sum_{i=1}^{n} a_i 1_{\mathrm{A}_i}, \tag{9.1}$$

où $a_i \in \mathbb{R}$ *et* $\mathrm{A}_i \in \mathcal{A}$, $1 \leqslant i \leqslant n$ (Cette formule définit clairement une v.a., c'est-à-dire une fonction mesurable ; inversement si X est mesurable et ne prend que les valeurs $a_1, \ldots, a_n$ elle se met sous la forme (9.1) avec $\mathrm{A}_i = \{\mathrm{X} = a_i\}$: une v.a. simple a bien sûr plusieurs représentations différentes du type (9.1).)

b) *Si* X *est une v.a. simple, son* espérance *(ou « espérance mathématique », ou « intégrale » par rapport à* P*) est le nombre*

$$\mathrm{E}(\mathrm{X}) = \sum_{i=1}^{n} a_i \mathrm{P}(\mathrm{A}_i). \tag{9.2}$$

(On l'écrit aussi $\int \mathrm{X}(\omega)\mathrm{P}(d\omega)$, *ou même plus simplement* $\int \mathrm{X}\, d\mathrm{P}$.*)*

Un calcul élémentaire montre que E(X) *ne dépend pas de la représentation* (9.1) *choisie.*

Soit X, Y deux v.a. simples et β un réel. On peut écrire X et Y comme dans (9.1), avec les mêmes événements A_i qui forment une partition de Ω, et avec les nombres a_i pour X et b_i pour Y. Ainsi, $\beta\mathrm{X}$ et $\mathrm{X}+\mathrm{Y}$ sont encore de la forme (9.1) avec les mêmes A_i et avec les nombres βa_i et $a_i + b_i$ respectivement. Donc $\mathrm{E}(\beta\mathrm{X})\beta\mathrm{E}(\mathrm{X})$ et $\mathrm{E}(\mathrm{X}+\mathrm{Y}) = \mathrm{E}(\mathrm{X}) + \mathrm{E}(\mathrm{Y})$, ce qui signifie que l'espérance est linéaire sur l'espace vectoriel des v.a. (réelles) simples. Si de plus $X \leqslant Y$ on a $a_i \leqslant b_i$ pour tout i, donc $E(X) \leqslant E(Y)$.

Ensuite, nous définissons l'espérance pour les v.a. positives, i.e. les v.a. à valeurs dans $[0, \infty]$, en incluant $+\infty$ comme valeur possible : cette extension anodine est nécessaire pour certains des résultats ultérieurs. On pose alors

$$E(X) = \sup(E(Y) : \; Y \text{ v.a. simple avec } 0 \leqslant Y \leqslant X). \tag{9.3}$$

Ce supremum existe toujours dans $[0, \infty]$. Comme $E(X) \leqslant E(Y)$ pour des variables simples vérifiant $0 \leqslant X \leqslant Y$, il est clair que (9.3) et (9.2) coïncident lorsque X est simple.

Noter que $E(X) \geqslant 0$, et qu'on peut avoir $E(X) = \infty$, même si X ne prend que des valeurs finies.

Finalement soit X une v.a. réelle arbitraire. Soit $X^+ = \max(X, 0)$ et $X^- = -\min(X, 0)$. On a $X = X^+ - X^-$, et X^+ et X^- sont des v.a. positives, et on a aussi $|X| = X^+ + X^-$.

Définition 9.2. (a) *On dit qu'une v.a.* X *a une* espérance finie *(ou « est intégrable ») si* $E(X^+)$ *et* $E(X^-)$ *sont toutes deux finies* (on verra plus tard que cela revient à dire que $E(|X|)$ est finie). *Dans ce cas, son espérance est le nombre*

$$E(X) = E(X^+) - E(X^-), \tag{9.4}$$

qu'on écrit aussi $\int X(\omega) dP(\omega)$ *ou* $\int X\, dP$. (Si $X \geqslant 0$ on a $X^- = 0$ et $X^+ = X$, donc, comme clairement $E(0) = 0$, cette définition coïncide avec (9.3)).

On note $\mathcal{L}^1$ *l'ensemble de toutes les v.a. réelles intégrables (parfois* $\mathcal{L}^1(\Omega, \mathcal{A}, P)$ *pour éviter toute ambiguïté).*

(b) *Une v.a.* X admet une espérance *si* $E(X^+)$ *et* $E(X^-)$ *ne sont pas tous deux simultanément infinis. Dans ce cas, son espérance est encore donnée par (9.4), avec les conventions* $+\infty + \alpha = +\infty$ *et* $-\infty + \alpha = -\infty$ *pour* $\alpha \in \mathbb{R}$. (Si $X \geqslant 0$ cette définition coïncide encore avec (9.3) ; noter que si X admet une espérance, on a $E(X) \in [-\infty, +\infty]$, et X est intégrable si et seulement si son espérance existe et est finie.)

Remarque. Lorsque Ω est fini ou dénombrable, on a donc apparemment deux définitions différentes pour l'espérance d'une v.a. X, celle ci-dessus et celle du chapitre 5. Bien entendu ces deux définitions coïncident : il suffit de le vérifier pour les v.a. simples, et pour celles-ci on observe que les deux formules (9.2) avec $A_i = \{X = a_i\}$ et (5.1) sont en fait identiques.

Le théorème suivant présente les propriétés les plus importantes de la fonctionnelle espérance. La preuve des parties (d), (e) et (f) est difficile et peut être omise. Dans son énoncé, l'assertion « X = Y presque sûrement », ou en abrégé « p.s. », signifie que l'événement $\{X \neq Y\}$ est négligeable, c'est-à-dire que $P(X \neq Y) = 0$ ou de manière équivalente $P(X = Y) = 1$; de même « $X \geqslant Y$ p.s. » signifie $P(X \geqslant Y) = 1$, et « X_n converge p.s. vers X » que l'ensemble des ω pour lesquels la suite numérique $X_n(\omega)$ ne converge pas vers $X(\omega)$ est négligeable.

Théorème 9.1. (a) *$\mathcal{L}^1$ est un espace vectoriel et l'espérance est une forme linéaire sur $\mathcal{L}^1$, et aussi une fonctionnelle positive (i.e., $X \geqslant 0 \Rightarrow E(X) \geqslant 0$). Si de plus X et Y sont deux v.a. avec $0 \leqslant X \leqslant Y$ et $Y \in \mathcal{L}^1$, alors $X \in \mathcal{L}^1$ et $E(X) \leqslant E(Y)$.*

(b) *$X \in \mathcal{L}^1$ si et seulement si $|X| \in \mathcal{L}^1$, ou de manière équivalente si $E(|X|) < \infty$, et dans ce cas $|E(X)| \leqslant E(|X|)$. En particulier toute v.a. bornée est intégrable.*

(c) *Si X = Y presque sûrement, alors $E(X) = E(Y)$.*

(d) *(Théorème de convergence monotone.) Si les v.a. X_n sont* positives *et si la suite (X_n) croît p.s. vers X, alors $\lim_{n\to\infty} E(X_n) = E(X)$ (même si $E(X) = \infty$).*

(e) *(Lemme de Fatou.) Si les v.a. X_n vérifient $X_n \geqslant Y$ p.s., avec Y intégrable, les v.a. X_n et $\liminf_{n\to\infty} X_n$ admettent une espérance, et on a $E(\liminf_{n\to\infty} X_n) \leqslant \liminf_{n\to\infty} E(X_n)$. Cela est vrai en particulier si $X_n \geqslant 0$ p.s. pour tout n.*

(f) *(Théorème de convergence dominée de Lebesgue.) Si les v.a. X_n convergent p.s. vers X et si $|X_n| \leqslant Y$ p.s. pour tout n, où Y est une v.a. intégrable, alors $X_n \in \mathcal{L}^1$, $X \in \mathcal{L}^1$, et $E(X_n) \to E(X)$.*

L'égalité p.s. entre v.a. est clairement une relation d'équivalence, et deux v.a. équivalentes (i.e. p.s. égales) ont la même espérance : On peut donc définir l'espace L^1 comme « $\mathcal{L}^1$ modulo cette relation d'équivalence ». En d'autres termes un élément de L^1 est une classe d'équivalence, c'est-à-dire l'ensemble de toutes les v.a. qui sont p.s. égales à une v.a. particulière. Vu (c) ci-dessus, on peut parler de « l'espérance » de cette classe d'équivalence, qui est l'espérance de n'importe laquelle des v.a. appartenant à cette classe. Comme de plus l'addition de deux v.a. et la multiplication d'une v.a. par une constante préservent l'égalité p.s., l'ensemble L^1 est aussi un espace vectoriel. Par suite on commet l'abus (anodin) d'identifier une v.a. avec sa classe d'équivalence, et on écrit d'habitude $X \in L^1$ au lieu de $X \in \mathcal{L}^1$.

Si $1 < p < \infty$, on définit $\mathcal{L}^p$ comme l'ensemble des v.a. X telles que $|X|^p \in \mathcal{L}^1$; L^p est défini de manière analogue à L^1 : c'est $\mathcal{L}^p$ modulo la relation d'équivalence « égalité p.s. » ; ou de manière équivalente, deux éléments de $\mathcal{L}^p$ qui sont p.s. égaux sont considérés comme deux représentants du même élément de L^p. Nous n'utiliserons vraiment dans ce livre que les espaces L^1 et L^2.

Avant de donner la preuve du théorème 9.1, nous montrons deux résultats auxiliaires.

Résultat 1. *Pour toute v.a. positive X il existe une suite $(X_n)_{n\geqslant 1}$ de v.a. positives simples qui croît vers X quand $n \to \infty$.* Un exemple d'une telle suite est donné par

$$X_n(\omega) = \begin{cases} k2^{-n} & \text{si } k2^{-n} \leqslant X(\omega) < (k+1)2^{-n} \text{ et } 0 \leqslant k \leqslant n2^n - 1, \\ n & \text{si } X(\omega) \geqslant n. \end{cases} \tag{9.5}$$

Résultat 2. *Si X est une v.a. positive et si $(X_n)_{n\geqslant 1}$ est une suite arbitraire de v.a. positives simples croissant vers X, alors $E(X_n)$ croît vers $E(X)$.*

Pour vérifier ceci, on observe d'abord que la suite $E(X_n)$ croît vers une limite a qui vérifie $a \leqslant E(X)$ par (9.3). Pour obtenir qu'en fait $a = E(X)$, et au vu de (9.3) encore, il suffit clairement de montrer que si Y est une v.a. simple telle que $0 \leqslant Y \leqslant X$, alors $E(Y) \leqslant a$.

La variable Y prend k valeurs différentes, disons $a_1, \ldots, a_k$, et on pose $A_k = \{Y = a_k\}$. Choisissons $\varepsilon \in]0,1]$. La v.a. $Y_{n,\varepsilon} = (1-\varepsilon)Y1_{\{(1-\varepsilon)Y\leqslant X_n\}}$ prend les valeurs $(1-\varepsilon)a_k$ sur l'ensemble $A_{k,n,\varepsilon} = A_k \cap \{(1-\varepsilon)Y \leqslant X_n\}$ et 0 sur l'ensemble $\{(1-\varepsilon)Y > X_n\}$. De plus il est évident que $Y_{n,\varepsilon} \leqslant X_n$, donc d'après (9.2) il vient

$$E(Y_{n,\varepsilon}) = (1-\varepsilon)\sum_{i=1}^{k} a_k P(A_{k,n,\varepsilon}) \leqslant E(X_n). \tag{9.6}$$

Maintenant, rappelons que $Y \leqslant \lim_n X_n$, donc $(1-\varepsilon)Y < \lim_n X_n$ dès que $Y > 0$, de sorte que $A_{k,n,\varepsilon} \to A_k$ quand $n \to \infty$. Une application du théorème 2.4 entraîne $P(A_{k,n,\varepsilon}) \to P(A_k)$, donc en passant à la limite dans (9.6) on obtient

$$(1-\varepsilon)\sum_{i=1}^{k} a_k \; P(A_k) = (1-\varepsilon)E(Y) \leqslant a.$$

Il reste à faire $\varepsilon \to 0$ ci-dessus pour obtenir $\mathrm{E}(\mathrm{Y}) \leqslant a$, d'où le résultat 2.

Preuve du théorème 9.1. (a) Soit X et Y deux v.a. positives, et $\alpha \in \mathbb{R}_+$. Le résultat 1 permet d'obtenir deux suites (X_n) et (Y_n) de v.a. positives simples, croissant vers X et Y respectivement. Les v.a. $\mathrm{X}_n + \mathrm{Y}_n$ et $\alpha \mathrm{X}_n$ sont simples et croissent vers $\mathrm{X}+\mathrm{Y}$ et $\alpha\mathrm{X}$ respectivement, tandis que $\mathrm{E}(\mathrm{X}_n + \mathrm{Y}_n) = \mathrm{E}(\mathrm{X}_n) + \mathrm{E}(\mathrm{Y}_n)$ et $\mathrm{E}(\alpha\mathrm{X}_n) = \alpha\mathrm{E}(\mathrm{X}_n)$: en appliquant le résultat 2 on obtient alors $\mathrm{E}(\mathrm{X}+\mathrm{Y}) = \mathrm{E}(\mathrm{X})+\mathrm{E}(\mathrm{Y})$ et $\mathrm{E}(\alpha\mathrm{X}) = \alpha\mathrm{E}(\mathrm{X})$. Si de plus $\mathrm{X} \leqslant \mathrm{Y}$, l'inégalité $\mathrm{E}(\mathrm{X}) \leqslant \mathrm{E}(\mathrm{Y})$ est une conséquence immédiate de (9.3).

Ce qui précède entraîne les deux dernières assertions de (a). Comme pour des v.a. réelles quelconques X et Y et pour $\alpha \in \mathbb{R}$ on a $(\alpha\mathrm{X})^+ + (\alpha\mathrm{X})^- \leqslant |\alpha|(\mathrm{X}^+ + \mathrm{X}^-)$ et $(\mathrm{X}+\mathrm{Y})^+ + (\mathrm{X}+\mathrm{Y})^- \leqslant \mathrm{X}^+ + \mathrm{X}^- + \mathrm{Y}^+ + \mathrm{Y}^-$, on en déduit ensuite que $\mathcal{L}^1$ est un espace vectoriel. Finalement, comme $\mathrm{E}(\mathrm{X}) = \mathrm{E}(\mathrm{X}^+) - \mathrm{E}(\mathrm{X}^-)$ on déduit que l'espérance est linéaire.

(b) Comme $|\mathrm{X}| = \mathrm{X}^+ + \mathrm{X}^-$ la première assertion est évidente. Si de plus $\mathrm{X} \in \mathcal{L}^1$, on a $|\mathrm{E}(\mathrm{X})| \leqslant \mathrm{E}(\mathrm{X}^+) + \mathrm{E}(\mathrm{X}^-) = \mathrm{E}(|\mathrm{X}|)$. Enfin, comme $\mathrm{E}(\mathrm{Y}) = \alpha$ si $\mathrm{Y} = \alpha \in [0, \infty[$, la dernière assertion est également évidente.

(c) Supposons que $\mathrm{X} = \mathrm{Y}$ p.s. et supposons d'abord que $\mathrm{X} \geqslant 0$ et $\mathrm{Y} \geqslant 0$. Si $\mathrm{A} = \{\omega : \mathrm{X}(\omega) \neq \mathrm{Y}(\omega)\} = \{\mathrm{X} \neq \mathrm{Y}\}$ on a $\mathrm{P}(\mathrm{A}) = 0$. De plus

$$\mathrm{E}(\mathrm{Y}) = \mathrm{E}(\mathrm{Y}1_\mathrm{A} + \mathrm{Y}1_{\mathrm{A}^c}) = \mathrm{E}(\mathrm{Y}1_\mathrm{A}) + \mathrm{E}(\mathrm{Y}1_{\mathrm{A}^c}) = \mathrm{E}(\mathrm{Y}1_\mathrm{A}) + \mathrm{E}(\mathrm{X}1_{\mathrm{A}^c}).$$

Soit une suite (Y_n) de v.a. simples positives croissant vers Y. Les v.a. $\mathrm{Y}_n 1_\mathrm{A}$ sont aussi simples et croissent vers $\mathrm{Y}1_\mathrm{A}$. Comme Y_n est simple elle est bornée, disons par N_n, et on a

$$0 \leqslant \mathrm{E}(\mathrm{Y}_n 1_\mathrm{A}) \leqslant \mathrm{E}(\mathrm{N}_n 1_\mathrm{A}) = \mathrm{N}_n \mathrm{P}(\mathrm{A}) = 0.$$

Donc $\mathrm{E}(\mathrm{Y}1_\mathrm{A}) = \lim_{n\to\infty} \mathrm{E}(\mathrm{Y}_n 1_\mathrm{A}) = 0$. De même $\mathrm{E}(\mathrm{X}1_\mathrm{A}) = 0$. Finalement, il vient

$$\begin{aligned}\mathrm{E}(\mathrm{Y}) = \mathrm{E}(\mathrm{Y}1_\mathrm{A}) + \mathrm{E}(\mathrm{X}1_{\mathrm{A}^c}) = 0 + \mathrm{E}(\mathrm{X}1_{\mathrm{A}^c}) = \mathrm{E}(\mathrm{X}1_\mathrm{A}) + \mathrm{E}(\mathrm{X}1_{\mathrm{A}^c}) \\ = \mathrm{E}(\mathrm{X}).\end{aligned}$$

Lorsque X et Y sont de signe quelconque, la propriété $\mathrm{X} = \mathrm{Y}$ p.s. entraîne aussi $\mathrm{Y}^+ = \mathrm{X}^+$ p.s. et $\mathrm{Y}^- = \mathrm{X}^-$ p.s., et (c) s'ensuit.

(d) Pour chaque n fixé on choisit une suite $(\mathrm{Y}_{n,k})_{k\geqslant 1}$ de v.a. simples positives croissant vers X_n (quand $k \to \infty$), d'après le résultat 1. Posons

$$\mathrm{Z}_k = \max_{n \leqslant k} \mathrm{Y}_{n,k}.$$

Alors $(Z_k)_{k\geqslant 1}$ est une suite croissante de v.a. simples positives, admettant donc une limite $Z = \lim_{k\to\infty} Z_k$. De plus

$$Y_{n,k} \leqslant Z_k \leqslant X_k \leqslant X \quad \text{p.s.} \quad \text{pour } n \leqslant k$$

ce qui entraine

$$X_n \leqslant Z \leqslant X \quad \text{p.s.}$$

En faisant $n \to \infty$ on obtient $Z = X$ p.s. L'espérance étant une fonctionnelle positive,

$$E(Y_{n,k}) \leqslant E(Z_k) \leqslant E(X_k) \qquad \text{pour } n \leqslant k.$$

Fixons n et faisons $k \to \infty$. Le résultat 2 implique

$$E(X_n) \leqslant E(Z) \leqslant \lim_{k\to\infty} E(X_k).$$

En faisant maintenant $n \to \infty$ on obtient :

$$\lim_{n\to\infty} E(X_n) \leqslant E(Z) \leqslant \lim_{k\to\infty} E(X_k),$$

et comme les deux membres extrêmes sont égaux, ils sont aussi égaux à celui du milieu. On déduit alors le résultat de (c) et de l'égalité $X = Z$ p.s.

(e) Posons $\widehat{X}_n = X_n - Y$, de sorte que $\widehat{X}_n \geqslant 0$. Soit $Z_n = \inf_{k\geqslant n} \widehat{X}_k$. Les Z_n sont de v.a. positives, croissant vers la v.a. positive $\liminf_{n\to\infty} \widehat{X}_n$, et on a aussi $Z_n \leqslant \widehat{X}_n$. Par suite (d) implique que

$$E(\liminf_{n\to\infty} \widehat{X}_n) = \lim_{n\to\infty} E(Z_n) = \liminf_{n\to\infty} E(Z_n) \leqslant \liminf_{n\to\infty} E(\widehat{X}_n). \tag{9.7}$$

Par ailleurs si une v.a. U vérifie $U \geqslant Y$ p.s. avec $Y \in \mathcal{L}^1$ on a $U^- \leqslant |Y|$ p.s., donc $E(U^-) \leqslant E(|Y|) < \infty$, de sorte que U admet une espérance : on en déduit que les X_n et $\liminf_{n\to\infty} X_n$ admettent toutes une espérance. Enfin, on a $E(X_n) = E(\widehat{X}_n) + E(Y)$ et $E(\liminf_{n\to\infty} X_n) = E(\lim_{n\to\infty} \widehat{X}_n) + E(Y)$, donc l'inégalité cherchée se déduit immédiatement de (9.7).

(f) Soit $U = \liminf_{n\to\infty} X_n$ et $V = \limsup_{n\to\infty} X_n$. Par hypothèse $U = V = X$ p.s. On a aussi $|X_n| \leqslant Y$ p.s. et $|X| \leqslant Y$ p.s., donc $E(|X_n|) \leqslant E(Y) < \infty$, donc X_n est intégrable, et de même X est intégrable. D'une part, $X_n \geqslant -Y$ p.s. et $-Y \in \mathcal{L}^1$, donc (e) entraine

$$E(U) \leqslant \liminf_{n\to\infty} E(X_n).$$

On a aussi $-X_n \geqslant -Y$ p.s. et $-V = \liminf_{n\to\infty} -X_n$, de sorte que

$$-E(V) = E(-V) \leqslant \liminf_{n\to\infty} E(-X_n) = -\limsup_{n\to\infty} E(X_n).$$

En rassemblant ces deux résultats, et grâce à (c), on arrive à

$$E(X) = E(U) \leqslant \liminf_{n\to\infty} E(X_n) \leqslant \limsup_{n\to\infty} E(X_n) \leqslant E(V) = E(X).$$

Les deux membres extrêmes étant égaux, tous les termes de l'inéquation ci-dessus sont égaux, et on déduit le résultat. ■

Une conséquence utile du théorème de convergence dominée de Lebesgue (partie (f) du théorème 9.1) est le résultat suivant, qui permet d'intervertir sommation et espérance. Comme la somme d'une série est la limite des sommes partielles et que l'espérance est en fait aussi une limite, cela revient à changer l'ordre dans lequel on prend les deux limites.

Théorème 9.2. *Soit X_n une suite de v.a.*

(a) *Si les X_n sont toutes positives, alors*

$$E\left(\sum_{n=1}^{\infty} X_n\right) = \sum_{n=1}^{\infty} E(X_n). \tag{9.8}$$

les deux membres étant simultanément finis ou infinis.

(b) *Si $\sum_{n=1}^{\infty} E(|X_n|) < \infty$, alors la série $\sum_{n=1}^{\infty} X_n$ converge p.s., sa somme est intégrable, et on a (9.8).*

Preuve. Soit $S_n = \sum_{k=1}^{n} |X_k|$ et $T_n = \sum_{k=1}^{n} X_k$. On a

$$E(S_n) = E\left(\sum_{k=1}^{n} |X_k|\right) = \sum_{k=1}^{n} E(|X_k|).$$

et la suite S_n croît vers la limite $S = \sum_{k=1}^{\infty} |X_k|$ (qui peut être infinie pour certaines valeurs de ω et finie pour d'autres). Donc le théorème de convergence monotone 9.1(d) implique :

$$E(S) = \lim_{n\to\infty} E(S_n) = \sum_{k=1}^{\infty} E(|X_k|) < \infty.$$

Si les X_n sont toutes positives on a $S_n = T_n$ et on a donc prouvé (a). Si les X_n ne sont pas nécessairement positives, mais si on a $\sum_{n=1}^{\infty} E(|X_n|) < \infty$, on obtient aussi que $E(S) < \infty$.

Maintenant, pour chaque $\varepsilon > 0$ on a $1_{\{S=\infty\}} \leqslant \varepsilon S$, donc

$$P(S = \infty) = E(1_{\{S=\infty\}}) \leqslant \varepsilon E(S).$$

Par suite, $E(S) < \infty$ et le fait que ε est arbitrairement proche de 0 entraînent que $P(S = \infty) = 0$; donc $\sum_{k=1}^{\infty} X_k$ est p.s. une série absolument convergente, et sa somme, disons T, est la limite des T_n. De plus

$$|T_n| \leqslant S_n \leqslant S$$

et $S \in \mathbf{L}^1$. Par suite le théorème de convergence dominée (théorème 9.1(f)) implique

$$E\left(\sum_{k=1}^{\infty} X_k\right) = E(\lim_{n\to\infty} T_n) = E(T),$$

ce qui est (9.8). ■

Rappelons que L^1 et L^2 sont les ensembles de classes d'équivalence de v.a. intégrables (resp. de carré intégrable), pour la relation d'équivalence « égalité p.s. ».

Théorème 9.3. (a) *Si* X *et* Y *sont dans* $\mathbf{L}^2$, *on a* $XY \in L^1$ *et* l'inégalité de Cauchy-Schwarz :

$$|E(XY)| \leqslant \sqrt{E(X^2)E(Y^2)};$$

(b) *On a* $L^2 \subset L^1$, *et si* $X \in \mathbf{L}^2$ *alors* $E(X)^2 \leqslant E(X^2)$;

(c) *L'espace* $\mathbf{L}^2$ *est un espace vectoriel : si* $X, Y \in L^2$ *et* $\alpha, \beta \in \mathbb{R}$, *alors* $\alpha X + \beta Y \in \mathbf{L}^2$ (nous verrons au chapitre 22 qu'en fait L^2 est un espace de Hilbert).

Preuve. (a) On a $|XY| \leqslant X^2/2 + Y^2/2$, donc $X, Y \in L^2$ implique $XY \in L^1$. Pour tout $x \in \mathbb{R}$ on a

$$0 \leqslant E((xX + Y)^2) = x^2 E(X^2) + 2xE(XY) + E(Y^2). \qquad (9.9)$$

Le discriminant du polynôme du second degré (en x) dans le membre de droite ci-dessus vaut

$$4\left\{(E(XY))^2 - E(X^2)E(Y^2)\right\},$$

et comme l'expression (9.9) est toujours positive le discriminant doit être négatif, ce qui donne l'inégalité de Cauchy-Schwarz.

(b) Soit $X \in L^2$. Comme $X = X \cdot 1$ et comme la fonction identiquement égale à 1 appartient à L^2 avec $E(1^2) = 1$, l'assertion se déduit de (a).

(c) Soit $X, Y \in L^2$. Pour toutes constantes α et β on a que $(\alpha X + \beta Y)^2 \leqslant \alpha^2 X^2/2 + \beta^2 Y^2/2$ est intégrable : donc $\alpha X + \beta Y \in L^2$ et L^2 est un espace vectoriel. ■

Si $X \in L^2$, la *variance* de X, notée aussi $\sigma^2(X)$ ou σ^2_X, est

$$\mathrm{Var}\,(X) = \sigma^2(X) = E((X - E(X))^2).$$

(Noter que $X \in L^2 \Rightarrow X \in L^1$, donc $E(X)$ existe.) Si $\mu = E(X)$ on a

$$\begin{aligned}\mathrm{Var}\,(X) = E((X-\mu)^2) &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2.\end{aligned}$$

On a donc aussi l'égalité suivante, triviale mais très utile :

$$\sigma^2(X) = E(X^2) - E(X)^2.$$

Théorème 9.4. (Inégalité de Bienaymé-Chebyshev.) *On a*

$$P(|X| \geqslant a) \leqslant \frac{E(X^2)}{a^2}.$$

Preuve. Sur l'ensemble $\{|X| \geqslant a\}$ on a $a^2 1_{\{|X| \geqslant a\}} = a^2 \leqslant X^2$, et sur le complémentaire de cet ensemble on a $a^2 1_{\{|X| \geqslant a\}} = 0 \leqslant X^2$. Par suite la variable aléatoire $a^2 1_{\{|X| \geqslant a\}}$ est majorée partout par la variable aléatoire X^2, et on a $E(a^2 1_{\{|X| \geqslant a\}}) \leqslant E(X^2)$. Donc

$$a^2 P(|X| \geqslant a) \leqslant E(X^2);$$

et le résultat s'obtient en divisant par a^2. ■

Cette inégalité est souvent écrite pour la variable $X - E(X)$, ce qui donne :

$$P(|X - E(X)| \geqslant a) \leqslant \frac{\sigma^2(X)}{a^2}.$$

Le théorème suivant est très utile, de même que le corollaire qui le suit, pour les calculs effectifs d'espérance. Il montre que l'espérance d'une variable aléatoire ne dépend que de sa loi, et il s'agit d'une version de la formule « de changement de variable » dans les intégrales, vue sous un aspect « abstrait ».

Théorème 9.5. *Soit* X *une v.a. sur* $(\Omega, \mathcal{A}, \mathrm{P})$. *à valeurs dans* $(\mathrm{E}, \mathcal{E})$, *et de loi* $\mathrm{P_X}$. *Soit* $h : (\mathrm{E}, \mathcal{E}) \mapsto (\mathbb{R}, \mathcal{B})$ *une application mesurable.*

(a) *On a* $h(\mathrm{X}) \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathrm{P})$ *si et seulement si* $h \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}, \mathrm{P_X})$.

(b) *Si* h *est positive, ou si elle satisfait les conditions équivalentes de* (a), *on a*
$$\mathrm{E}(h(\mathrm{X})) = \int h(x)\mathrm{P_X}(dx). \tag{9.10}$$

Preuve. En se rappelant que la loi $\mathrm{P_X}$ est définie par $\mathrm{P_X(B)}=\mathrm{P(X^{-1}(B))}$, on voit que
$$\mathrm{E}(1_\mathrm{B}(\mathrm{X})) = \mathrm{P}(\mathrm{X}^{-1}(\mathrm{B})) = \mathrm{P^X(B)} = \int 1_\mathrm{B}(x)\mathrm{P^X}(dx).$$

Par suite si h est simple, on a (9.10) par linéarité de l'espérance. Si h est positive on choisit une suite (h_n) de fonctions simples et positives, qui croît vers h. On a
$$\begin{aligned}
\mathrm{E}(h(\mathrm{X})) &= \mathrm{E}(\lim_{n\to\infty} h_n(\mathrm{X})) \\
&= \lim_{n\to\infty} \mathrm{E}(h_n(\mathrm{X})) \\
&= \lim_{n\to\infty} \int h_n(x)\mathrm{P_X}(dx) \\
&= \int \lim_{n\to\infty} h_n(x)\mathrm{P_X}(dx) \\
&= \int h(x)\mathrm{P_X}(dx)
\end{aligned}$$

en utilisant deux fois le théorème de convergence monotone. Cela prouve (b) quand h est positive. En l'appliquant à $|h|$, cela prouve aussi (a) (appelons que $\mathrm{X} \in \mathcal{L}^1$ si et seulement si $\mathrm{E}(|\mathrm{X}|) < \infty$). Enfin en écrivant $h = h^+ - h^-$ on déduit (9.10) pour h de signe quelconque de la même formule pour h positive par soustraction. ∎

Le résultat suivant peut se déduire du théorème 9.5, mais nous le prouverons en fait au chapitre 11 (corollaire 11.1) et nous omettons donc la preuve ici.

Corollaire 9.1. *Soit* X *une v.a. de densité* f *(i.e.* $\mathrm{P}(\mathrm{X} \leqslant x) = \int_{-\infty}^{x} f(u)du$ *pour tout* $x \in \mathbb{R}$*). Si* $\mathrm{E}(|h(\mathrm{X})|) < \infty$ *si* h *est positive, on a* $\mathrm{E}(h(\mathrm{X})) = \int h(x)f(x)dx$.

Exemples.

1. Soit X de loi exponentielle de paramètre α. Alors

$$\mathrm{E}(h(\mathrm{X})) = \int_0^\infty h(x)\alpha e^{-\alpha x} dx.$$

En particulier, si $h(x) = x$ il vient

$$\mathrm{E}(\mathrm{X}) = \int_0^\infty \alpha x e^{-\alpha x} dx = \frac{1}{\alpha},$$

par intégration par parties. Ainsi, la moyenne d'une variable exponentielle est $1/\alpha$.

2. Soit X de loi $\mathrm{N}(\mu, \sigma^2)$ (loi normale, ou gaussienne). Alors $\mathrm{E}(\mathrm{X}) = \mu$. Pour voir ceci, on écrit

$$\mathrm{E}(\mathrm{X}) = \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}\sigma}\ x\ e^{-(x-\mu)^2/2\sigma^2} dx.$$

et on fait le changement de variable $y = x - \mu$, donc $x = y + \mu$, et il vient

$$\begin{aligned}\mathrm{E}(\mathrm{X}) &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}\sigma}(y+\mu)e^{-y^2/2\sigma^2} dy \\ &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}\sigma} y e^{-y^2/2\sigma^2} dy + \mu \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2} dy.\end{aligned}$$

La première intégrale est l'intégrale d'une fonction impaire intégrable, donc est nulle ; la seconde égale $\mu \int_{-\infty}^\infty f(x)dx = \mu \cdot 1 = \mu$.

3. Soit X une variable de Cauchy de densité $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$. On a $\mathrm{E}(|\mathrm{X}|) = \infty$, puisque

$$\mathrm{E}(|\mathrm{X}|) = \int_{-\infty}^\infty \frac{|x|}{\pi(1+x^2)} dx = \frac{2}{\pi}\int_{-\infty}^\infty \frac{x}{1+x^2} dx \geqslant \frac{2}{\pi}\int_1^\infty \frac{1}{2x} dx = \infty$$

(en effet $\frac{x}{1+x^2} \geqslant 0$ pour $x \geqslant 0$ et $\frac{x}{1+x^2} \geqslant \frac{1}{2x}$ pour $x > 1$). Par suite l'espérance $\mathrm{E}(\mathrm{X})$ *n'existe pas.*

Exercices des chapitres 8 et 9

1. Soit $X : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}, \mathcal{B})$ une v.a. Posons
$$\mathcal{F} = \{A : A = X^{-1}(B), B \in \mathcal{B}\} = X^{-1}(\mathcal{B}).$$
Montrer que X est mesurable en tant qu'application de $(\Omega, \mathcal{F})$ dans $(\mathbb{R}, \mathcal{B})$.

2.* Soit $(\Omega, \mathcal{A}, P)$ un espace de probabilité, et soit $\mathcal{F}$ et $\mathcal{G}$ deux tribus sur Ω. Supposons que $\mathcal{F} \subset \mathcal{A}$ et $\mathcal{G} \subset \mathcal{A}$ (on dit alors que $\mathcal{F}$ et $\mathcal{G}$ sont des *sous-tribus* de $\mathcal{A}$). Les tribus $\mathcal{F}$ et $\mathcal{G}$ sont dites *indépendantes* si pour tous $A \in \mathcal{F}$ et $B \in \mathcal{G}$ on a $P(A \cap B) = P(A)P(B)$. Supposons alors $\mathcal{F}$ et $\mathcal{G}$ indépendantes, et soit $X : \Omega \mapsto (\mathbb{R}, \mathcal{B})$ une fonction mesurable à la fois pour la tribu $\mathcal{F}$ et pour la tribu $\mathcal{G}$. Montrer que X est p.s. constante, i.e. $P(X = c) = 1$ pour une certaine constante c.

3.* Soit $(\Omega, \mathcal{A}, P)$ un espace de probabilité et $\mathcal{A}' = \{A \cup N : A \in \mathcal{A}, N \subset \mathcal{N}\}$, où $\mathcal{N}$ est la classe des ensembles négligeables (comme dans le théorème 6.4). Soit X et Y deux applications de Ω dans $\mathbb{R}$ telles que l'ensemble $\{X \neq Y\}$ appartienne à $\mathcal{N}$ (on écrit encore $X = Y$ p.s.). Montrer que $X : (\Omega, \mathcal{A}') \mapsto (\mathbb{R}, \mathcal{B})$ est mesurable si et seulement si $Y : (\Omega, \mathcal{A}') \mapsto (\mathbb{R}, \mathcal{B})$ est mesurable.

4.* Soit $X \in \mathcal{L}^1$ sur $(\Omega, \mathcal{A}, P)$ et considérons une suite d'événements (A_n) telle que $\lim_{n\to\infty} P(A_n) = 0$. Montrer que $\lim_{n\to\infty} E(X 1_{A_n}) = 0$. (*Attention* : nous ne supposons pas que $\lim_{n\to\infty} X 1_{A_n} = 0$ p.s.)

5.* Soit $(\Omega, \mathcal{A}, P)$ un espace de probabilité et X une v.a. réelle telle que $X \geqslant 0$ p.s. et $E(X) = 1$. Définissons $Q : \mathcal{A} \to \mathbb{R}$ par $Q(A) = E(X 1_A)$. Montrer que Q est une mesure de probabilité sur $(\Omega, \mathcal{A})$.

6. Pour Q comme dans l'exercice 5, montrer que si $P(A) = 0$, alors $Q(A) = 0$. Donner un exemple montrant que $Q(A) = 0$ n'implique pas en général que $P(A) = 0$.

7.* Pour Q comme dans l'exercice 5, et en notant E_Q l'espérance par rapport à Q, montrer que $E_Q(Y) = E_P(YX)$ pour toute v.a. Y positive ou bornée.

8. Soit Q comme dans l'exercice 5, et supposons que $P(X > 0) = 1$.
   (a) Montrer que $\frac{1}{X}$ est intégrable pour Q.
   (b) Définissons $R : \mathcal{A} \to \mathbb{R}$ par $R(A) = E_Q(\frac{1}{X} 1_A)$. Montrer que R est exactement la probabilité P. (*Indication* : utiliser l'exercice 7.)

9. Soit Q comme dans l'exercice 8. Montrer que $Q(A) = 0$ implique $P(A) = 0$ (comparer avec l'exercice 6).
10. Soit X une v.a. de loi uniforme sur $[a, b]$. Montrer que $E(X) = \frac{a+b}{2}$.
11. Soit X une v.a. de densité f, et $\mu = E(X)$ (qu'on suppose exister et être finie). Montrer que

$$\mathrm{Var}(X) = \sigma^2(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx.$$

12. Soit X une v.a. de loi uniforme sur $[a, b]$. Montrer que $\sigma^2(X) = \frac{(b-a)^2}{12}$.
13. Soit X une v.a. de Cauchy de densité $\frac{1}{\pi(1+(x-\alpha)^2)}$. Montrer que $\sigma^2(X)$ n'est pas définie et que $E(X^2) = \infty$.
14. La *fonction bêta* est $B(r, s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}$, où Γ est la fonction gamma. De manière équivalente,

$$B(r, s) = \int_0^1 t^{r-1}(1-t)^{s-1}dt \quad (r > 0, s > 0).$$

Une v.a. X est dite *de loi bêta* de paramètre r et s si elle admet la densité

$$f(x) = \begin{cases} \dfrac{x^{r-1}(1-x)^{s-1}}{B(r,s)} & \text{si } 0 \leqslant x \leqslant 1, \\ 0 & \text{si } x < 0 \text{ ou } x > 1. \end{cases}$$

Montrer alors que

$$E(X^k) = \frac{B(r+k, s)}{B(r,s)} = \frac{\Gamma(r+k)\Gamma(r+s)}{\Gamma(r)\Gamma(r+s+k)},$$

pour $k \geqslant 0$. En déduire que

$$E(X) = \frac{r}{r+s},$$
$$\sigma^2(X) = \frac{rs}{(r+s)^2(r+s+1)}.$$

La famille des lois bêta est souvent utilisée pour modéliser des proportions aléatoires, puisqu'une v.a. de loi bêta prend ses valeurs entre 0 et 1.

15. Soit X une v.a. de loi lognormale de paramètres (μ, σ^2). Montrer que

$$\mathrm{E}(\mathrm{X}^r) = e^{r\mu + \frac{1}{2}\sigma^2 r^2}$$

et déduire que $\mathrm{E}(\mathrm{X}) = e^{\mu + \frac{1}{2}\sigma^2}$ et que $\sigma_{\mathrm{X}}^2 = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$. (*Indication :* $\mathrm{E}(\mathrm{X}^r) = \int_0^\infty x^r f(x)dx$ où f est la densité lognormale ; faire le changement de variables $y = \log x - \mu$ pour obtenir

$$\mathrm{E}(\mathrm{X}^r) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{(r\mu + ry - y^2/2\sigma^2)} dy.)$$

16. On considère souvent une sous-famille des lois gamma : on dit qu'une v.a. X admet la *loi gamma standard* de paramètre α si elle admet la densité

$$f(x) = \begin{cases} \dfrac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)} & \text{si } x \geqslant 0, \\ 0 & \text{si } x < 0. \end{cases}$$

(Le second paramètre β est pris égal à 1 : rappelons que $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$.) Montrer que

$$\mathrm{E}(\mathrm{X}^k) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)} \quad (k \geqslant 0).$$

En déduire que X a la moyenne α et la variance α.

17.* Soit X une v.a. positive de moyenne μ et de variance σ^2 finies. Montrer que pour tout $b > 0$,

$$\mathrm{P}(\mathrm{X} \geqslant \mu + b\sigma) \leqslant \frac{1}{1+b^2}.$$

(*Indication :* Considérer la fonction $g(x) = \frac{\{(x-\mu)b+\sigma\}^2}{\sigma^2(1+b^2)^2}$ et établir que $\mathrm{E}(((\mathrm{X}-\mu)b + \sigma)^2) = \sigma^2(b^2+1)$.)

18. Soit X une v.a. de moyenne μ et de variance σ^2 finies. Montrer que

$$\mathrm{P}(\mu - d\sigma < \mathrm{X} < \mu + d\sigma) \geqslant 1 - \frac{1}{d^2}.$$

(Noter que cette inégalité n'a d'intérêt que si $d > 1$.)

19. Soit X une v.a. normale N(0,1). Montrer que $\mathrm{P}(\mathrm{X} > x) \leqslant \frac{1}{x\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, pour $x > 0$.

20. Soit X une v.a. de loi exponentielle. Montrer que $P(X > s + t \mid X > s) = P(X > t)$ pour $s > 0$, $t > 0$. Ceci est connu sous le nom de « propriété de non vieillissement » de la loi exponentielle.

21.* Soit X une v.a. strictement positive vérifiant $P(X > s+t \mid X > s) = P(X > t)$ pour tous $s > 0$, $t > 0$. Montrer que si $h(t) = P(X > t)$, alors h vérifie *l'équation de Cauchy* :

$$h(s + t) = h(s)h(t) \qquad (s > 0, t > 0)$$

et montrer que X admet une loi exponentielle (*Indication* : utiliser le fait que h est continue à droite, et que l'équation de Cauchy admet pour seules solutions continues à droite les exponentielles $h(t) = e^{at}$ pour $a \in \mathbb{R}$).

22. Soit α un entier strictement positif et supposons que la v.a. X admette la loi gamma de paramètres (α, β). Montrer que $P(X \leqslant x) = P(Y \geqslant \alpha)$, où Y suit une loi de Poisson de paramètre $\lambda = x\beta$. (Indication : On a $\Gamma(\alpha) = (\alpha - 1)!$; écrire $P(X \leqslant x)$, et utiliser une intégration par parties avec $u = t^{\alpha-1}$ et $dv = e^{-t\beta}dt$.)

23. *L'intensité* d'une v.a. strictement positive X est la fonction

$$h_X(t) = \lim_{\varepsilon \to 0} \frac{P(t \leqslant X < t + \varepsilon \mid X \geqslant t)}{\varepsilon}$$

lorsque la limite existe. Elle représente la probabilité pour qu'un objet de durée de vie X « meure » immédiatement après l'instant t, sachant qu'il est en vie à l'instant t. La propriété de non vieillissement des lois exponentielles implique que l'intensité soit constante. Une v.a. de loi de Weibull est aussi souvent utilisée pour modéliser les durées de vie. Montrer que

a) Si X est exponentielle (λ), son intensité est $h_X(t) = \lambda$;

b) Si X suit la loi de Weibull (α, β), son intensité est $h_X(t) = \alpha\beta^\alpha t^{\alpha-1}$.

24. Une v.a. strictement positive X admet la *loi logistique* si sa fonction de répartition est donnée par

$$F(x) = P(X \leqslant x) = \frac{1}{1 + e^{-(x-\mu)/\beta}},$$

avec les paramètres $\mu \in \mathbb{R}$ et $\beta > 0$.

a) Monter que si $\mu = 0$ et $\beta = 1$, alors X admet la densité

$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2};$$

b) Montrer que si les paramètres sont (μ, β), alors X admet l'intensité
$$h_X(t) = \frac{1}{\beta} F(t).$$

Chapitre 10

Variables aléatoires indépendantes

Rappelons que deux événements A et B sont indépendants si la connaissance du fait que B est réalisé ne change pas la probabilité de A, ce qui revient à dire que $P(A \cap B) = P(A)P(B)$. De même une famille quelconque $(A_i)_{i\in I}$ d'événements est indépendante si $P(\bigcap_{i\in J} A_i) = \prod_{i\in J} P(A_i)$ pour tout sous-ensemble fini J de I (définition 3.1).

De manière analogue, deux v.a. X et Y seront indépendantes si la connaissance de la valeur prise par Y ne modifie pas la loi de X, ce qui, grossièrement, revient à dire que les événements $\{Y \in A\}$ et $\{X \in B\}$ sont indépendants pour tout choix de A et B dans les tribus correspondantes. Cela s'exprime de manière plus naturelle et plus facile à mettre en œuvre mathématiquement sur les tribus engendrées par X et Y : si X est mesurable de $(\Omega, \mathcal{A})$ dans $(E, \mathcal{E})$, la *tribu engendrée par* X est $X^{-1}(\mathcal{E}) = \{X^{-1}(A) : A \in \mathcal{E}\}$.

Définition 10.1. (a) *Soit $(\mathcal{A}_i)_{i\in I}$ une famille de sous-tribus de $\mathcal{A}$. Les tribus $\mathcal{A}_i$ sont* indépendantes *si pour toute partie finie* J *de* I *et tous* $A_i \in \mathcal{A}_i$ *on a*

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

(b) *Soit $(X_i)_{i\in I}$ une famille de variables aléatoires, à valeurs respectives dans les $(E_i, \mathcal{E}_i)$. Les variables X_i sont* indépendantes *si les tribus engendrées $X_i^{-1}(\mathcal{E}_i)$ sont indépendantes.*

Dans la suite de ce chapitre nous considérons essentiellement des *couples* de v.a., de façon à garder des notations simples. Tous les résultats s'étendent de manière évidente aux familles *finies* de v.a.

Noter que X et Y prennent leurs valeurs dans $(E, \mathcal{E})$ et $(F, \mathcal{F})$ respectivement : les espaces E et F peuvent être différents.

Théorème 10.1. *Pour que* X *et* Y *soient indépendantes, il faut et il suffit que l'une des conditions équivalentes ci-dessous soit satisfaite :*

(a) $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ *pour tous* $A \in \mathcal{E}$, $B \in \mathcal{F}$;

(b) $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ *pour tous* $A \in \mathcal{C}$, $B \in \mathcal{D}$, *où* $\mathcal{C}$ *et* $\mathcal{D}$ *sont des classes de parties de* E *et* F *stables par intersection finie et engendrant les tribus* $\mathcal{E}$ *et* $\mathcal{F}$ *respectivement ;*

(c) $f(X)$ *et* $g(Y)$ *sont indépendantes pour tout couple* (f, g) *de fonctions mesurables réelles sur* E *et* F *respectivement ;*

(d) $E(f(X)g(Y)) = E(f(X))E(g(Y))$ *pour tout couple* (f, g) *de fonctions mesurables réelles, bornées (ou positives) sur* E *et* F *respectivement ;*

(e) *Si de plus* E *et* F *sont des espaces métriques avec leurs tribus boréliennes respectives* $\mathcal{E}$ *et* $\mathcal{F}$, *alors* $E(f(X)g(Y)) = E(f(X))E(g(Y))$ *pour tout couple* (f, g) *de fonctions continues réelles bornées sur* E *et* F *respectivement.*

Preuve. Comme $X^{-1}(\mathcal{E})$ est l'ensemble des $\{X \in A\}$ pour les $A \in \mathcal{E}$, et de même pour $Y^{-1}(\mathcal{F})$, il est clair que (a) est juste une autre manière d'écrire la définition de l'indépendance de X et Y.

(a)⇒(b) : Implication triviale.

(b)⇒(a) . La classe $\mathcal{C}_B$ des ensembles $A \in \mathcal{E}$ vérifiant $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ pour un $B \in \mathcal{D}$ donné est stable par limites croissantes et par différence, et contient $\mathcal{C}$ par hypothèse, tandis que $\mathcal{C}$ est stable par intersection. Le théorème des classes monotones 6.2 entraîne alors que $\mathcal{C}_B = \mathcal{E}$. En d'autres termes on a (b) avec $\mathcal{C} = \mathcal{E}$. De manière analogue (en fixant $A \in \mathcal{E}$) on obtient qu'on a (b) avec de plus $\mathcal{D} = \mathcal{F}$: en d'autres termes, on a (a).

(c)⇒(a) : Il suffit de prendre $f(x) = x$ et $g(y) = y$ (les applications identité sur E et F respectivement).

(a)⇒(c) : Étant données f et g, on voit que

$$f(X)^{-1}(\mathcal{E}) = X^{-1}(f^{-1}(\mathcal{E})) \subset X^{-1}(\mathcal{E}).$$

et de même $g(Y)^{-1}(\mathcal{F}) \subset Y^{-1}(\mathcal{F})$: le résultat est alors évident.

(d)⇒(a) : Prendre $f(x) = 1_A(x)$ et $g(x) = 1_B(x)$.

(a)⇒(d) : Par hypothèse on a (d) pour f et g fonctions indicatrices, et donc pour f et g fonctions simples (i.e. $f(x) = \sum_{i=1}^{k} a_i 1_{A_i}(x)$) par linéarité. Si f et g sont positives on peut trouver des suites (f_n) et (g_n) de fonctions simples positives croissant vers f et g respectivement. Noter que les produits $f_n(X)g_n(Y)$ croissent vers $f(X)g(Y)$. En appliquant le théorème de convergence monotone, il vient alors

$$\begin{aligned} E(f(X)g(Y)) &= E\left(\lim_{n\to\infty} f_n(X)g_n(Y)\right) \\ &= \lim_{n\to\infty} E(f_n(X)g_n(Y)) \end{aligned}$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \mathrm{E}(f_n(\mathrm{X}))\mathrm{E}(g_n(\mathrm{Y}))\\
&= \lim_{n\to\infty} \mathrm{E}(f_n(\mathrm{X})) \lim_{n\to\infty} \mathrm{E}(g_n(\mathrm{Y}))\\
&= \mathrm{E}(f(\mathrm{X}))\mathrm{E}(g(\mathrm{Y})).
\end{aligned}
$$

On a donc le résultat lorsque f et g sont positives. Lorsque f et g sont bornées on écrit $f = f^+ - f^-$ et $g = g^+ - g^-$ et on conclut par linéarité.

(d)⇒(e) : Évident.

(e)⇒(b) : Il suffit de montrer (b) lorsque $\mathcal{C}$ et $\mathcal{D}$ sont les classes de tous les fermés de E et F (ces classes sont stables par intersection). Soit par exemple A un fermé de E ; si $f_n(x) = \inf(1, nd(x, \mathrm{A}))$, où $d(x, \mathrm{A})$ désigne la distance du point x au fermé A, alors f_n est continue, vérifie $0 \leqslant f_n \leqslant 1$ et décroît vers l'indicatrice 1_{A}. Pour B fermé de F soit de même une suite (g_n) de fonctions continues avec $0 \leqslant g_n \leqslant 1$ et convergeant vers 1_{B}. On peut reproduire la fin de la preuve de l'implication (a)⇒(d) en remplaçant le théorème de convergence monotone par celui de convergence dominée, pour obtenir (b). ■

Exemple. Soit E et F finis ou dénombrables. Alors X et Y sont indépendantes si et seulement si $\mathrm{P}(\mathrm{X} = i, \mathrm{Y} = j) = \mathrm{P}(\mathrm{X} = i)\mathrm{P}(\mathrm{Y} = j)$ pour tous $i \in \mathrm{E}$, $j \in \mathrm{F}$.

Nous présenterons d'autres exemples dans le chapitre 12.

Nous allons maintenant introduire les fonctions mesurables sur un produit d'espaces. En général si $\mathcal{E}$ et $\mathcal{F}$ sont deux tribus sur les espaces E et F respectivement, le produit cartésien $\mathcal{E} \times \mathcal{F} = \{\mathrm{A} \subset \mathrm{E} \times \mathrm{F} : \mathrm{A} = \Lambda \times \Gamma, \Lambda \subset \mathcal{E} \text{ et } \Gamma \in \mathcal{F}\}$ n'est *pas* une tribu. La plus petite tribu de $\mathrm{E} \times \mathrm{F}$ contenant $\mathcal{E} \times \mathcal{F}$, soit $\sigma(\mathcal{E} \times \mathcal{F})$, est très fréquemment rencontrée et est notée ainsi :

$$\mathcal{E} \otimes \mathcal{F} = \sigma(\mathcal{E} \times \mathcal{F}).$$

Théorème 10.2. *Soit f mesurable : $(\mathrm{E} \times \mathrm{F}, \mathcal{E} \otimes \mathcal{F}) \longmapsto (\mathbb{R}, \mathcal{R})$. Pour tout $x \in \mathrm{E}$ la « section » $y \mapsto f(x, y)$ est une fonction $\mathcal{F}$-mesurable ; de même pour tout $y \in \mathrm{F}$ la fonction $x \mapsto f(x, y)$ est $\mathcal{E}$-mesurable.*

Note. La réciproque du théorème 10.2 est fausse en général.

Preuve. Supposons d'abord f de la forme $f(x, y) = 1_{\mathrm{C}}(x, y)$, pour $\mathrm{C} \in \mathcal{E} \otimes \mathcal{F}$. Soit $\mathcal{H} = \{\mathrm{C} \in \mathcal{E} \otimes \mathcal{F} : y \mapsto 1_{\mathrm{C}}(x, y)$ est $\mathcal{F}$-mesurable pour chaque $x \in \mathrm{E}$ fixé$\}$. Alors $\mathcal{H}$ est une tribu (vérification immédiate) contenant $\mathcal{E} \times \mathcal{F}$, donc $\sigma(\mathcal{E} \times \mathcal{F}) \subset \mathcal{H}$. Comme par construction $\mathcal{H} \subset \sigma(\mathcal{E} \times \mathcal{F})$, il vient $\mathcal{H} = \mathcal{E} \otimes \mathcal{F}$. On a donc montré le résultat pour les fonctions indicatrices, donc par linéarité pour les fonctions mesurables simples. Par limite croissante (toute fonction mesurable positive étant limite croissante de fonctions mesurables simples, et toute limite croissante de fonctions mesurables étant mesurable), on l'obtient aussi pour les fonctions mesurables positives. Enfin par différence (i.e. en écrivant $f = f^+ - f^-$) on l'obtient pour les fonctions mesurables quelconques. ∎

Théorème 10.3. (Tonelli-Fubini.) *Soit* P *et* Q *des probabilités sur* $(\mathrm{E}, \mathcal{E})$ *et* $(\mathrm{F}, \mathcal{F})$ *respectivement.*

(a) *Posons* $\mathrm{R}(\mathrm{A} \times \mathrm{B}) = \mathrm{P}(\mathrm{A})\mathrm{Q}(\mathrm{B})$ *pour* $\mathrm{A} \in \mathcal{E}$ *et* $\mathrm{B} \in \mathcal{F}$. *Alors* R *s'étend de manière unique en une probabilité sur* $(\mathrm{E} \times \mathrm{F}, \mathcal{E} \otimes \mathcal{F})$, *notée* $\mathrm{P} \otimes \mathrm{Q}$.

(b) *Pour toute fonction réelle* f *sur* $\mathrm{E} \times \mathrm{F}$ *qui est* $\mathcal{E} \otimes \mathcal{F}$*-mesurable, et positive ou intégrable par rapport à* $\mathrm{P} \otimes \mathrm{Q}$, *la fonction* $x \to \int f(x, y)\mathrm{Q}(dy)$ *est* $\mathcal{E}$*-mesurable, la fonction* $y \to \int f(x, y)\mathrm{P}(dx)$ *est* $\mathcal{F}$*-mesurable, et on a*

$$\int f\, d(\mathrm{P} \otimes \mathrm{Q}) = \int \left\{ \int f(x, y)\mathrm{Q}(dy) \right\} \mathrm{P}(dx)$$
$$= \int \left\{ \int f(x, y)\mathrm{P}(dx) \right\} \mathrm{Q}(dy).$$

Preuve. (a) Soit $\mathrm{C} \in \mathcal{E} \otimes \mathcal{F}$, et écrivons $\mathrm{C}(x) = \{y : (x, y) \in \mathrm{C}\}$. Pour $\mathrm{C} = \mathrm{A} \times \mathrm{B}$, on a $\mathrm{C}(x) = \mathrm{B}$ si $x \in \mathrm{A}$ et $\mathrm{C}(x) = \emptyset$ sinon, donc

$$\mathrm{R}(\mathrm{C}) = \mathrm{P} \otimes \mathrm{Q}(\mathrm{C}) = \mathrm{P}(\mathrm{A})\mathrm{Q}(\mathrm{B}) = \int \mathrm{P}(dx)\mathrm{Q}[\mathrm{C}(x)].$$

Soit $\mathcal{H} = \{\mathrm{C} \in \mathcal{E} \otimes \mathcal{F} : x \to \mathrm{Q}[\mathrm{C}(x)] \text{ est } \mathcal{E}\text{-mesurable}\}$. Clairement $\mathcal{H}$ est stable par limite croissante et par différence, tandis que $\mathcal{E} \times \mathcal{F} \subset \mathcal{H} \subset \mathcal{E} \otimes \mathcal{F}$, donc d'après le théorème des classes monotones on a $\mathcal{H} = \mathcal{E} \otimes \mathcal{F}$. Pour chaque $\mathrm{C} \in \mathcal{H} = \mathcal{E} \otimes \mathcal{F}$, et comme $x \mapsto \mathrm{Q}[\mathrm{C}(x)]$ est mesurable et positive, on peut alors poser

$$\mathrm{R}(\mathrm{C}) = \int \mathrm{P}(dx)\mathrm{Q}[\mathrm{C}(x)].$$

Nous devons montrer que R est une probabilité. On a d'abord

$$\mathrm{R}(\Omega) = \mathrm{R}(\mathrm{E} \times \mathrm{F}) = \int_{\mathrm{E}} \mathrm{P}(dx)\mathrm{Q}[\mathrm{F}] = 1.$$

Soit $\mathrm{C}_n \in \mathcal{E} \otimes \mathcal{F}$ des ensembles deux à deux disjoints, et $\mathrm{C} = \bigcup_{n=1}^{\infty} \mathrm{C}_n$. Comme Q est une probabilité et que les $\mathrm{C}_n(x)$ sont aussi deux à deux disjoints, on a $Q[C(x)] = \sum_{n=1}^{\infty} Q[C_n(x)]$. Appliquons le théorème 9.2 à P et aux fonctions $f_n(x) = Q[C_n(x)]$:

$$\begin{aligned}\sum_{n=1}^{\infty} R(C_n) &= \sum_{n=1}^{\infty} \int f_n \, dP \\ &= \int \Big(\sum_{n=1}^{\infty} f_n\Big) dP \\ &= \int P(dx) Q[C(x)] = R(C).\end{aligned}$$

Donc R est une probabilité. L'unicité de R provient du corollaire 6.1.

(b) Nous avons déjà établi le résultat pour f de la forme $f(x,y) = 1_C(x,y)$, avec $C \in \mathcal{E} \otimes \mathcal{F}$. Il est vrai par linéarité pour les fonctions simples. Si f est positive et $\mathcal{E} \otimes \mathcal{F}$-mesurable, soit f_n des fonctions mesurables croissant vers f. On a

$$\begin{aligned}\int f \, dR &= \lim_{n \to \infty} \int f_n \, dR = \lim_{n \to \infty} \int \left\{ \int f_n(x,y) Q(dy) \right\} P(dx) \\ &= \int \left\{ \lim_{n \to \infty} \int f_n(x,y) Q(dy) \right\} P(dx) \\ &= \int \left\{ \int \lim_{n \to \infty} f_n(x,y) Q(dy) \right\} P(dx) = \int \left\{ \int f(x,y) Q(dy) \right\} P(dx).\end{aligned}$$

Un argument similaire montre que l'expression ci-dessus égale aussi

$$= \int \left\{ \int f(x,y) P(dx) \right\} Q(dy).$$

Finalement pour une fonction f de signe arbitraire (intégrable par rapport à R) il suffit de décomposer en $f = f^+ - f^-$. ■

Corollaire 10.1. *Soit X et Y deux v.a. sur $(\Omega, \mathcal{A}, P)$, à valeurs respectives dans $(E, \mathcal{E})$ et $(F, \mathcal{F})$. Le couple $Z = (X, Y)$ peut être considéré comme une v.a. à valeurs dans $(E \times F, \mathcal{E} \otimes \mathcal{F})$, et les deux v.a. X, Y sont indépendantes si et seulement si la loi $P_{(X,Y)}$ du couple égale le produit $P_X \otimes P_Y$ des lois de X et Y.*

Preuve. Il est évident que $Z^{-1}(A \times B) = X^{-1}(A) \cap Y^{-1}(B) \in \mathcal{A}$ si $A \in \mathcal{E}$ et $B \in \mathcal{F}$, de sorte que la mesurabilité de Z découle de la définition de la tribu $\mathcal{E} \otimes \mathcal{F}$ et du théorème 8.1.

L'indépendance de X et Y revient au fait que pour tous $A \in \mathcal{E}$ et $B \in \mathcal{F}$ on ait

$$P((X, Y) \in A \times B) = P(X \in A)P(Y \in B),$$

ce qui équivant à

$$P_{(X,Y)}(A \times B) = P_X(A)P_Y(B),$$

de sorte que l'unicité de la probabilité produit dans le théorème de Fubini nous donne la seconde assertion. ■

Nous faisons maintenant une légère digression en *construisant un modèle pour des variables aléatoires indépendantes*, ce qui permet en particulier de vérifier que cette notion n'est pas vide.

Soit d'abord μ une probabilité sur un espace mesurable $(E, \mathcal{E})$. Il est facile de construire une v.a. X à valeurs dans E et de loi μ : prendre simplement $\Omega = E$, $\mathcal{A} = \mathcal{E}$, $P = \mu$, et prendre pour X l'identité ($X(x) = x$).

Un peu plus compliquée est la construction d'un couple de deux v.a. indépendantes X et Y, à valeurs dans $(E, \mathcal{E})$ et $(F, \mathcal{F})$, et de lois respectives μ et ν (probabilités données *a priori* sur E et F) : on peut prendre $\Omega = E \times F$, $\mathcal{A} = \mathcal{E} \otimes \mathcal{F}$, $P = \mu \otimes \nu$, et $X(x, y) = x$ et $Y(x, y) = y$, où $(x, y) \in E \times F$.

Il est malheureusement bien plus compliqué, mais indispensable pour les applications (la loi des grands nombres par exemple qu'on verra plus tard, ou même l'étude d'une suite infinie de jets de dés, etc.), de construire une *suite infinie de variables indépendantes* de lois données. Plus précisément, pour chaque entier n on se donne une v.a. X_n définie sur un espace de probabilité $(\Omega_n, \mathcal{A}_n P_n)$, à valeurs dans $(E_n, \mathcal{E}_n)$ et de loi μ_n (pour construire chaque X_n on peut opérer comme ci-dessus). Ensuite, on pose

$$\Omega = \prod_{n=1}^{\infty} \Omega_n \qquad \text{(produit cartésien dénombrable)}$$

$$\mathcal{A} = \bigotimes_{n=1}^{\infty} \mathcal{A}_n$$

où $\otimes_{n=1}^{\infty} \mathcal{A}_n$ désigne la plus petite tribu de Ω à laquelle appartiennent tous les ensembles suivants :

$$A_1 \times A_2 \times \cdots \times A_k \times \Omega_{k+1} \times \Omega_{k+2} \times \cdots, \qquad A_i \in \mathcal{A}_i, \qquad k = 1, 2, 3, \ldots$$

Le théorème suivant constitue un résultat non trivial de la théorie de la mesure, et généralise le théorème de Fubini. Nous l'énonçons sans démonstration.

Théorème 10.4. *Avec les notations ci-dessus, il existe une probabilité* P *et une seule sur* $(\Omega, \mathcal{A})$, *telle que*

$$\mathrm{P}(\mathrm{A}_1 \times \cdots \times \mathrm{A}_k \times \Omega_{k+1} \times \cdots) = \prod_{i=1}^{k} \mathrm{P}_i(\mathrm{A}_i)$$

pour tous $k = 1, 2, \ldots$ *et* $\mathrm{A}_i \in \mathcal{A}_i$.

Maintenant, on note $\widetilde{\mathrm{X}}_n$ *l'extension naturelle* de X_n à Ω, c'est-à-dire que pour chaque $\omega = (\omega_1, \omega_2, \ldots, \omega_n, \ldots) \in \Omega$ avec $\omega_i \in \Omega_i$ on pose

$$\widetilde{\mathrm{X}}_n(\omega) = \mathrm{X}_n(\omega_n).$$

Corollaire 10.2. *Les variables* $\widetilde{\mathrm{X}}_n$ *ci-dessus sont indépendantes, et chaque* $\widetilde{\mathrm{X}}_n$ *a la loi* μ_n.

Preuve. Soit $\mathrm{B}_n \in \mathcal{E}_n$ On a

$$\widetilde{\mathrm{X}}_n^{-1}(\mathrm{B}_n) = \Omega_1 \times \cdots \times \Omega_{n-1} \times \mathrm{X}_n^{-1}(\mathrm{B}_n) \times \Omega_{n+1} \times \Omega_{n+2} \times \cdots,$$

et le théorème 10.4 donne pour $k = 1, 2, \ldots$:

$$\mathrm{P}\left(\bigcap_{n=1}^{k} \widetilde{\mathrm{X}}_n^{-1}(\mathrm{B}_n)\right) = \mathrm{P}\left(\mathrm{X}_1^{-1}(\mathrm{B}_1) \times \cdots \times \mathrm{X}_k^{-1}(\mathrm{B}_k) \times \Omega_{k+1} \times \cdots\right)$$
$$= \prod_{n=1}^{k} \mathrm{P}_n(\mathrm{X}_n \in \mathrm{B}_n) = \prod_{n=1}^{k} \mu_n(\mathrm{B}_n).$$

En particulier $\mathrm{P}(\widetilde{\mathrm{X}}_n \in \mathrm{B}_n) = \mu_n(\mathrm{B}_n)$, et on a les résultats souhaités. ■

Nous allons maintenant discuter quelques propriétés importantes de l'indépendance. Soit d'abord une suite (A_n) d'événements. On rappelle la définition de l'événement $\limsup_{n\to\infty} \mathrm{A}_n$ donnée en (2.2) :

$$\limsup_{n\to\infty} \mathrm{A}_n = \bigcap_{n=1}^{\infty}\left(\bigcup_{m\geqslant n} \mathrm{A}_m\right) = \lim_{n\to\infty}\left(\bigcup_{m\geqslant n} \mathrm{A}_m\right).$$

Théorème 10.5. (Lemme de Borel-Cantelli.) *Soit* (A_n) *une suite d'événements sur* $(\Omega, \mathcal{A}, \mathrm{P})$.

(a) *Si* $\sum_{n=1}^{\infty} P(A_n) < \infty$, *alors* $P(\limsup_{n\to\infty} A_n) = 0$.

(b) *Si* $P(\limsup_{n\to\infty} A_n) = 0$ *et si les événements* A_n *sont indépendants, alors on a* $\sum_{n=1}^{\infty} P(A_n) < \infty$.

Remarque. Nous verrons plus bas (comme cas particulier de la « loi zéro-un ») que si les A_n sont indépendants, $P(\limsup_{n\to\infty} A_n)$ ne peut être égal que à 0 ou à 1. Par suite si la série $\sum_{n=1}^{\infty} P(A_n)$ converge on a $P(\limsup_{n\to\infty} A_n) = 0$, et si elle diverge on a $P(\limsup_{n\to\infty} A_n) = 1$.

Preuve. (a) Soit $a_n = P(A_n) = E(1_{A_n})$. Grâce au théorème 9.2(b), $\sum_{n=1}^{\infty} a_n < \infty$ implique $\sum_{n=1}^{\infty} 1_{A_n} < \infty$ p.s., tandis que $\sum_{n=1}^{\infty} 1_{A_n}(\omega) = \infty$ si et seulement si $\omega \in \limsup_{n\to\infty} A_n$: donc on a (a).

(b) Supposons les A_n indépendants. On a

$$
\begin{aligned}
P(\limsup_{n\to\infty} A_n) &= \lim_{n\to\infty}\lim_{k\to\infty} P\left(\bigcup_{m=n}^{k} A_m\right)\\
&= \lim_{n\to\infty}\lim_{k\to\infty}\left(1 - P\left(\bigcap_{m=n}^{k} A_m^c\right)\right)\\
&= 1 - \lim_{n\to\infty}\lim_{k\to\infty}\left(\prod_{m=n}^{k}(1 - P(A_m))\right) \quad \text{(par indépendance)}\\
&= 1 - \lim_{n\to\infty}\lim_{k\to\infty}\prod_{m=n}^{k}(1 - a_m).
\end{aligned}
$$

Par hypothèse $P(\limsup_{n\to\infty} A_n) = 0$, donc $\lim_{n\to\infty}\lim_{k\to\infty}\prod_{m=n}^{k}(1 - a_m) = 1$ et en prenant les logarithmes on arrive à

$$\lim_{n\to\infty}\lim_{k\to\infty}\sum_{m=n}^{k}\log(1 - a_m) = 0,$$

donc

$$\lim_{n\to\infty}\sum_{m\geqslant n}\log(1 - a_m) = 0,$$

ce qui signifie que la série $\sum_m \log(1 - a_m)$ converge. Comme $|\log(1 - x)| \geqslant x$ pour $0 < x < 1$, la série $\sum_m a_m$ converge également. ∎

Soit maintenant des v.a. X_n définies sur $(\Omega, \mathcal{A}, P)$. Soit les tribus

$$\mathcal{B}_n = \sigma(\mathrm{X}_n)$$

$$\mathcal{C}_n = \sigma\left(\bigcup_{p\geqslant n} \mathcal{B}_p\right)$$

$$\mathcal{C}_\infty = \bigcap_{n=1}^{\infty} \mathcal{C}_n$$

$\mathcal{C}_\infty$ est appelée la *tribu asymptotique*.Un événement est dans cette tribu s'il ne dépend que du comportement de la suite X_n lorsque $n \to \infty$ (d'où son nom) ; nous donnons des exemples de v.a. $\mathcal{C}_\infty$-mesurables juste après le théorème suivant.

Théorème 10.6. (Loi zéro-un.) *Soit des v.a. indépendantes définies sur un espace de probabilité* $(\Omega, \mathcal{A}, \mathrm{P})$*, et soit* $\mathcal{C}_\infty$ *la tribu asymptotique associée. Pour tout* $\mathrm{C} \in \mathcal{C}_\infty$ *on a* $\mathrm{P(C)} = 0$ *ou* $\mathrm{P(C)} = 1$.

Preuve. Soit $\mathcal{D}_n = \sigma(\bigcup_{p<n} \mathcal{B}_p)$. Par hypothèse les tribus $\mathcal{C}_n$ et $\mathcal{D}_n$ sont indépendantes, donc si $\mathrm{A} \in \mathcal{C}_n$ et $\mathrm{B} \in \mathcal{D}_n$ on a

$$\mathrm{P(A \cap B) = P(A)P(B)}. \tag{10.1}$$

Si $\mathrm{A} \in \mathcal{C}_\infty$ nous avons donc (10.1) pour tout $\mathrm{B} \in \bigcup \mathcal{D}_n$, donc aussi pour $\mathrm{B} \in \mathcal{D} = \sigma(\bigcup \mathcal{D}_n)$ par le théorème des classes monotones 6.2. Cependant $\mathcal{C}_\infty \subset \mathcal{D}$, donc on a (10.1) pour $\mathrm{B} = \mathrm{A} \in \mathcal{C}_\infty$, ce qui implique que $\mathrm{P(A) = P(A)P(A) = P(A)^2}$, donc $\mathrm{P(A)} = 0$ ou $\mathrm{P(A)} = 1$. ■

Conséquence. Si les v.a. $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n, \ldots$ sont indépendantes, on a

1. $\{\omega : \lim_{n\to\infty} \mathrm{X}_n(\omega) \text{ existe}\} \in \mathcal{C}_\infty$; donc soit la suite X_n converge p.s., soit elle diverge p.s.
2. Chaque v.a. qui est $\mathcal{C}_\infty$-mesurable est p.s. constante. En particulier,

$$\limsup_{n\to\infty} \mathrm{X}_n, \qquad \liminf_{n\to\infty} \mathrm{X}_n,$$

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{p\leqslant n} \mathrm{X}_p, \qquad \liminf_{n\to\infty} \frac{1}{n}\sum_{p\leqslant n} \mathrm{X}_p$$

sont toutes p.s. constantes.

Exercices

1. Soit $f = (f_1, f_2) : \Omega \to \mathrm{E} \times \mathrm{F}$. Montrer que $f : (\Omega, \mathcal{A}) \to (\mathrm{E} \times \mathrm{F}, \mathcal{E} \otimes \mathcal{F})$ est mesurable si et seulement si f_1 est mesurable de $(\Omega, \mathcal{A})$ dans $(\mathrm{E}, \mathcal{E})$ et f_2 est mesurable de $(\Omega, \mathcal{A})$ dans $(\mathrm{F}, \mathcal{F})$.
2. Soit $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, et soit $\mathcal{B}^2$ la tribu borélienne de $\mathbb{R}^2$, tandis que $\mathcal{B}$ désigne la tribu borélienne de $\mathbb{R}$. Montrer que $\mathcal{B}^2 = \mathcal{B} \otimes \mathcal{B}$.
3. Soit $\Omega = [0,1]$ et $\mathcal{A}$ la tribu borélienne de Ω, et soit $\mathrm{P}(\mathrm{A}) = \int 1_{\mathrm{A}}(x)dx$ pour $\mathrm{A} \in \mathcal{A}$. Soit $\mathrm{X}(x) = x$. Montrer que X admet la loi uniforme sur $[0,1]$.
4. Soit Ω=$\mathbb{R}$ et $\mathcal{A}$=$\mathcal{B}$. Soit P donnée par $\mathrm{P}(\mathrm{A})$=$\frac{1}{\sqrt{2\pi}} \int 1_{\mathrm{A}}(x) e^{-x^2/2} dx$, et $\mathrm{X}(x) = x$. Montrer que X admet la loi normale $\mathrm{N}(0,1)$.
5. Construire un exemple montrant que $\mathrm{E}(\mathrm{XY}) = \mathrm{E}(\mathrm{X})\mathrm{E}(\mathrm{Y})$ n'implique pas en général que X et Y soient indépendantes (on suppose que X, Y et XY sont intégrables).
6. Soit X et Y des v.a. indépendantes à valeurs dans $\mathbb{N}$, avec
$$\mathrm{P}(\mathrm{X} = i) = \mathrm{P}(\mathrm{Y} = i) = \frac{1}{2^i} \qquad (i = 1, 2, \dots).$$
Trouver les probabilités des événements suivants :
   a) $\mathrm{P}(\min(\mathrm{X}, \mathrm{Y}) \leqslant i)$ [Rép. : $1 - \frac{1}{4^i}$.]
   b) $\mathrm{P}(\mathrm{X} = \mathrm{Y})$ [Rép. : $\frac{1}{3}$.]
   c) $\mathrm{P}(\mathrm{Y} > \mathrm{X})$ [Rép. : $\sum_{i \geqslant 0} \frac{1}{2^i(2^i - 1)}$.]
   d) $\mathrm{P}(\mathrm{X} \text{ divise } \mathrm{Y})$ [Rép. : $\frac{1}{3}$.]
   e) $\mathrm{P}(\mathrm{X} \geqslant k\mathrm{Y})$ pour un entier $k \geqslant 1$ [Rép. : $\frac{1}{2^{1+k} - 1}$.]
7. Soit X et Y deux v.a. indépendantes de loi géométrique de paramètres respectifs λ et μ. Soit $\mathrm{Z} = \min(\mathrm{X}, \mathrm{Y})$. Montrer que Z suit une loi géométrique et trouver son paramètre. [Rép. : $\lambda\mu$.]
8. Soit X et Y dans L^2. On définit la *covariance* de X et Y par
$$\mathrm{Cov}(\mathrm{X}, \mathrm{Y}) = \mathrm{E}((\mathrm{X} - \mu)(\mathrm{Y} - \nu))$$
où $\mathrm{E}(\mathrm{X}) = \mu$ et $\mathrm{E}(\mathrm{Y}) = \nu$. Montrer que
$$\mathrm{Cov}(\mathrm{X}, \mathrm{Y}) = \mathrm{E}(\mathrm{XY}) - \mu\nu$$
et que, si de plus X et Y sont indépendantes, alors $\mathrm{Cov}(\mathrm{X}, \mathrm{Y}) = 0$.

9. Soit X et Y dans L^1. Si X et Y sont indépendantes, montrer que $XY \in L^1$. Donner un exemple montrant que XY n'est pas nécessairement dans L^1 en général (i.e., si X et Y ne sont pas indépendantes).

10.* Soit n un entier premier plus grand que 2. Soit X et Y deux v.a. indépendantes et uniformément distribuées sur l'ensemble fini $\{0, 1, \dots, n-1\}$ (i.e., $P(X = i) = P(Y = i) = \frac{1}{n}$ pour $i = 0, 1, \dots, n-1$). Pour chaque entier r avec $0 \leqslant r \leqslant n-1$, soit $Z_r = X + rY \pmod n$.

   a) Montrer que les v.a. $\{Z_r : r = 0, \dots, n-1\}$ sont deux à deux indépendantes (i.e. Z_r et Z_s sont indépendantes si $r \neq s$).

   b) Le même résultat est-il encore vrai si l'entier n n'est pas premier ? [Rép. : Non.]

11. Soit X et Y deux v.a. indépendantes de loi $P(X = 1) = P(Y = 1) = \frac{1}{2}$ et $P(X = -1) = P(Y = -1) = \frac{1}{2}$. Soit $Z = XY$. Montrer que X, Y, Z sont indépendantes deux à deux, mais qu'elles ne sont pas globalement indépendantes.

12. Soit (A_n) une suite d'événements. Montrer que

$$P(\limsup_{n\to\infty} A_n) \geqslant \limsup_{n\to\infty} P(A_n).$$

13. Une suite (X_n) de v.a. est dite *complètement convergente* vers X si

$$\sum_{n=1}^{\infty} P(|X_n - X| > \varepsilon) < \infty \text{ pour tout } \varepsilon > 0.$$

Montrer que si la suite (X_n) est indépendante, la convergence complète équivaut à la convergence p.s.

En vue des deux exercices suivants, nous introduisons la notion de « mesure » (positive) : une mesure μ sur un espace mesurable $(E, \mathcal{E})$ est une application de $\mathcal{E}$ dans $[0, \infty]$ qui vérifie l'axiome de σ-additivité (2.) de la définition 2.3. Cette mesure est dite *finie* si $\mu(E) < \infty$ (une probabilité est donc une mesure finie), elle est dite *σ-finie* s'il existe une suite $(A_n)_{n\geqslant 1}$ avec $A_n \in \mathcal{E}$ et $\cup_{n\geqslant 1} A_n = E$ et $\mu(A_n) < \infty$ pour tout n (un exemple — fondamental — de mesure σ-finie sur $\mathbb{R}$ est la mesure de Lebesgue, définie dans le chapitre suivant).

14. Soit μ et ν deux mesures finies sur $(E, \mathcal{E})$ et $(F, \mathcal{F})$ respectivement. On définit $\lambda = \mu \otimes \nu$ sur $(E \times F, \mathcal{E} \times \mathcal{F})$ par $\lambda(A \times B) = \mu(A)\nu(B)$ pour $A \in \mathcal{E}$ et $B \in \mathcal{F}$.

a) Montrer que λ s'étend de manière unique en une mesure finie sur $\mathcal{E} \otimes \mathcal{F}$;

b) Soit $f : \mathrm{E} \times \mathrm{F} \to \mathbb{R}$ mesurable relativement à $\mathcal{E} \otimes \mathcal{F}$. Montrer le *théorème de Fubini* : si f est λ-intégrable, alors $x \to \int f(x, y)\nu(dy)$ et $y \to \int f(x, y)\mu(dx)$ sont respectivement $\mathcal{E}$ et $\mathcal{F}$ mesurables, et de plus

$$\int f\,d\lambda = \int \left(\int f(x, y)\mu(dx) \right) \nu(dy) = \int \left(\int f(x, y)\nu(dy) \right) \mu(dx).$$

(*Indication :* Utiliser le théorème 10.3.)

15.* Montrer que si μ et ν sont supposées σ-finies, et en supposant que $\lambda = \mu \otimes \nu$ existe (même définition que dans l'exercice précédent), alors

a) $\lambda = \mu \otimes \nu$ est σ-finie ;

b) (*Théorème de Fubini.*) Si $f : \mathrm{E} \times \mathrm{F} \to \mathbb{R}$ est mesurable relativement à $\mathcal{E} \otimes \mathcal{F}$ et est λ-intégrable, alors $x \to \int f(x, y)\nu(dy)$ et $y \to \int f(x, y)\mu(dx)$ sont respectivement $\mathcal{E}$ et $\mathcal{F}$ mesurables, et de plus

$$\int f\,d\lambda = \int \left(\int f(x, y)\mu(dx) \right) \nu(dy) = \int \left(\int f(x, y)\nu(dy) \right) \mu(dx).$$

(*Indication :* Utiliser l'exercice 14 sur les ensembles $\mathrm{E}_j \times \mathrm{F}_k$, où $\mu(\mathrm{E}_j) < \infty$ et $\nu(\mathrm{F}_k) < \infty$.)

16.* On jette indéfiniment une pièce de monnaie, avec P(Face)$= p$. Soit A_k l'événement selon lequel au moins k faces consécutifs apparaissent au cours des jets numérotés $2^k, 2^k + 1, \ldots, 2^{k+1} - 1$. Montrer que $\mathrm{P}(\limsup_{k\to\infty} \mathrm{A}_k)$ vaut 1 ou 0 selon que $p \geqslant \frac{1}{2}$ ou que $p < \frac{1}{2}$.

17. Soit $\mathrm{X}_0, \mathrm{X}_1, \mathrm{X}_2, \ldots$ une suite de v.a. indépendantes telles que $\mathrm{P}(\mathrm{X}_n = 1) = \mathrm{P}(\mathrm{X}_n = -1) = \frac{1}{2}$. Soit $\mathrm{Z}_n = \Pi_{i=0}^n \mathrm{X}_i$. Montrer que $\mathrm{Z}_1, \mathrm{Z}_2, \mathrm{Z}_3, \ldots$ sont indépendantes.

18. Soit X et Y des v.a. indépendantes, et supposons que $\mathrm{P}(\mathrm{X} + \mathrm{Y} = \alpha) = 1$, où α est une constante. Montrer que X et Y sont p.s. des constantes.

Chapitre 11
Lois de probabilité sur $\mathbb{R}$

On a déjà vu qu'une probabilité sur $(\mathbb{R}, \mathcal{B})$ ($\mathcal{B}$ est la tribu borélienne) est caractérisée par sa fonction de répartition

$$\mathrm{F}(x) = \mathbf{P}(]-\infty, x]).$$

Nous allons maintenant utiliser les outils développés précédemment pour étudier la *mesure de Lebesgue* sur $\mathbb{R}$.

Définition 11.1. *La* mesure de Lebesgue *est une application $m : \mathcal{B} \to [0, \infty]$ qui vérifie les propriétés :*

(i) *(σ-additivité) si $\mathrm{A}_1, \mathrm{A}_2, \mathrm{A}_3, \ldots$ sont des boréliens deux à deux disjoints, alors*
$$m\left(\bigcup_{i=1}^{\infty} \mathrm{A}_i\right) = \sum_{i=1}^{\infty} m(\mathrm{A}_i)\ ;$$

(ii) *si $a, b \in \mathbb{R}$, $a < b$, alors $m(]a, b]) = b - a$.*

Comme $\emptyset \subset]a, b]$ on a alors $m(\emptyset) \leqslant b-a$ pour tous $a < b$, donc $m(\emptyset) = 0$.

Théorème 11.1. *La mesure de Lebesgue est unique.*

Preuve. On suppose que m existe. Fixons $a < b$ dans $\mathbb{R}$, et posons

$$m_{a,b}(\mathrm{A}) = \frac{m(\mathrm{A} \cap]a, b])}{b - a}, \quad \forall \mathrm{A} \in \mathcal{B}.$$

Alors $m_{a,b}$ est une probabilité sur $(\mathbb{R}, \mathcal{B})$ (la « loi uniforme sur $[a, b]$ »), et sa fonction de répartition est

$$\mathrm{F}_{a,b}(x) = m_{a,b}(]-\infty, x]) = \begin{cases} 0 & \text{si } x < a \\ \dfrac{x-a}{b-a} & \text{si } a \leqslant x < b \\ 1 & \text{si } b \leqslant x. \end{cases} \tag{11.1}$$

Donc $m_{a,b}$ est déterminée de manière unique. De plus comme

$$m(\mathrm{A}) = \sum_{n \in \mathbb{Z}} m_{n,n+1}(\mathrm{A}), \qquad \forall \mathrm{A} \in \mathcal{B}. \tag{11.2}$$

($\mathbb{Z}$ désigne l'ensemble des entiers relatifs) on voit que m est déterminée de manière unique également. ■

Maintenant que nous savons que la mesure de Lebesgue est unique, il nous reste à montrer qu'elle existe !

Théorème 11.2. *La mesure de Lebesgue existe.*

Preuve. Comme $F_{a,b}$ donnée par (11.1) existe et est croissante, continue, et vaut 0 pour x assez petit et 1 pour x assez grand, la mesure $m_{a,b}$ existe d'après le théorème 7.1. Il suffit donc de définir m par (11.2). La vérification de la σ-additivité et de $m(]a,b]) = b - a$ est immédiate. ■

La théorie de l'intégration décrite au chapitre 9 reste valide pour la mesure de Lebesgue ; la seule différence est que $m(\mathbb{R})$ n'est pas égal à 1, mais à $+\infty$: tous les résultats du chapitre 9 restent vrais, à l'exception du fait que toute fonction borélienne bornée est intégrable (théorème 9.1(b)) et du fait que $L^2 \subset L^1$ (théorème 9.3(b)), qui sont maintenant faux : par exemple la fonction identiquement égale à 1 est borélienne bornée mais pas intégrable ; la fonction $f(x) = \frac{1}{x}\, 1_{[1,\infty[}(x)$ est de carré intégrable mais pas intégrable.

Si f est une fonction borélienne intégrable par rapport à la mesure de Lebesgue, son intégrale est notée $\int f(x)dx$. Rappelons que f est intégrable si et seulement si $\int f^+(x)dx < \infty$ et $\int f^-(x)dx < \infty$, où $f = f^+ - f^-$. Toute fonction borélienne à support compact (i.e., nulle en dehors d'un intervalle $[-N, N]$) et intégrable au sens de Riemann est aussi intégrable au sens de Lebesgue (la réciproque est fausse), et dans ce cas les deux intégrales coïncident.

Définition 11.2. *La* densité *d'une probabilité* P *sur* $(\mathbb{R}, \mathcal{B})$ *est une fonction borélienne positive* f *qui vérifie pour tout* $x \in \mathbb{R}$:

$$P(]-\infty, x]) = \int_{-\infty}^{x} f(y)dy = \int f(y) 1_{]-\infty,x]}(y)dy. \tag{11.3}$$

Si $P = P_X$, *la loi d'une v.a. réelle* X, *on dit alors que* X admet la densité f.

Attention : comme on l'a déjà vu, il n'est *pas vrai* que toute probabilité sur $(\mathbb{R}, \mathcal{B})$ admette une densité, puisque dans ce cas la fonction de répartition F est continue, ce qui n'est pas vrai de toutes les fonctions de répartition. Il existe même des fonctions de répartition continues qui n'admettent pas de densité.

Théorème 11.3. (a) *Une fonction borélienne positive f sur $\mathbb{R}$ est la densité d'une probabilité* P *sur* $(\mathbb{R}, \mathcal{B})$ *si et seulement si elle vérifie* $\int f(x)dx = 1$. *Dans ce cas elle détermine la probabilité de manière unique, et toute autre fonction borélienne positive g vérifiant $m(f \neq g) = 0$ est aussi une densité de* P.

(b) *Inversement si une probabilité admet une densité, celle-ci est déterminée de manière unique à un ensemble de mesure de Lebesgue nulle près* (i.e., si f et g sont deux densités pour la même probabilité, on a $m(f \neq g) = 0$).

Preuve. (a) Soit f la densité de P. Par (11.3) on a $\int_{-\infty}^{x} f(y)dy = \mathrm{P}((-\infty, x])$. Si on fait tendre x vers l'infini, on voit que

$$\int f(y)dy = \int_{-\infty}^{\infty} f(y)dy = \lim_{x\to\infty} \int_{-\infty}^{x} f(y)dy = 1.$$

donc $\int f(x)dx = 1$.

Pour la condition suffisante, on pourrait utiliser la fonction de répartition, mais nous allons faire une démonstration directe, qui s'étend immédiatement au cas des probabilités sur $\mathbb{R}^n$ étudiées au chapitre suivant. Soit donc f une fonction borélienne positive vérifiant $\int f(x)dx = 1$. Pour tout borélien A on pose

$$\mathrm{P}(\mathrm{A}) = \int_{\mathrm{A}} f(y)dy = \int f(y)1_{\mathrm{A}}(y)dy. \tag{11.4}$$

Cela définit une application $\mathrm{P} : \mathcal{B} \to \mathbb{R}_+$ qui vérifie clairement $\mathrm{P}(\mathbb{R}) = 1$. De plus si les $\mathrm{A}_n \in \mathcal{B}$ sont deux à deux disjoints, alors

$$\begin{aligned}\mathrm{P}\left(\bigcup_{i=1}^{\infty} \mathrm{A}_i\right) &= \int f(x)1_{\{\bigcup_{i=1}^{\infty} \mathrm{A}_i\}}(x)dx \\ &= \int \left(\sum_{i=1}^{\infty} f(x)1_{\mathrm{A}_i}(x)\right) dx \\ &= \sum_{i=1}^{\infty} \int f(x)1_{\mathrm{A}_i}(x)dx = \sum_{i=1}^{\infty} \mathrm{P}(\mathrm{A}_i)\end{aligned}$$

en utilisant le théorème 9.2. Il s'ensuit que P est une probabilité sur $(\mathbb{R}, \mathcal{B})$. Si $\mathrm{A} =]-\infty, x]$ il vient

$$\mathrm{P}(]-\infty, x]) = \int_{-\infty}^{x} f(y)dy,$$

donc P admet f pour densité. L'unicité de P admettant f pour densité provient immédiatement du théorème 7.1. Enfin si g est une autre fonction borélienne telle que $m(f \neq g) = 0$, dans (11.4) on peut remplacer f par g sans changer $\mathrm{P(A)}$, ce qui montre que g est aussi une densité de P.

(b) Montrons maintenant que si P admet deux densités f et g, alors $m(f \neq g) = 0$. On a alors (11.4) pour f et aussi pour g (pour vérifier ceci, on définit P′ par (11.4) avec g au lieu de f, et on observe que P et P′ admettent la même fonction de répartition, donc sont égales). Soit alors $n \in \mathbb{N}$ et $\mathrm{A}_n = \left\{x : f(x) + \frac{1}{n} \leqslant g(x)\right\}$. On a

$$\mathrm{P}(\mathrm{A}_n) + \frac{1}{n}m(\mathrm{A}_n) = \int \left(f(x) + \frac{1}{n}\right) 1_{\mathrm{A}_n}(x)dx \leqslant \int g((x) 1_{\mathrm{A}_n}(x)dx = \mathrm{P}(\mathrm{A}_n),$$

donc nécessairement $m(\mathrm{A}_n) = 0$. Comme A_n croît vers $\{f < g\}$ quand $n \to \infty$, on obtient $m(\{g < f\}) = 0$ par le théorème de limite monotone. On obtient de même $m(\{g > f\}) = 0$, d'où le résultat. ■

Remarque 11.1. La densité f et la fonction de répartition F étant liées par $\mathrm{F}(x) = \int_{-\infty}^{x} f(y)dy$, on est tenté de conclure que F est dérivable et que sa dérivée vaut $\mathrm{F}'(x) = f(x)$. C'est vrai en tout point x où la fonction f est continue. On peut montrer (c'est un résultat – difficile – dû à Lebesgue) que c'est vrai pour m-presque tout x. Comme f n'est définie de manière unique qu'à un ensemble de mesure de Lebesgue nulle près, « concrètement » si on connaît F et si on sait que f existe (par exemple si F est continue partout et dérivable par morceaux), on peut prendre pour f la dérivée de F partout où celle-ci existe, et des valeurs arbitraires (par exemple 0) ailleurs.

Corollaire 11.1. *Soit* X *une v.a. réelle admettant la densité* f. *Si* g *est une fonction borélienne sur* $\mathbb{R}$, *elle est intégrable (resp. admet une intégrale) par rapport à la loi* $\mathrm{P_X}$ *de* X *si et seulement si la fonction produit* fg *est intégrable (resp. admet une intégrale) par rapport à la mesure de Lebesgue, et on a alors*

$$\mathrm{E}(g(\mathrm{X})) = \int g(x)\mathrm{P_X}(dx) = \int g(x)f(x)dx. \tag{11.5}$$

Preuve. L'égalité (11.5) et satisfaite pour les fonctions g indicatrices grâce au théorème 11.3, car elle se réduit alors à (11.4). Par linéarité elle est donc vraie pour les fonctions boréliennes simples, puis par limite croissante pour les fonctions boréliennes positives (les trois termes de (11.5) étant simultanément finis, ou infinis). Pour les fonctions de signe quelconque, on déduit le résultat par linéarité encore, en utilisant les décompositions $g = g^+ - g^-$ et $fg = fg^+ - fg^-$. ■

Nous avons présenté divers exemples de densités au chapitre 7. Dans tous ces exemples la densité était continue ou continue par morceaux, tandis qu'ici nous considérons le cas plus général où la densité est seulement borélienne. La plupart des exemples concrets donnent lieu à des densités régulières, mais dès qu'on fait des opérations relativement simples sur les v.a. ayant des densités régulières (par exemple si l'on prend des espérances conditionnelles, voir plus loin), on obtient de nouvelles variables pouvant avoir des densités simplement mesurables.

Soit X une v.a. réelle de densité f. Soit $Y = g(X)$ pour une autre fonction borélienne g donnée. Peut-on exprimer la densité de Y, si elle existe, en termes de f et g ? On le peut, dans les « bons cas », comme nous allons le voir maintenant. Commençons par un résultat trivial :

Théorème 11.4. *Soit X une v.a. de densité f_X et g une fonction borélienne. Soit $Y = g(X)$. La fonction de répartition de Y est*

$$F_Y(y) = P(Y \leqslant y) = \int_{A_y} f_X(u)du \qquad \text{où } A_y = \{u : g(u) \leqslant y\}.$$

Si F_Y est continue partout et dérivable sauf en un nombre fini de points, on peut utiliser le résultat ci-dessus pour obtenir la densité f_Y de Y.

Exemple. Soit X uniforme sur $[0, 1]$ et $Y = -\frac{1}{\lambda}\log X$, où $\lambda > 0$. Alors

$$\begin{aligned} F_Y(y) &= P\left(-\frac{1}{\lambda}\log X \leqslant y\right) \\ &= P(\log X \geqslant -\lambda y) \\ &= P(X \geqslant \exp(-\lambda y)) \\ &= \begin{cases} 1 - e^{-\lambda y} & \text{si } y \geqslant 0 \\ 0 & \text{sinon.} \end{cases} \end{aligned}$$

Donc (cf. la remarque 11.1)

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \begin{cases} \lambda e^{-\lambda y} & \text{si } y > 0 \\ 0 & \text{si } y \leqslant 0 \end{cases}$$

et on voit que Y est exponentielle de paramètre λ.

Attention : cet exemple est très simple car g est injective. Le résultat général pour g injective est donné ci-dessous, le cas non injectif étant donné plus bas.

Corollaire 11.2. *Supposons que* X *admette une densité continue* f_X. *Soit* $g : \mathbb{R} \to \mathbb{R}$ *une fonction continûment dérivable, dont la dérivée ne s'annule pas (elle est donc strictement monotone), et soit* $h = g^{-1}$ *sa fonction réciproque, qui est définie sur l'image* $g(\mathbb{R})$ *de* $\mathbb{R}$ *par* g *et continûment dérivable sur* $g(\mathbb{R})$. *Alors* $Y = g(X)$ *admet la densité*

$$f_Y(y) = \begin{cases} f_X(h(y))|h'(y)| & \text{si } y \in g(\mathbb{R}) \\ 0 & \text{sinon.} \end{cases}$$

Preuve. Supposons g croissante. L'image $g(\mathbb{R})$ est un intervalle $]a,b[$, avec éventuellement $a = -\infty$ et/ou $b = +\infty$. Soit $F_Y(y) = P(Y \leqslant y)$. Si $a < y < b$ on a

$$F_Y(y) = P\left(g(X) \leqslant y\right) = P\left(h(g(X)) \leqslant h(y)\right),$$

car h est strictement croissante. Donc ce qui précède vaut

$$= P\left(X \leqslant h(y)\right) = F_X(h(y)) = \int_{-\infty}^{h(y)} f(x)dx.$$

On sait par ailleurs que la dérivée de h est $h'(x) = \frac{1}{f'(h(x))}$. Donc F_Y est dérivable sur $]a,b[$ et

$$\frac{d}{dy}F_Y(y) = f(h(y))h'(y) = f(h(y))|h'(y)|.$$

Enfin $F_Y(y) = 0$ si $y \leqslant a$, quand $a > -\infty$, et de même $F_Y(y) = 1$ si $y \geqslant b$, quand $b < \infty$; donc F_Y est également dérivable, et de dérivée nulle, sur $]-\infty, a[$ et sur $]b, \infty[$. On en déduit le résultat.

Si g est décroissante, pour $y \in g(\mathbb{R})$ le même argument conduit à

$$\frac{d}{dy}F_Y(y) = f(h(y))(-h'(y)) = f(h(y))|h'(y)|.$$

et on conclut comme ci-dessus. ∎

Corollaire 11.3. *Supposons que* X *admette une densité continue par morceaux* f_X. *Soit* $g : \mathbb{R} \to \mathbb{R}$ *une fonction continûment dérivable et strictement monotone par morceaux, i.e. il existe une partition* $I_1, \ldots, I_n$ *de* $\mathbb{R}$ *constituée d'intervalles et telle que* g *soit continûment dérivable avec* $g' \neq 0$ *sur chaque intervalle ouvert* I_i' *ayant mêmes extrémités que* I_i. *Pour chaque* i *on note* h_i *la fonction réciproque de la restriction de* g *à* I_i', *et* $\Delta_i = g(I_i)$ *l'image de l'intervalle* I_i *par* g. *Alors la v.a.* $Y = g(X)$ *admet la densité*

$$f_Y(y) = \sum_{i=1}^{n} f_X(h_i(y))\ |h_i'(y)|\ 1_{\Delta_i}(y).$$

Remarque. La preuve, analogue à celle du corollaire précédent, est laissée au lecteur. Cette méthode de démonstration utilise la continuité de f_X, mais le résultat reste vrai quand f_X est simplement borélienne.

Exemple. Soit X une v.a. de loi N(0, 1), et $Y = X^2$. On applique le résultat précédent à $g(x) = x^2$, qui n'est pas injective. Prenons $I_1 = [0, \infty[$ et $I_2 =]-\infty, 0[$. Alors g est injective et strictement monotone sur I_1 et sur I_2, de sorte que $h_1 :]0, \infty[\to \mathbb{R}$ avec $h_1(y) = \sqrt{y}$ et $h_2 :]0, \infty[\to \mathbb{R}$ avec $h_2(y) = -\sqrt{y}$. On a

$$|h_i'(y)| = \left|\frac{1}{2\sqrt{y}}\right| = \frac{1}{2\sqrt{y}}, \qquad \text{pour } i = 1, 2.$$

Donc le corollaire 11.3 entraine

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}} e^{-y/2} \frac{1}{2\sqrt{y}} + \frac{1}{\sqrt{2\pi}} e^{-y/2} \frac{1}{2\sqrt{y}} \qquad (y > 0) \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-y/2} 1_{(0,\infty)}(y). \end{aligned}$$

La v.a. Y est appelée une variable χ^2 (ou, « chi-deux ») à 1 degré de liberté.

L'exemple ci-dessus est suffisamment simple pour pouvoir être traité « à la main », sans le corollaire 11.3. En fait

$$\begin{aligned} F_Y(y) &= P(Y \leqslant y) = P(X^2 \leqslant y) \\ &= P(-\sqrt{y} \leqslant X \leqslant \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}); \end{aligned}$$

et

$$F_X(\sqrt{y}) = \int_{-\infty}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Donc en dérivant on obtient

$$\frac{d}{dy} F_X(\sqrt{y}) = \frac{1}{\sqrt{2\pi}} e^{-y/2} \frac{1}{2\sqrt{y}} 1_{(y>0)}.$$

De même

$$-F_X(-\sqrt{y}) = -\int_{-\infty}^{-\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx,$$

donc

$$\begin{aligned}\frac{d}{dy}(-\mathrm{F}(-\sqrt{y})) &= -\frac{1}{\sqrt{2\pi}}e^{-y/2}\frac{-1}{2\sqrt{y}}1_{(y>0)}\\ &= \frac{1}{\sqrt{2\pi}}e^{-y/2}\frac{1}{2\sqrt{y}}1_{(y>0)},\end{aligned}$$

et en additionnant on obtient le même résultat qu'auparavant.

Remarque. Les *lois du chi-deux* jouent un rôle important en statistique. Soit p un entier supérieur ou égal à 1. La probabilité sur $\mathbb{R}$ de densité

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}}x^{p/2-1}e^{-\frac{x}{2}}, \qquad 0 < x < \infty$$

est appelée *loi du chi-deux à p degrés de liberté*, ou χ_p^2. Il s'agit d'un cas particulier des lois gamma, celles de paramètres $(\frac{p}{2}, \frac{1}{2})$. On vient de voir que si X admet la loi χ_1^2, alors $\mathrm{X} = \mathrm{Z}^2$ où Z est $\mathrm{N}(0,1)$. On verra au chapitre 15 (exemple 6) que si X est χ_p^2 on peut l'écrire comme $\mathrm{X} = \sum_{i=1}^p \mathrm{Z}_i^2$, où les Z_i sont indépendantes de loi $\mathrm{N}(0,1)$. De telles lois apparaissent en statistique lorsqu'on cherche à estimer la variance (inconnue) d'une famille de v.a. indépendantes de loi normale (voir l'exercice 13 du chapitre 15).

Noter aussi que χ_2^2 est simplement la loi exponentielle de paramètre $\lambda = \frac{1}{2}$.

Exercices

1. Utiliser la densité de la loi du chi-deux pour montrer que $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.
2. Soit X une v.a. uniforme sur $[-1,1]$. Trouver la densité de $\mathrm{Y} = \mathrm{X}^k$ pour k entier $\geqslant 1$. [Rép. : pour k impair, $f_\mathrm{Y}(y) = \frac{1}{2k}y^{\frac{1}{k}-1}1_{[-1,1]}(y)$; pour k pair, $f_\mathrm{Y}(y) = \frac{1}{k}y^{\frac{1}{k}-1}1_{[0,1]}(y)$.]
3. Soit X une v.a. de fonction de répartition F. Quelle est la fonction de répartition de $\mathrm{Y} = |\mathrm{X}|$? Quand X admet une densité f_X, montrer que Y admet aussi une densité f_Y, et exprimer f_Y en fonction de f_X.
4. Soit X une v.a. de Cauchy de paramètres α et 1. Soit $\mathrm{Y} = \frac{a}{\mathrm{X}}$ avec $a \neq 0$. Montrer que Y est aussi de Cauchy et trouver ses paramètres. [Rép. : $\frac{a\alpha}{1+\alpha^2}$, $\sqrt{\frac{|a|}{1+\alpha^2}}$].

5. Soit X une v.a. de densité f_X, et soit $Y = \frac{a}{X}$ avec $a \neq 0$. Trouver la densité f_Y de Y en fonction de f_X. [Rép. : $f_Y(y) = \frac{|a|}{y^2} f_X(\frac{a}{y})$.]
6. Soit X une v.a. uniforme sur $]-\pi, \pi[$, et $Y = \sin(X+\theta)$, avec $\theta \in \mathbb{R}$. Montrer que Y admet la densité $f_Y(y) = \frac{2}{2\pi\sqrt{1-y^2}} 1_{[-1,1]}(y)$.
7.* Soit X une v.a. de densité f_X, et $Y = a\sin(X+\theta)$ avec $a > 0$ et $\theta \in \mathbb{R}$. Montrer que Y admet la densité

$$f_Y(y) = \frac{1}{\sqrt{a^2 - y^2}} \sum_{i=-\infty}^{\infty} \big(f_X(h_i(y)) + f_X(k_i(x))\big) 1_{[-a,a]}(y)$$

pour des fonctions h_i et k_i qu'on déterminera.
8. Soit X une v.a. uniforme sur $]-\pi, \pi[$, et $Y = a\tan X$ avec $a > 0$. Trouver la densité f_Y de Y. [Rép. : $f_Y(y) = \frac{a/\pi}{a^2+y^2}$.]
9.* Soit X une v.a. de densité f_X, et

$$Y = ce^{-\alpha X} 1_{\{X>0\}}, \qquad (\alpha > 0, c > 0).$$

Trouver la densité f_Y de Y en fonction de f_X. [Rép. : cette densité n'existe pas si $P(X \leqslant 0) > 0$, et sinon elle existe et vaut $f_Y(y) = \frac{f_X(-\frac{1}{\alpha}\ln(\frac{y}{c}))}{\alpha y} 1_{]0,c[}(y)$.]
10. Une *densité f est appelée symétrique* si $f(-x) = f(x)$, i.e. si c'est une fonction paire. Une v.a. X est dite *symétrique* si X et $-X$ ont même loi. Montrer qu'une v.a. X admettant une densité est symétrique si et seulement si elle admet une densité symétrique. Dans ce cas, admet-elle aussi une densité non symétrique ? [Rép. : Oui, il suffit de modifier la densité symétrique sur un ensemble de mesure de Lebesgue nulle sur $]0, \infty[$.] [*Note*. Exemples de densités symétriques : la densité uniforme sur $[-a, a]$; la densité normale $N(0, \sigma^2)$; la densité exponentielle double de paramètres 0 et β ; la densité de Cauchy de paramètres 0 et β.]
11. Soit X une v.a. positive de densité f, et $Y = \frac{1}{X+1}$. Trouver la densité de Y.
12. Soit X une v.a. normale $N(\mu, \sigma^2)$. Montrer que $Y = e^X$ est de loi lognormale.
13. Soit X une v.a. de fonction de répartition F continue. Montrer que la v.a. $Y = F(X)$ est uniforme sur $[0, 1]$.

14. Soit F une fonction de répartition continue et injective, de sorte que sa fonction réciproque F^{-1} soit bien définie sur $]0,1[$. Soit U une v.a. uniforme sur $]0,1[$. Montrer que $X = F^{-1}(U)$ admet F pour fonction de répartition.

15.* Soit F une fonction de répartition continue, et U une v.a. uniforme sur $]0,1[$. Posons $G(u) = \inf\{x : F(x) \geqslant u\}$. Montrer que $X = G(U)$ admet F pour fonction de répartition.

16. Soit $Y = -\frac{1}{\lambda} \log U$, où U est uniforme sur $]0,1[$. Montrer que Y est exponentielle de paramètre λ en inversant la fonction de répartition de la loi exponentielle (*Indication :* si U est uniforme sur $]0,1[$, il en est de même de $1-U$). Cela donne une méthode pour « simuler » une v.a. exponentielle à partir d'une v.a. uniforme.

Chapitre 12

Probabilités sur $\mathbb{R}^n$

Dans le chapitre 11 nous avons étudié les lois (probabilités) sur $(\mathbb{R}, \mathcal{B})$. Le cas des probabilités sur $(\mathbb{R}^n, \mathcal{B}^n)$ pour $n \geqslant 2$ est à la fois similaire et plus compliqué. ($\mathcal{B}^n$ désigne la tribu borélienne de $\mathbb{R}^n$.)

D'abord, et avec une démonstration essentiellement identique à celle du théorème 2.1, on voit que $\mathcal{B}^n$ est la tribu engendrée par les « rectangles » de la forme

$$\prod_{i=1}^{n}]-\infty, a_i] ; \qquad a_i \in \mathbb{Q}.$$

En particulier $\mathcal{B} \otimes \mathcal{B} \otimes \cdots \otimes \mathcal{B} = \mathcal{B}^n$, i.e., $\mathcal{B}^n$ est aussi la tribu engendrée par le produit cartésien $\mathcal{B} \times \cdots \times \mathcal{B}$.

La *fonction de répartition* d'une probabilité sur $(\mathbb{R}^n, \mathcal{B}^n)$ est la fonction sur $\mathbb{R}^n$ définie par

$$\mathrm{F}(x_1, \ldots, x_n) = \mathrm{P}\left(\prod_{i=1}^{n}]-\infty, x_i]\right).$$

La caractérisation des fonctions de répartition lorsque $n \geqslant 2$ est nettement plus délicate que dans le cas $n = 1$, aussi ne sont-elles que très rarement utilisées.

Nous avons aussi observé que la densité d'une probabilité sur $\mathbb{R}$, lorsqu'elle existe, est un outil très efficace. Contrairement aux fonctions de répartition, la notion de densité se généralise au cas $n \geqslant 2$ en offrant la même facilité d'utilisation.

Définition 12.1. *La* mesure de Lebesgue m_n *sur* $(\mathbb{R}^n, \mathcal{B}^n)$ *est définie sur les produits cartésiens* $\mathrm{A}_1 \times \mathrm{A}_2 \times \cdots \times \mathrm{A}_n$ *par*

$$m_n\left(\prod_{i=1}^{n} \mathrm{A}_i\right) = \prod_{i=1}^{n} m(\mathrm{A}_i), \qquad \forall \mathrm{A}_i \in \mathcal{B}, \tag{12.1}$$

où m est la mesure de Lebesgue sur $(\mathbb{R}, \mathcal{B})$. Comme dans le théorème 10.3, cette formule permet de définir une unique mesure m_n sur $(\mathbb{R}^n, \mathcal{B}^n)$ (vérifiant (12.1)), et m_n est même caractérisée par le fait qu'on ait

$$m_n\left(\prod_{i=1}^{n}\,]a_i,b_i]\right)=\prod_{i=1}^{n}(b_i-a_i),\qquad \forall\ \infty<a_i<b_i<\infty. \tag{12.2}$$

Si $\mathrm{A}\in\mathcal{B}^n$, on peut se représenter $m_n(\mathrm{A})$ comme le « volume » de l'ensemble A, comme le suggère la formule (12.2). On note

$$\int f(x)\,dx=\int f(x_1,\ldots,x_n)\,dx_1dx_2\ldots dx_n$$

l'intégrale de f par rapport à m_n, et $\int_{\mathrm{A}} f(x)dx$ celle de $f1_{\mathrm{A}}$ quand $\mathrm{A}\in\mathcal{B}^n$, exactement comme dans le cas uni-dimensionnel.

Définition 12.2. *Une probabilité* P *sur* $(\mathbb{R}^n,\mathcal{B}^n)$ *admet la* densité f *si cette fonction est borélienne positive sur* $\mathbb{R}^n$ *et vérifie*

$$\begin{aligned}\mathrm{P}(\mathrm{A})&=\int_{\mathrm{A}} f(x)dx=\int f(x)1_{\mathrm{A}}(x)dx\\&=\int f(x_1,\ldots,x_n)1_{\mathrm{A}}(x_1,\ldots,x_n)\,dx_1\ldots dx_n.\end{aligned}$$

pour tout $\mathrm{A}\in\mathcal{B}^n$.

Une fois de plus, nous insistons sur le fait que certaines probabilités sur $(\mathbb{R}^n,\mathcal{B}^n)$ n'ont pas de densité : par exemple, et exactement comme pour $\mathbb{R}$, la masse de Dirac en un point $a\in\mathbb{R}^n$ n'admet pas de densité.

Le théorème suivant est l'exact analogue du théorème 11.3, avec une preuve identique, que nous ne répétons pas.

Théorème 12.1. (a) *Une fonction borélienne positive f sur $\mathbb{R}^n$ est la densité d'une probabilité* P *sur* $(\mathbb{R}^n,\mathcal{B}^n)$ *si et seulement si elle vérifie* $\int f(x)dx=1$. *Dans ce cas elle détermine la probabilité de manière unique, et toute autre fonction borélienne positive g vérifiant* $m_n(f\neq g)=0$ *est aussi une densité de* P.

(b) *Inversement si une probabilité admet une densité, celle-ci est déterminée de manière unique à un ensemble de mesure de Lebesgue nulle près* (i.e., si f et g sont deux densités pour la même probabilité, on a $m_n(f\neq g)=0$).

Afin de garder des notations simples le résultat suivant est écrit pour $n=2$; il s'étend, sans aucune difficulté, au cas $n\geqslant 3$.

Soit X une v.a. à valeurs dans $\mathbb{R}^2$, de composantes Y et Z : i.e., $\mathrm{X}=(\mathrm{Y},\mathrm{Z})$. On dit que X admet la densité f si sa loi (donc une probabilité sur $\mathbb{R}^2$) admet f pour densité.

Théorème 12.2. *Si* $\mathrm{X}=(\mathrm{Y},\mathrm{Z})$ *admet la densité f, alors :*

(a) *Les v.a. réelles* Y *et* Z *admettent les densités suivantes :*

$$f_Y(y) = \int_{-\infty}^{\infty} f(y,z)dz; \qquad f_Z(z) = \int_{-\infty}^{\infty} f(y,z)dy. \qquad (12.3)$$

(b) *Les v.a.* Y *et* Z *sont indépendantes si et seulement si*

$$f(y,z) = f_Y(y)f_Z(z) \qquad (m_2\text{-}p.p.).$$

(c) *Pour chaque y tel que $f_Y(y) \neq 0$, la formule ci-dessous définit une densité sur $\mathbb{R}$:*

$$f_{Y=y}(z) = \frac{f(y,z)}{f_Y(y)}.$$

Dans (b), « m_2-p.p. » (où « p.p. » abrège « presque partout ») signifie « en dehors d'un ensemble de m_2-mesure nulle ». Avant de prouver ce théorème, nous allons faire quelques commentaires concernant (a) et (c). D'abord, les densités f_Y et f_Z sont appelées les *densités marginales* de f. Ces deux densités marginales ne permettent pas, à elles seules, de retrouver la densité « jointe » f (il faut pour cela disposer d'une information additionnelle, par exemple que X et Y sont indépendantes dans le cas de (b)).

La fonction $f_{Y=y}(z)$ s'appelle la *densité conditionnelle de* Z *sachant que* $Y = y$. Cette assertion n'est pas à prendre au sens littéral, puisque $P(Y = y) = 0$ pour chaque y et donc $P(A \mid Y = y)$ n'a *a priori* pas de sens. Néanmoins cette terminologie admet la justification heuristique suivante : soit Δy et Δz des quantités positives « très petites ». Alors

$$f(y,z)\Delta y \Delta z \approx P(y \leqslant Y \leqslant y + \Delta y; z \leqslant Z \leqslant z + \Delta z)$$

et

$$f_Y(y)\Delta y \approx P(y \leqslant Y \leqslant y + \Delta y);$$

si de plus $P(y \leqslant Y \leqslant y + \Delta y) > 0$ (ce qui est vrai si par exemple $f_Y(y) > 0$ et si f_Y est continue en y), on obtient alors :

$$\begin{aligned} f_{Y=y}(z)\,\Delta z &\approx \frac{P(y \leqslant Y \leqslant y + \Delta y; z \leqslant Z \leqslant z + \Delta z)}{P(y \leqslant Y \leqslant y + \Delta y)} \\ &\approx P(z \leqslant Z \leqslant z + \Delta z \mid Y \approx y). \end{aligned}$$

Preuve du théorème 12.2. (a) Pour tout borélien $A \in \mathcal{B}$ on a

$$\begin{aligned}
P(Y \in A) = P(X \in A \times \mathbb{R}) &= \iint_{A\times\mathbb{R}} f(y,z)dy\,dz \\
&= \int_A dy \int_{-\infty}^{\infty} f(y,z)dz \\
&= \int_A dy\, f_Y(y).
\end{aligned}$$

Comme les densités sur $\mathbb{R}$ sont caractérisées par (11.4), f_Y définies dans (12.1) est une densité de Y. La preuve pour f_Z est la même.

(b) Supposons que $f(y,z) = f_Y(y)f_Z(z)$. On a

$$\begin{aligned}
P(Y \in A, Z \in B) &= \iint 1_{A\times B}(y,z)f(y,z)dy\,dz \\
&= \iint 1_A(y)1_B(z)f_Y(y)f_Z(z)dy\,dz \\
&= \int 1_A(y)f_Y(y)dy \int 1_B(z)f_Z(z)dz \\
&= P(Y \in A)P(Z \in B),
\end{aligned}$$

et comme A et B sont des boréliens arbitraires on a que Y et Z sont indépendantes.

Supposons inversement Y et Z indépendantes. Soit

$$\mathcal{H} = \left\{C \in \mathcal{B}^2 : \iint_C f(y,z)dy\,dz = \iint_C f_Y(y)f_Z(z)dy\,dz\right\}.$$

À cause de l'indépendance, si $C = A \times B$ avec $A \in \mathcal{B}$ et $B \in \mathcal{B}$, on a

$$\begin{aligned}
\iint_C f(y,z)dy\,dz &= P((Y,Z) \in C) \\
&= P(Y \in A, Z \in B) \\
&= P(Y \in A)P(Z \in B) \\
&= \int_A f_Y(y)dy \int_B f_z(z)dz \\
&= \iint_C f_Y(y)f_Z(z)dy\,dz
\end{aligned}$$

par le théorème de Fubini 10.3. Donc $\mathcal{H}$ contient la classe de tous les produits $C = A \times B$ avec $A, B \in \mathcal{B}$; cette classe est stable par intersection finie et engendre la tribu $\mathcal{B}^2$. Comme de plus $\mathcal{H}$ est stable par limites croissantes et par différences, le théorème des classes monotones (théorème 6.2) entraîne $\mathcal{H} = \mathcal{B}^2$. Donc

$$P(X \in C) = \int_C f(y,z)dy\,dz = \int_C f_Y(y) f_Z(z) dy\,dz$$

pour tout $C \in \mathcal{B}^2$. L'unicité de la densité (théorème 12.1) donne

$$f(y,z) = f_Y(y) f_Z(z), \qquad m_2\text{-p.p.}$$

(c) On a

$$\begin{aligned} \int f_{Y=y}(z)dz &= \int_{-\infty}^{\infty} \frac{f(y,z)}{f_Y(y)} dz \\ &= \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f(y,z)dz = \frac{1}{f_Y(y)} f_Y(y) = 1. \end{aligned}$$

Comme $f_{Y=y}$ est positive, borélienne, et d'intégrale 1, c'est donc une densité. ∎

Définition 12.3. *Soit* X *et* Y *deux v.a. réelles, ayant toutes deux une variance finie. La* covariance *de* X *et* Y *est définie comme étant le nombre*

$$\mathrm{Cov}(X,Y) = E\Big((X - E(X))(Y - E(Y))\Big) = E(XY) - E(X)E(Y).$$

Noter que E(XY) existe : comme X et Y ont des variances finies elles sont toutes deux dans L^2, et le théorème 9.3(a) implique que $XY \in L^1$. Remarquer aussi que

$$\mathrm{Cov}(X,X) = \mathrm{Var}\,(X) = \sigma^2(X).$$

Théorème 12.3. *Si* X *et* Y *sont indépendantes, alors* $\mathrm{Cov}(X,Y) = 0$.

Preuve. X et Y indépendantes implique $E(XY) = E(X)E(Y)$, d'où le résultat. ∎

Attention : la réciproque du théorème 12.3 est fausse ; on peut avoir $\mathrm{Cov}(X,Y) = 0$ sans que X et Y soient indépendantes.

Définition 12.4. *Soit* X *et* Y *deux v.a. de variances finies non nulles. Le* coefficient de corrélation *de* X *et* Y *est le nombre*

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

($\sigma(X) = \sqrt{\sigma^2(X)}$ et $\sigma(Y) = \sqrt{\sigma^2(Y)}$.)

Noter que d'après l'inégalité de Cauchy-Schwarz (théorème 9.3(a)) on a toujours $-1 \leqslant \rho \leqslant 1$, et si X et Y sont indépendantes alors $\rho = 0$ (théorème 12.3).

Définition 12.5. *Soit* $X = (X_1, \ldots, X_n)$ *une v.a. à valeurs dans* $\mathbb{R}^n$. *La* matrice de covariance de X *est la matrice* $n \times n$ *dont le terme général est*

$$c_{ij} = \mathrm{Cov}(X_i, X_j).$$

Théorème 12.4. *Une matrice de covariance est symétrique et semi-définie positive : i.e.,* $c_{ij} = c_{ji}$ *et* $\sum a_i a_j c_{ij} \geqslant 0$ *pour tous réels* a_i.

Preuve. La symétrie est évidente, puisque $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$. Un calcul simple montre que

$$\sum a_i a_j c_{ij} = \sigma^2 \left(\sum_{i=1}^{n} a_i X_i \right),$$

et comme une variance est toujours positive on a le résultat. ■

Théorème 12.5. *Soit* X *une v.a. à valeurs dans* $\mathbb{R}^n$, *de matrice de covariance* C. *Soit* A *une matrice* $m \times n$, *et posons* $Y = AX$. *Alors* Y *est une v.a. à valeurs dans* $\mathbb{R}^m$ *et sa matrice de covariance est* $C' = ACA^*$, *où* A^* *est la transposée de* A.

Preuve. Il s'agit d'un simple calcul. ■

Passons maintenant à l'étude des fonctions de v.a. vectorielles, i.e. à valeurs dans $\mathbb{R}^n$. On s'intéresse au problème suivant : soit $g : \mathbb{R}^n \to \mathbb{R}^n$ une fonction borélienne. Étant donnée une v.a. $X = (X_1, \ldots, X_n)$ à valeurs dans $\mathbb{R}^n$, admettant la densité f, quelle est la densité de $Y = g(X)$? et pour commencer, existe-t-elle ?

Nous aurons besoin d'un résultat classique et important, mais de démonstration assez difficile, sur les changements de variables dans les intégrales multiples (voir par exemple [6] ou [25, p. 83]). Rappelons d'abord que si g est une fonction différentiable d'un ouvert G de $\mathbb{R}^n$ dans $\mathbb{R}^n$, sa *matrice jacobienne* $J_g(x)$ au point $x \in G$ est $J_g(x) = \frac{\partial g}{\partial x}(x)$ (i.e., $J_g(x)_{ij} = \frac{\partial g_i}{\partial x_j}(x)$, où $g = (g_1, g_2, \ldots, g_n)$). Le *jacobien* de g au point x est le déterminant de la matrice jacobienne $J_g(x)$. Si le jacobien en x est non nul, alors la fonction g est inversible sur un voisinage de x, et le jacobien de l'inverse (ou fonction réciproque) g^{-1} au point $y = g(x)$ est l'inverse du jacobien de g en x.

Théorème 12.6. (Formule de changement de variable.) *Soit* G *un ouvert de* $\mathbb{R}^n$ *et* $g : \mathrm{G} \to \mathbb{R}^n$ *une fonction continûment différentiable. Supposons de plus que* g *soit injective sur* G *et que son jacobien de s'annule jamais sur* G. *Soit* f *une fonction borélienne telle que la fonction produit* $f\mathbf{1}_{g(\mathrm{G})}$ *soit positive ou intégrable par rapport à la mesure de Lebesgue* m_n, *avec*

$$g(\mathrm{G}) = \{y \in \mathbb{R}^n \text{ : il existe } x \in \mathrm{G} \text{ avec } g(x) = y\}.$$

On a alors

$$\int_{g(\mathrm{G})} f(y)dy = \int_{\mathrm{G}} f(g(x))|\det(\mathrm{J}_g(x))|dx.$$

Ce qui suit est une simple application du théorème 12.6 aux densités de v.a.

Théorème 12.7. *Soit* $\mathrm{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_n)$ *une v.a. à valeurs dans* $\mathbb{R}^n$, *admettant une densité* f_X. *Soit* $g : \mathbb{R}^n \to \mathbb{R}^n$ *une fonction continûment différentiable et injective, dont le jacobien ne s'annule pas. Alors la v.a.* $\mathrm{Y} = g(\mathrm{X})$ *admet la densité*

$$f_\mathrm{Y}(y) = \begin{cases} f_\mathrm{X}(g^{-1}(y))|\det \mathrm{J}_{g^{-1}}(y)| & \text{si } y \text{ est dans l'image de } g \\ 0 & \text{sinon.} \end{cases}$$

Preuve. On note $\mathrm{G} = g(\mathbb{R}^n)$ l'image de g. Les propriétés de g impliquent que G est un ouvert et que la fonction réciproque g^{-1} est bien définie sur G et continûment différentiable avec un jacobien ne s'annulant pas. Soit $\mathrm{B} \in \mathcal{B}^n$, et $\mathrm{A} = g^{-1}(\mathrm{B})$. On a

$$\begin{aligned}
\mathrm{P}(\mathrm{X} \in \mathrm{A}) &= \int_\mathrm{A} f_\mathrm{X}(x)dx \\
&= \int_{g^{-1}(\mathrm{B})} f_\mathrm{X}(x)dx \\
&= \int_\mathrm{B} f_\mathrm{X}(g^{-1}(x))|\det \mathrm{J}_{g^{-1}}(x)|dx,
\end{aligned}$$

par le théorème 12.6 appliqué à g^{-1}. On a aussi $\mathrm{P}(\mathrm{Y} \in \mathrm{B}) = \mathrm{P}(\mathrm{X} \in \mathrm{A})$, donc

$$\mathrm{P}(\mathrm{X} \in \mathrm{A}) = \int_\mathrm{B} f_\mathrm{Y}(y)dy.$$

Comme $\mathrm{B} \in \mathcal{B}^n$ est arbitraire on en déduit que

$$f_\mathrm{X}(g^{-1}(x))|\det \mathrm{J}_{g^{-1}}(x))| = f_\mathrm{Y}(x),$$

m_n-p.p., d'où le résultat. ∎

De manière analogue au corollaire 11.3 du chapitre 11, on peut aussi traiter le cas où g n'est pas injective mais seulement « injective par morceaux ».

Corollaire 12.1. *Soit* $S_0, S_1, \dots, S_m$ *des boréliens deux à deux disjoints de* $\mathbb{R}^n$*, avec* $m_n(S_0) = 0$ *et* S_i *ouvert pour* $i = 1, \dots, m$. *Soit* X *une v.a. à valeurs dans* $\mathbb{R}^n$*, admettant une densité* f_X *nulle en dehors de* $S = \bigcup_{i=0}^m S_i$ (donc X est en fait à valeurs dans S). *Soit* g *une fonction de* S *dans* $\mathbb{R}^n$ *telle que pour chaque* $i = 1, \dots, m$ *la restriction* g_i *de* g *à* S_i *soit continûment différentiable, injective, de jacobien ne s'annulant pas. La v.a.* $Y = g(X)$ *admet alors la densité*

$$f_Y(y) = \sum_{i=1}^m f_X(g_i^{-1}(y)) |\det J_{g_i^{-1}}(y)| 1_{g_i(S_i)}(y).$$

Exemples.

1. Soit X et Y des v.a. indépendantes $N(0,1)$. Calculons la densité du couple $(U, V) = (X + Y, X - Y)$. On a

$$g(x, y) = (x + y, x - y) = (u, v),$$
$$g^{-1}(u, v) = \left(\frac{u+v}{2}, \frac{u-v}{2}\right).$$

La matrice jacobienne ne dépend pas ici de (u, v) et vaut

$$J_{g^{-1}}(u, v) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix}.$$

et

$$\det J_{g^{-1}} = \frac{1}{2}\left(-\frac{1}{2}\right) - \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = -\frac{1}{2}.$$

Donc

$$\begin{aligned} f_{(U,V)}(u, v) &= f_{(X,Y)}\left(\frac{u+v}{2}, \frac{u-v}{2}\right) |\det J| \\ &= f_X\left(\frac{u+v}{2}\right) f_Y\left(\frac{u-v}{2}\right) |J| \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{u+v}{2})^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{u-v}{2})^2} \cdot \frac{1}{2} \\ &= \frac{1}{\sqrt{4\pi}} e^{-\frac{u^2}{4}} \frac{1}{\sqrt{4\pi}} e^{-\frac{v^2}{4}} \end{aligned}$$

pour $u, v \in \mathbb{R}$. On en déduit que les v.a. U et V sont indépendantes et N(0, 2).

2. Soit (X, Y) une v.a. bidimensionnelle de densité f. On veut trouver la densité de $U = XY$. Dans ce cas $h(x, y) = xy$ applique $\mathbb{R}^2$ dans $\mathbb{R}^1$, et on ne peut donc pas utiliser le théorème 12.7. On peut cependant s'y ramener par une astuce simple. Posons
$$g(x, y) = (xy, x).$$
Soit $S_0 = \{(x, y) : x = 0, y \in \mathbb{R}\}$ et $S_1 = \mathbb{R}^2 \backslash S_0$. Alors $m_2(S_0) = 0$ et g est injective de S_1 dans $\mathbb{R}^2$ et $g^{-1}(u, v) = (v, \frac{u}{v})$. La matrice jacobienne est
$$J_{g^{-1}}(u, v) = \begin{pmatrix} 0 & \dfrac{1}{v} \\ 1 & -\dfrac{u}{v^2} \end{pmatrix},$$
et $\det(J_{g^{-1}}) = -\frac{1}{v}$. Donc si $V = X$ et si on considère la v.a. bidimensionnelle (U, V), le corollaire 12.1 donne
$$f_{(U,V)}(u, v) = \begin{cases} f\left(v, \frac{u}{v}\right) \frac{1}{|v|} & \text{si } v \neq 0 \\ 0 & \text{si } v = 0. \end{cases}$$
Rappelons qu'on cherche la densité f_U de U, qui est donnée par
$$f_U(u) = \int_{-\infty}^{\infty} f_{(U,V)}(u, v)\, dv = \int_{-\infty}^{\infty} f\left(v, \frac{u}{v}\right) \frac{1}{|v|}\, dv.$$

3. Parfois on peut calculer une densité directement, sans passer par le théorème 12.6 ou l'astuce de l'exemple 2 ci-dessus.
Soit par exemple X et Y indépendantes de loi N(0, 1), et $Z = X^2 + Y^2$. Quelle est la densité f_Z de Z ?
Le théorème 12.2(b) entraîne que la densité du couple (X, Y) est
$$f(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right),$$
et donc
$$\begin{aligned} P(Z \in A) &= E(1_A(Z)) = E(1_A(X^2 + Y^2)) \\ &= \iint 1_A(x^2 + y^2) f(x, y) dx\, dy. \\ &= \iint 1_A(x^2 + y^2) \frac{1}{2\pi} e^{-(\frac{x^2 + y^2}{2})} dx\, dy \end{aligned}$$
En passant en coordonnées polaires il vient

$$= \frac{1}{2\pi}\int_0^\infty \int_0^{2\pi} 1_{\mathrm{A}}(r^2)e^{-r^2/2}\; r\,dr\,d\theta$$
$$= \frac{1}{2\pi}\int_0^{2\pi} d\theta \int_0^\infty 1_{\mathrm{A}}(r^2)e^{-r^2/2}\; r\,dr$$
$$= \int_0^\infty 1_{\mathrm{A}}(r^2)e^{-r^2/2}\; r\,dr$$

En posant $z = r^2$, donc $dz = 2r\,dr$:

$$= \int_0^\infty 1_{\mathrm{A}}(z)\frac{1}{2}e^{-(z/2)}dz,$$

et on voit que Z admet la densité $\frac{1}{2}e^{-(z/2)}$ de l'exponentielle de paramètre $\frac{1}{2}$. Noter que la transformation en coordonnées polaires n'est pas injective sur $\mathbb{R}^2$, de sorte que pour justifier l'argument précédent il faut utiliser de nouveau le corollaire 12.1 : cette transformation est injective sur $\mathrm{S}_1 = \mathbb{R}^2\backslash\{0\}$, tandis que l'ensemble $\mathrm{S}_0 = \{0\}$ est de mesure de Lebesgue nulle. Cet argument sera utilisé sans mention explicite dans la suite.

4. Quand une fonction transforme n v.a. en une seule on peut parfois éviter de combiner le théorème 12.6 et l'astuce de l'exemple 2 en utilisant à la place la fonction de répartition. Plus précisément si $\mathrm{Y} = g(\mathrm{X}_1, \ldots, \mathrm{X}_n)$, avec g borélienne de $\mathbb{R}^n$ dans $\mathbb{R}$, et si f est la densité de $(\mathrm{X}_1, \ldots, \mathrm{X}_n)$, alors

$$\mathrm{F_Y}(y) = \mathrm{P}(\mathrm{Y} \leqslant y) = \int_{g(x_1,\ldots,x_n)\leqslant y} f(x)dx.$$

Supposons qu'il existe une fonction $h(y; x_2, \ldots, x_n)$ telle que

$$g(x_1, x_2, \ldots, x_n) \leqslant y \quad \text{si et seulement si} \quad x_1 \leqslant h(y; x_2, \ldots, x_n).$$

Alors

$$\begin{aligned}\mathrm{F_Y}(y) &= \int_{\{x : g(x)\leqslant y\}} f(x)\,dx \\ &= \int_{-\infty}^{\infty}\!\!\ldots\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{h(y;x_2,\ldots,x_n)} f(x_1, \ldots, x_n)dx_1\right\}dx_2, \ldots, dx_n.\end{aligned}$$

Si nous dérivons les deux membres par rapport à y (ce qui suppose en particulier que h soit continûment dérivable en y et f soit continue), on obtient

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{\partial h(y; x_2, \ldots, x_n)}{\partial y} \times \tag{12.4}$$
$$\times f\left(h(y, x_2, \ldots, x_n), x_2, \ldots, x_n\right) dx_2 \ldots dx_n.$$

Un exemple d'utilisation de cette technique concerne le problème suivant, important en statistique. Soit X et Y des v.a. indépendantes, avec X de loi $N(0, 1)$ et Y de loi gamma de paramètres $\alpha = n/2$ avec n entier non nul et $\beta = \frac{1}{2}$. [Y suit donc la loi du chi-deux à n degrés de liberté.] Quelle est la loi de

$$Z = \frac{X}{\sqrt{Y/n}}\ ?$$

On a $g(x, y) = \frac{x}{\sqrt{y/n}}$, et comme $\frac{x}{\sqrt{y/n}} \leqslant z$ si et seulement si $x \leqslant z\sqrt{y/n}$, on peut prendre $h(z; y) = z\sqrt{y/n}$. À cause de l'indépendance la densité f de (X, Y) est

$$f(x, y) = f_X(x)f_Y(y) = \left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\right)\left(\frac{y^{\frac{n}{2}-1}e^{-\frac{1}{2}y}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\right) 1_{]0,\infty[}(y).$$

On peut alors appliquer (12.4) :

$$f_Z(z) = \frac{1}{\sqrt{2\pi n}\Gamma(\frac{n}{2})2^{\frac{n}{2}}} \int_0^{\infty} y^{\frac{n-1}{2}} e^{-y\frac{(1+\frac{z^2}{n})}{2}} dy \tag{12.5}$$
$$= \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi n}(1 + \frac{z^2}{n})^{\frac{1}{2}(n+1)}}.$$

La loi de densité f_Z dans (12.5) s'appelle la *loi t de Student à n degrés de liberté.*

La loi t de Student, introduite pour l'étude statistique de la moyenne des lois normales lorsque les variances sont inconnues, a été introduite par W. Gosset (1876-1937) qui a utilisé pour cela le pseudonyme « Student ».

5. Soit X et Y des v.a. indépendantes de loi $N(0, \sigma^2)$. Soit $Z = \sqrt{X^2 + Y^2}$ et $W = \frac{X}{Y}$ si $Y \neq 0$ et $W = 0$ si $Y = 0$. On veut trouver la densité $f_{(Z,W)}$ de (Z, W). Ici,

$$g(x, y) = \left(\sqrt{x^2 + y^2}, \frac{x}{y}\right) = (z, w)$$

et g n'est pas injective. On a

$$g^{-1}(z,w) = \left(\frac{zw}{\sqrt{1+w^2}}, \frac{z}{\sqrt{1+w^2}}\right) = (x,y).$$

Une autre fonction « inverse » serait

$$h^{-1}(z,w) = \left(\frac{-zw}{\sqrt{1+w^2}}, \frac{-z}{\sqrt{1+w^2}}\right).$$

La matrice jacobienne $J_{g^{-1}}$ de g^{-1} est

$$\begin{pmatrix} \frac{w}{\sqrt{1+w^2}} & \frac{1}{\sqrt{1+w^2}} \\ \frac{z}{(1+w^2)^{\frac{3}{2}}} & \frac{-zw}{(1+w^2)^{\frac{3}{2}}} \end{pmatrix}$$

de déterminant $\frac{-z}{1+w^2}$. Le corollaire 12.1 entraîne alors

$$\begin{aligned} f_{(Z,W)}(z,w) = \frac{z}{1+w^2}\Bigg\{ & f_{(X,Y)}\left(\frac{zw}{\sqrt{1+w^2}}, \frac{z}{\sqrt{1+w^2}}\right) \\ & + f_{(X,Y)}\left(\frac{-zw}{\sqrt{1+w^2}}, \frac{-z}{\sqrt{1+w^2}}\right)\Bigg\}. \end{aligned}$$

La densité normale étant symétrique, on a

$$f_{(X,Y)}(x,y) = f_{(X,Y)}(-x,-y),$$

de sorte que

$$f_{(Z,W)}(z,w) = \frac{2z}{1+w^2}\frac{1}{2\pi\sigma^2}e^{-\frac{z^2}{2\sigma^2}}1_{(z>0)}.$$

Noter que cette densité se factorise :

$$f_{(Z,W)}(z,w) = \frac{1}{\pi\sigma^2}\left(\frac{1}{1+w^2}\right)\left(ze^{-\frac{z^2}{2\sigma^2}}1_{(z>0)}\right).$$

On en déduit que Z et W sont indépendantes, ce qui n'est pas *a priori* évident, et on peut lire immédiatement les densités de Z et W : en fait, comme $\frac{1}{\pi(1+w^2)}$ est la densité d'une loi de Cauchy de paramètres $\alpha = 0$ et $\beta = 1$, il vient :

$$f_Z(z) = \frac{z}{\sigma^2}e^{-\frac{z^2}{2\sigma^2}}1_{(z>0)}$$

et

$$f_W(w) = \frac{1}{\pi(1+w^2)},$$

La loi de densité f_Z ci-dessus est connue sous le nom de *loi de Rayleigh* de paramètre $\sigma^2 > 0$. Cet exemple montre aussi que le rapport de deux v.a. indépendantes normales $N(0, \sigma^2)$ suit une loi de Cauchy de paramètres $\alpha = 0$ et $\beta = 1$.

Exercices

1. Montrer que

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\frac{-(x^2+y^2)}{2\sigma^2}}\,dx\,dy = 2\pi\sigma^2.$$

et donc que $\frac{1}{2\pi\sigma^2}\, e^{-(x^2+y^2)/2\sigma^2}$ est une densité.

2. Soit (X, Y) une v.a. à valeurs dans $\mathbb{R}^2$ admettant une densité $f_{(X,Y)}$ qui se factorise : $f_{(X,Y)}(x, y) = g(x)h(y)$. Trouver les densités marginales f_X et f_Y.
3. Soit (X, Y) une v.a. à valeurs dans $\mathbb{R}^2$ admettant la densité

$$\begin{aligned} f(x,y) \quad &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \times \exp\left(-\frac{1}{2(1-r^2)}\left\{\frac{(x-\mu_1)^2}{\sigma_1^2} - \right.\right. \\ &\qquad \left.\left. \frac{2r(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right\}\right), \end{aligned}$$

où $r \in]-1, 1[$, $\sigma_1, \sigma_2 > 0$ et $\mu_1, \mu_2 \in \mathbb{R}$. Trouver $f_{X=x}(y)$.
[Rép. : $\frac{1}{\sigma_2\sqrt{2\pi(1-r^2)}} \exp(-\frac{1}{2\sigma_2^2(1-r^2)}\{y - \mu_2 - \frac{r\sigma_2}{\sigma_1}(x-\mu_1)\}^2)$.]

4. Soit $\rho_{X,Y}$ le coefficient de corrélation des v.a. réelles X et Y. Soit $a > 0$, $c > 0$ et $b \in \mathbb{R}$. Montrer que

$$\rho_{aX+b,cY+b} = \rho_{X,Y}.$$

(Cela montre le résultat utile suivant : le coefficient de corrélation ne dépend pas des « échelles » avec lesquelles on mesure X et Y.)

5. Pour $a \neq 0$, montrer que

$$\rho_{X,aX+b} = \frac{a}{|a|}\ ;$$

donc si $Y = aX + b$ est une fonction affine non constante de X, alors $\rho_{X,Y} = \pm 1$.

6. Soit X, Y des v.a. ayant des variances finies non nulles, et soit

$$Z = \left(\frac{1}{\sigma_Y}\right) Y - \left(\frac{\rho_{X,Y}}{\sigma_X}\right) X.$$

Montrer que $\sigma_Z^2 = 1 - \rho_{X,Y}^2$, et déduire que si $\rho_{X,Y} = \pm 1$, alors Y est une fonction affine non constante de X.

7.* (Gut (1995), p. 27.) Soit (X, Y) une v.a. à valeurs dans $\mathbb{R}^2$, de *loi uniforme sur la boule unité* :

$$f_{(X,Y)}(x,y) = \begin{cases} \dfrac{1}{\pi} & \text{si } x^2+y^2 \leqslant 1 \\ 0 & \text{si } x^2+y^2 > 1. \end{cases}$$

Trouver la loi de $R = \sqrt{X^2+Y^2}$. (*Indication* : Introduire la v.a. auxiliaire $S = \operatorname{Arctan} \frac{Y}{X}$.) [Rép. : $f_R(r) = 2r1_{[0,1]}(r)$.]

8. Soit (X, Y) une v.a. à valeurs dans $\mathbb{R}^2$, de densité f. Trouver la densité de $Z = X + Y$. (*Indication* : Trouver d'abord la densité du couple (Z, W), où $W = Y$.) [Rép. : $f_Z(z) = \int_{-\infty}^{\infty} f_{(X,Y)}(z - w, w)dw$.]

9. Soit X une v.a. normale $N(0, \sigma^2)$, et soit Θ une v.a. uniforme sur $]0, \pi[$: i.e. Θ a la densité $f(\theta) = \frac{1}{\pi}1_{]0,\pi[}(\theta)$. Supposons X et Θ indépendantes. Trouver la loi de $Z = X + a\cos\Theta$. (Ceci est utilisé en électricité). [Rép. : $f_Z(z) = \frac{1}{\pi\sigma\sqrt{2\pi}} \int_0^{\pi} e^{-(z-a\cos w)^2/2\sigma^2} dw$.]

10. Soit X et Y des v.a. réelles indépendantes de densités respectives f_X et f_Y, et soit $Z = g(X)$ et $W = h(Y)$, où g et h sont injectives et dérivables. Trouver la densité $f_{(Z,W)}$ du couple (Z, W).

11. Soit X et Y deux v.a. indépendantes de loi $N(0, \sigma^2)$. Soit

$$Z = \sqrt{X^2+Y^2} \quad \text{et} \quad W = \operatorname{Arctan} \frac{X}{Y}, \qquad -\frac{\pi}{2} < W \leqslant \frac{\pi}{2}.$$

Montrer que Z admet une loi de Rayleigh, que W est uniforme sur $]-\frac{\pi}{2}, \frac{\pi}{2}[$, et que Z et W sont indépendantes.

12. Soit $(X_1, \ldots, X_n)$ des v.a. réelles. Posons

$$\begin{aligned} Y_1 &= \min(X_i ; 1 \leqslant i \leqslant n) \\ Y_2 &= \text{ seconde plus petite valeur parmi } X_1, \ldots, X_n \\ &\ \vdots \\ Y_n &= \max(X_i : 1 \leqslant i \leqslant n). \end{aligned}$$

Montrer que $Y_1, \ldots, Y_n$ sont des v.a. et que $Y_1 \leqslant Y_2 \leqslant \cdots \leqslant Y_n$ (on dit que les Y_i forment le *réarrangement croissant*, ou la *statistique d'ordre* des X_i et on les note en général

$$Y_k = X_{(k)}.\ )$$

Supposons que les X_i sont indépendantes et toutes avec la même densité f. Montrer que la densité de la v.a. n-dimensionnelle $(Y_1, \ldots, Y_n)$ est donnée par

$$g(y_1, \ldots, y_n) = \begin{cases} n! \prod_{i=1}^{n} f(y_i) & \text{si } y_1 < y_2 < \cdots < y_n \\ 0 & \text{sinon.} \end{cases}$$

13. Soit $(X_1, \ldots, X_n)$ des v.a. indépendantes uniformes sur $]0, a[$. Montrer que la statistique d'ordre $(X_{(1)}, \ldots, X_{(n)})$ (cf. l'exercice précédent) admet la densité

$$g(y_1, \ldots, y_n) = \begin{cases} \frac{n!}{a^n} & \text{si } 0 < y_1 < y_2 < \cdots < y_n < a \\ 0 & \text{sinon.} \end{cases}$$

14. Dans la situation de l'exercice 12, en notant f la densité commune des X_i (supposées indépendantes) et F leur fonction de répartition, montrer que $X_{(k)}$ admet la densité

$$f_{(k)}(y) = k C_n^k f(y)(1 - F(y))^{n-k} F(y)^{k-1}.$$

15. (Simulation d'une variable aléatoire normale.) Soit U_1 et U_2 deux v.a. indépendantes uniformes sur $]0, 1[$. Soit $\theta = 2\pi U_1$ et $S = -\log U_2$.

    a) Montrer que S suit une loi exponentielle, et que $\sqrt{2S}$ suit une loi de Rayleigh.

    b) Posons $X = R\cos\theta$, $Y = R\sin\theta$. Montrer que X et Y sont indépendantes normales.

    [*Indication :* Pour a), la loi exponentielle est un cas particulier des lois gamma, c'est la loi χ_2^2. Pour b), utiliser la méthode de l'exercice 11 en sens inverse !]

    [*Remarque.* La méthode décrite dans cet exercice pour simuler les v.a. normales centrées est connue sous le nom de méthode de Box-Muller.]

Chapitre 13

Fonctions caractéristiques

Il arrive souvent en mathématiques qu'on puisse résoudre un problème ou obtenir certaines propriétés d'objets mathématiques en les « transformant » dans un autre espace, en résolvant le problème dans cet autre espace, et en revenant ensuite dans l'espace initial. Deux parmi les transformations importantes sont la transformée de Laplace et la transformée de Fourier. Ces transformées sont largement utilisées pour l'étude des équations différentielles ou la « théorie du signal » ; elles sont aussi extrêmement utiles en probabilités : elles servent à analyser les variables aléatoires (par exemple, à calculer leurs moments), ou à donner des preuves « élémentaires » et élégantes de certains théorèmes, comme le théorème-limite central (voir le chapitre 21). La transformée de Fourier est la plus sophistiquée des deux, et aussi la plus utile.

Avant de l'introduire il nous faut dire un mot de l'intégrale des fonctions complexes. Soit $(\mathrm{E}, \mathcal{E}, \mu)$ un espace probabilisé, et f une application de E dans l'ensemble des nombres complexes. On peut l'écrire $f = g+ih$, où $i = \sqrt{-1}$ est la racine carrée de -1 et où g et h sont des fonctions réelles (la *partie réelle* et la *partie imaginaire* de f, respectivement). La fonction module est $|f| = \sqrt{g^2 + h^2}$ (racine carrée positive). On dit alors que f est *mesurable* si les deux fonctions g et h sont mesurables, et *intégrable* si g et h sont toutes deux intégrables. L'intégrale de f est alors le nombre complexe

$$\int f(x)\mu(dx) = \int g(x)\mu(dx) + i \int h(x)\mu(dx). \tag{13.1}$$

Il est alors évident que *tous* les résultats du chapitre 9 sont valides pour les fonctions à valeurs complexes, en prenant séparément la partie réelle et la partie imaginaire. Noter en particulier que la fonction complexe mesurable f est intégrable si et seulement si la fonction réelle $|f|$ l'est : en effet on a $|g| \leqslant |f|$, $|h| \leqslant |f|$, et $|f| \leqslant |g|+|h|$. Il est également important de noter que l'intégrale est linéaire sur le corps des complexes : si a est un nombre complexe, alors $\int (af(x))\mu(dx) = a \int f(x)\mu(dx)$. Le seul résultat un peu délicat est la majoration suivante pour les modules :

$$\left|\int f(x)\mu(dx)\right| \leqslant \int |f(x)|\mu(dx). \tag{13.2}$$

Pour l'obtenir, on écrit le complexe $\int f(x)\mu(dx)$ sous la forme $\rho e^{i\theta}$; à cause de la linéarité de l'intégrale il vient

$$\left|\int f(x)\mu(dx)\right| = \rho = \int \left(f(x)e^{-i\theta}\right)\mu(dx),$$

et l'expression de droite, étant réelle, vaut aussi $\int k(x)\mu(dx)$, où k est la partie réelle de $fe^{-i\theta}$. Mais le théorème 9.1(b) implique $\left|\int k(x)\mu(dx)\right| \leqslant \int |k(x)|\mu(dx)$ et comme on a $|k| \leqslant |f|$, on obtient (13.2).

Une notation nous sera aussi nécessaire : nous écrivons $\langle x, y\rangle$ pour le produit scalaire de $x, y \in \mathbb{R}^n$. C'est-à-dire, si $x = (x_1, \dots, x_n)$ et $y = (y_1, \dots, y_n)$, alors

$$\langle x, y\rangle = \sum_{j=1}^{n} x_j y_j.$$

(Le produit scalaire est aussi souvent écrit $x \cdot y$).

Définition 13.1. *Soit μ une mesure de probabilité sur $\mathbb{R}^n$. Sa transformée de Fourier, notée $\widehat{\mu}$, est la fonction sur $\mathbb{R}^n$ définie par*

$$\widehat{\mu}(u) = \int e^{i\langle u, x\rangle}\mu(dx).$$

Ci-dessus on intègre la fonction complexe mesurable $x \mapsto e^{i\langle u, x\rangle}$, de module égal à 1, de sorte que d'après ce qui précède elle est intégrable. D'après (13.1), la transformée de Fourier s'écrit aussi

$$\widehat{\mu}(u) = \int \cos(\langle u, x\rangle)\mu(dx) + i\int \sin(\langle u, x\rangle)\mu(dx).$$

Si $n = 1$, on a $\langle u, x\rangle = ux$, donc $\widehat{\mu}(u) = \int e^{iux}\,\mu(dx)$.

Définition 13.2. *Soit X une v.a. à valeurs dans $\mathbb{R}^n$. Sa fonction caractéristique φ_X est la fonction sur $\mathbb{R}^n$ définie par*

$$\varphi_X(u) = E(e^{i\langle u, X\rangle}).$$

En d'autres termes,

$$\varphi_X(u) = \int e^{i\langle u, x\rangle}P_X(dx) = \widehat{P_X}(u) \tag{13.3}$$

où P_X est la loi de X. Ainsi, la fonction caractéristique de X est la transformée de Fourier de sa loi.

Théorème 13.1. *Soit μ une probabilité sur $\mathbb{R}^n$. La fonction $\widehat{\mu}$ est bornée (de module inférieur ou égal à 1), continue, et vérifie $\widehat{\mu}(0) = 1$.*

Preuve. On a déjà vu que $\widehat{\mu}$ est bien définie. Comme $|e^{i\langle u,x\rangle}| = 1$ pour tous $u, x \in \mathbb{R}^n$, on a

$$|\widehat{\mu}(u)| \leqslant \int |e^{i\langle u,x\rangle}|\ \mu(dx) = \int 1\ \mu(dx) = 1.$$

De plus

$$\widehat{\mu}(0) = \int e^{i\langle 0,x\rangle}\mu(dx) = \int 1\ \mu(dx) = 1.$$

Pour la continuité de $\widehat{\mu}$ il suffit de montrer que pour toute suite (u_p) convergeant vers une limite u (dans $\mathbb{R}^n$), on a $\widehat{\mu}(u_p) \to \widehat{\mu}(u)$. Pour cela, on remarque que $e^{i\langle u_p,x\rangle} \to e^{i\langle u,x\rangle}$ pour tout $x \in \mathbb{R}^n$, et comme les fonctions $x \mapsto e^{i\langle u,x\rangle}$ sont bornées en module par la fonction constante égale à 1, le résultat découle du théorème de convergence dominée de Lebesgue (théorème 9.16) [si on ne veut pas utiliser ce théorème pour les fonctions complexes, on peut l'utiliser séparément pour les parties réelle $x \mapsto \cos(\langle u, x\rangle)$ et imaginaire $x \mapsto \sin(\langle u, x\rangle)$, qui sont bornées en valeur absolue par 1]. ∎

On peut même montrer que $\widehat{\mu}$ est *uniformément continue*, mais nous n'aurons pas besoin de ce résultat ici.

Si $X = (X_1, \ldots, X_n)$ est une v.a. à valeurs dans $\mathbb{R}^n$, et si $E(|X|^m) < \infty$ pour un entier m ($|X|$ désigne le module du vecteur aléatoire X, c'est une v.a. réelle), alors $E(|X|^k) < \infty$ pour $k = 1, 2, \ldots, m$, et on dit que X *admet des moments jusqu'à l'ordre m*. Dans ce cas, pour tout choix d'indices $j_1, \ldots, j_m$ entre 1 et n, la v.a. produit (réelle) $X_{j_1} \ldots X_{j_m}$ est intégrable (car majorée par $|X|^m$), et l'espérance de cette v.a. s'appelle le $(j_1, \ldots, j_m)$-*moment* de X. Lorsque $n = 1$, on a évidemment $j_1 = \cdots = j_m = 1$, le produit ci-dessus est X^m, et son espérance $E(X^m)$ est le m-ième moment de X.

Théorème 13.2. *Soit X une v.a. à valeurs dans $\mathbb{R}^n$, admettant des moments jusqu'à l'ordre m. La fonction caractéristique φ_X de X est alors m fois continûment différentiable, et*

$$\frac{\partial^m}{\partial u_{j_1} \ldots \partial u_{j_m}}\ \varphi_X(u) = i^m E(X_{j_1} \ldots X_{j_m} e^{i\langle u,X\rangle}).$$

Preuve. On va en fait prouver la formulation équivalente pour les mesures de probabilité μ sur $\mathbb{R}^n$ (si $\mu = P_X$, rappelons que $\varphi_X = \widehat{\mu}$), l'hypothèse sur μ étant que la fonction $f(x) = |x|^m$ est intégrable par rapport à μ.

On veut montrer que $\widehat{\mu}$ est m fois continûment différentiable et que

$$\frac{\partial^m \widehat{\mu}}{\partial u_{j_1} \dots \partial u_{j_m}}(u) = i^m \int x_{j_1} \dots x_{j_m} e^{i\langle u,x\rangle} \mu(dx).$$

On donne la preuve pour $m = 1$ seulement, le cas général se montrant en répétant le même raisonnement m fois. Nous voulons donc montrer que pour tout j, la dérivée partielle $\frac{\partial \widehat{\mu}}{\partial u_j}$ existe au point u et vaut

$$\frac{\partial \widehat{\mu}}{\partial u_j}(u) = i \int e^{i\langle u,x\rangle} x_j \mu(dx), \tag{13.4}$$

et que de plus la fonction ci-dessus est continue en u.

Pour l'existence de $\frac{\partial \widehat{\mu}}{\partial u_j}$ et la formule (13.4), il suffit de montrer que pour toute suite de réels t_p convergeant vers 0, si $v = (v_1, \dots, v_n)$ désigne le vecteur unitaire de $\mathbb{R}^n$ dans la direction j (i.e. de coordonnées $v_k = 0$ pour $k \neq j$ et $v_j = 1$), alors la suite

$$\frac{1}{t_p}\{\widehat{\mu}(u + t_p v) - \widehat{\mu}(u)\} = \int e^{i\langle u,x\rangle} \frac{e^{it_p x_j} - 1}{t_p} \mu(dx). \tag{13.5}$$

converge vers le membre de droite de (13.4). La suite de fonctions $x \to \frac{e^{it_p x_j} - 1}{t_p}$ (où x_j est la $j^{\text{ième}}$ coordonnée de $x \in \mathbb{R}^n$) converge simplement vers $x \mapsto ix_j$ (appliquer le fait que la fonction exponentielle est dérivable). De plus

$$\left|\frac{e^{it_p x_j} - 1}{t_p}\right| \leqslant 2|x|,$$

et

$$\int 2|x|\mu(dx) < \infty$$

par hypothèse. Donc le théorème de convergence dominée de Lebesgue implique que (13.5) converge vers

$$i \int x_j e^{i\langle u,x\rangle} \mu(dx).$$

Donc on a (13.4).

La preuve de la continuité en u de la fonction $\partial\widehat{\mu}/\partial u_j$ dans (13.4) est exactement la même que la preuve de la continuité dans le théorème 13.1. ■

Corollaire 13.1. *Soit X une v.a. réelle admettant des moments jusqu'à l'ordre m. La fonction caractéristique φ_X de X est alors m fois continûment différentiable, et*

$$\frac{\partial^m}{\partial u^m}\,\varphi_X(u) = i^m \mathrm{E}(X^m e^{iuX}).$$

Une application immédiate de ce résultat est l'utilisation de la fonction caractéristique pour calculer les moments d'une v.a. réelle X. Pour les deux premiers moments, de loin les plus importants, on obtient immédiatement :

$$\mathrm{E}(X) = i\varphi_X'(0) \quad \text{si} \quad \mathrm{E}(|X|) < \infty \tag{13.6}$$

$$\mathrm{E}(X^2) = -\varphi_X''(0) \quad \text{si} \quad \mathrm{E}(X^2) < \infty. \tag{13.7}$$

Exemples.

1. *Variable de Bernoulli* (p). Si X est une v.a. de Bernoulli de paramètre p, i.e. $\mathrm{P}(X = 1) = p$ et $\mathrm{P}(X = 0) = 1 - p$, il vient

$$\varphi_X(u) = \mathrm{E}(e^{iuX}) = e^{iu0}(1-p) + e^{iu}p = \boxed{pe^{iu} + 1 - p}\,.$$

2. *Variable binomiale* (n, p). Si X est une v.a. binomiale de paramètres n et p, alors

$$\varphi_X(u) = \mathrm{E}(e^{iuX}) = \sum_{j=0}^{n} C_n^j\, e^{iuj} p^j (1-p)^{n-j} = \boxed{(pe^{iu} + 1 - p)^n}\,.$$

On aurait aussi pu remarquer que X s'écrit

$$X = \sum_{j=1}^{n} Y_j,$$

où les $Y_1, \ldots, Y_n$ sont indépendantes de loi de Bernoulli (p). Alors

$$\varphi_X(u) = \mathrm{E}\left(e^{iuX}\right) = \mathrm{E}\left(e^{iu\sum_{j=1}^n Y_j}\right) = \mathrm{E}\left(\prod_{j=1}^{n} e^{iuY_j}\right) = \prod_{j=1}^{n} \mathrm{E}\left(e^{iuY_j}\right)$$

par l'indépendance des Y_j ;

$$= \prod_{j=1}^{n} \varphi_{Y_j}(u) = (pe^{iu} + 1 - p)^n.$$

3. *Variable de Poisson* (λ).

$$\begin{aligned}\varphi_X(u) &= E(e^{iuX}) = \sum_{k=0}^{\infty} e^{iuk} P(X=k) \\ &= \sum_{k=0}^{\infty} e^{iuk} \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=0}^{\infty} \frac{(\lambda e^{iu})^k}{k!} e^{-\lambda} \\ &= e^{-\lambda} e^{\lambda e^{iu}} = \boxed{e^{\lambda(e^{iu}-1)}}\,.\end{aligned}$$

4. *Variable uniforme sur* $[-a, a]$.

$$\varphi_X(u) = E(e^{iuX}) = \frac{1}{2a}\int_{-a}^{a} e^{iux}dx = \frac{e^{iua} - e^{-iua}}{2aiu};$$

en utilisant les propriétés $e^z = \cos z + i \sin z$ et $\cos(a) = \cos(-a)$, on voit que ceci égale

$$= \frac{2i \sin au}{2aiu} = \boxed{\frac{\sin au}{au}}$$

si $u \neq 0$, et bien sûr $\varphi_X(0) = 1$.

5. *Variable normale* $N(0, 1)$. Le calcul de la fonction caractéristique d'une v.a. normale est un peu plus difficile. On peut le faire par un calcul d'intégrale le long d'un contour fermé du plan complexe et le théorème des résidus, ou par prolongement analytique (voir l'exercice 17 du chapitre 14). Voici une autre méthode, peu intuitive mais élémentaire. On a

$$\begin{aligned}\varphi_X(u) &= \int e^{iux} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \int \frac{\cos ux}{\sqrt{2\pi}} e^{-x^2/2}dx + i \int \frac{\sin ux}{\sqrt{2\pi}} e^{-x^2/2}dx.\end{aligned}$$

Comme $x \mapsto \sin ux\, e^{-x^2/2}$ est une fonction impaire et intégrable, son intégrale est nulle. Par suite

$$\varphi_X(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \cos ux\, e^{-x^2/2}dx.$$

Grâce au théorème 13.2 on peut dériver les deux membres par rapport à u, et on obtient

$$\varphi_X'(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} -x \sin ux\, e^{-x^2/2}dx.$$

Ensuite on intègre par parties :

$$= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u \cos ux\, e^{-x^2/2} dx = -u\varphi_X(u).$$

Par suite φ_X satisfait à l'équation différentielle $\varphi'_X(u) = -u\varphi_X(u)$, dont la solution générale est

$$\varphi_X(u) = Ce^{-u^2/2}$$

pour une constante C. Comme $\varphi_X(0) = 1$ on a $C = 1$, de sorte que

$$\boxed{\varphi_X(u) = e^{-u^2/2}}.$$

Théorème 13.3. *Soit* X *une v.a. à valeurs dans* $\mathbb{R}^n$ *et soit* $a \in \mathbb{R}^m$. *Soit* A *une matrice* $m \times n$. *Alors*

$$\varphi_{a+AX}(u) = e^{i\langle u,a\rangle}\varphi_X(A^*u),$$

pour tout $u \in \mathbb{R}^m$, *où* A^* *désigne la transposée de* A.

Preuve. On a

$$e^{i\langle u,a+AX\rangle} = e^{i\langle u,a\rangle}e^{i\langle A^*u,X\rangle},$$

et on obtient le résultat en prenant l'espérance des deux membres. ■

Exemples (suite).

6. *Variable normale* $N(\mu, \sigma^2)$. Si X est $N(\mu, \sigma^2)$, on vérifie facilement (voir exercice 18 du chapitre 14) que $Y = \frac{X-\mu}{\sigma}$ est $N(0, 1)$, et $X = \mu + \sigma Y$. D'après le théorème 13.3 et l'exemple 5, il vient

$$\boxed{\varphi_X = e^{iu\mu - u^2\sigma^2/2}}.$$

7. *Variable exponentielle* (λ). Soit X une v.a. exponentielle de paramètre λ. Alors

$$\varphi_X(u) = \int_0^\infty e^{iux}\lambda e^{-\lambda x} dx.$$

Un calcul formel donne

$$= \int_0^\infty \lambda e^{(iu-\lambda)x} dx = \boxed{\frac{\lambda}{\lambda - iu}}, \tag{13.8}$$

ce qui n'est pas vraiment rigoureux sur le plan mathématique. Une manière simple de justifier cela consiste à intégrer séparément les fonctions partie réelle et partie imaginaire, i.e. $x \mapsto \lambda e^{-\lambda x} \cos ux$ et $x \mapsto \lambda e^{-\lambda x} \sin ux$, en faisant une intégration par parties : un calcul élémentaire montre que la formule (13.8) est exacte.

8. *Variable gamma* (α, β). Soit X une v.a. de loi gamma de paramètres α et β. En utilisant une intégration dans le plan complexe et le théorème des résidus, on peut montrer que

$$\boxed{\varphi_X(u) = \frac{\beta^\alpha}{(\beta - iu)^\alpha}}\,.$$

On peut aussi calculer cette fonction caractéristique par une méthode élémentaire : voir l'exercice 19 du chapitre 14.

Chapitre 14

Propriétés des fonctions caractéristiques

Nous avons vu comment calculer les fonctions caractéristiques sur divers exemples (de manière équivalente : comment calculer des transformées de Fourier de probabilités). Pour que cela soit réellement utile nous avons besoin de savoir que cette transformée de Fourier caractérise la probabilité. C'est ce que nous prouvons dans le théorème ci-dessus. Sa démonstration utilise le théorème de Stone-Weierstrass et est donc un peu difficile pour ce livre ; vu l'importance du résultat, nous la donnons néanmoins.

Théorème 14.1. (Théorème d'unicité.) *La transformée de Fourier $\widehat{\mu}$ d'une probabilité μ sur $\mathbb{R}^n$ caractérise μ* : si deux probabilités admettent la même transformée de Fourier, elles sont égales.

Preuve. Soit

$$f_\sigma(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-|x|^2/2\sigma^2},$$

et

$$\hat{f}_\sigma(u) = e^{-|u|^2\sigma^2/2}.$$

Alors $f_\sigma(x)$ est la densité de $X = (X_1, \dots, X_n)$, où les X_j sont $N(0, \sigma^2)$ et indépendantes. L'exemple 6 du chapitre 13 et le théorème de Fubini entraînent :

$$\begin{aligned}
\int_{\mathbb{R}^n} f_\sigma(x) e^{i\langle u,x\rangle} dx &= \int_{\mathbb{R}^n} \prod_{j=1}^n \frac{1}{\sigma\sqrt{2\pi}}\ e^{\left(\frac{-x_j^2}{2\sigma^2} + iu_j x_j\right)} dx_1 \dots dx_n \\
&= \prod_{j=1}^n \int_{\mathbb{R}} \frac{1}{\sigma\sqrt{2\pi}}\ e^{\left(\frac{-x_j^2}{2\sigma^2} + iu_j x_j\right)} dx_j \\
&= \prod_{j=1}^n e^{-\frac{u_j^2\sigma^2}{2}} = \hat{f}_\sigma(u).
\end{aligned}$$

Donc

$$
\begin{aligned}
f_\sigma(u-v) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \hat{f}_\sigma\left(\frac{u-v}{\sigma^2}\right) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \int_{\mathbb{R}^n} f_\sigma(x) e^{i\langle\frac{u-v}{\sigma^2},x\rangle} dx.
\end{aligned}
$$

Supposons maintenant que μ_1 et μ_2 soient deux probabilités sur $\mathbb{R}^n$ ayant même transformée de Fourier $\widehat{\mu}_1 = \widehat{\mu}_2 = \widehat{\mu}$. Alors

$$
\begin{aligned}
\int f_\sigma(u-v)\mu_1(du) &= \int \frac{1}{(2\pi\sigma^2)^{n/2}} \left\{\int f_\sigma(x) e^{i\langle\frac{u-v}{\sigma^2},x\rangle} dx\right\} \mu_1(du) \\
&= \int f_\sigma(x) \frac{1}{(2\pi\sigma^2)^{n/2}} \widehat{\mu}\left(\frac{x}{\sigma^2}\right) e^{-i\frac{\langle v,x\rangle}{\sigma^2}} dx,
\end{aligned}
$$

(le lecteur vérifiera qu'on peut appliquer le théorème de Fubini ci-dessus), et la même égalité est vraie pour μ_2. Par suite

$$
\int g(x)\mu_1(dx) = \int g(x)\mu_2(dx)
$$

pour toute fonction $g \in \mathcal{H}$, où $\mathcal{H}$ est l'espace vectoriel engendré par toutes les fonctions de la forme $u \mapsto f(\sigma, u-v)$. On peut alors appliquer le théorème de Stone-Weierstrass[1] pour obtenir que $\mathcal{H}$ est dense dans C_0 pour la convergence uniforme, où C_0 est l'ensemble des fonctions continues « nulles à l'infini », i.e. l'ensemble des fonctions continues sur $\mathbb{R}^n$ telles que $\lim_{\|x\|\to\infty} |f(x)| = 0$. On obtient alors

$$
\int_{\mathbb{R}^n} g(x)\mu_1(dx) = \int_{\mathbb{R}^n} g(x)\mu_2(dx)
$$

pour tout $g \in C_0$. Comme la fonction indicatrice d'un ouvert est limite croissante d'une suite de fonctions positives dans C_0, le théorème de convergence monotone (théorème 9.1(d)) donne

$$
\mu_1(\mathrm{A}) = \mu_2(\mathrm{A}), \quad \text{pour tout ouvert } \mathrm{A} \subset \mathbb{R}^n.
$$

Finalement le théorème des classes monotones (théorème 6.2) donne

$$
\mu_1(\mathrm{A}) = \mu_2(\mathrm{A}) \quad \text{pour tout borélien } \mathrm{A} \subset \mathbb{R}^n,
$$

donc $\mu_1 = \mu_2$. ∎

1. Voir par exemple [23, p. 160].

Corollaire 14.1. *Soit $X = (X_1, \ldots, X_n)$ une v.a. à valeurs dans $\mathbb{R}^n$. Les variables réelles $(X_j)_{1 \leqslant j \leqslant n}$ sont indépendantes si et seulement si pour tout $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$ on a*

$$\varphi_X(u_1, \ldots, u_n) = \prod_{j=1}^{n} \varphi_{X_j}(u_j) \tag{14.1}$$

Preuve. Si les $(X_j)_{1 \leqslant j \leqslant n}$ sont indépendantes, alors

$$\begin{aligned}
\varphi_X(u) &= (e^{i\langle u, X\rangle}) = \mathrm{E}\left(e^{i\sum_{j=1}^{n} u_j X_j}\right) \\
&= \mathrm{E}\left(\prod_{j=1}^{n} e^{iu_j X_j}\right) = \prod_{j=1}^{n} \mathrm{E}(e^{iu_j X_j}) \quad \text{(par l'indépendance)} \\
&= \prod_{j=1}^{n} \varphi_{X_j}(u_j).
\end{aligned}$$

Supposons à l'inverse qu'on ait (14.1). Soit μ_X la loi de X sur $\mathbb{R}^n$ et soit μ_{X_j} les lois des X_j sur $\mathbb{R}$. Alors

$$\widehat{\mu}_X = (\widehat{\mu_{X_1} \otimes \mu_{X_2} \otimes \cdots \otimes \mu_{X_n}}),$$

et le théorème 14.1 entraîne que

$$\mu_X = \mu_{X_1} \otimes \mu_{X_2} \otimes \cdots \otimes \mu_{X_n},$$

qui équivaut à l'indépendance des X_j. ■

Attention : le fait que $\varphi_X(u, u, \ldots, u) = \prod_{j=1}^{n} \varphi_{X_j}(u)$ pour tout $u \in \mathbb{R}$ *n'est pas suffisant* pour que les v.a. X_j soient indépendantes.

Exercices des chapitres 13 et 14

Les trois premiers exercices nécessitent une intégration dans le plan complexe et le théorème des résidus.

1.* Soit $f(x) = \frac{1}{\pi(1+x^2)}$ la densité d'une v.a. X de Cauchy. Montrer que

$$\varphi_X(u) = e^{-|u|},$$

en intégrant le long d'un demi-cercle de diamètre $[-R, R]$ sur l'axe réel vers la gauche, puis du point $(-R, 0)$ au point $(R, 0)$ le long de l'axe réel ; puis passer à la limite ($R \to \infty$).

2.* Soit X une v.a. de loi gamma de paramètres α et β. Montrer que $\varphi_X(u) = \frac{\beta^\alpha}{(\beta - iu)^\alpha}$. [*Indication :* Intégrer le long du segment horizontal de $(d, 0)$ à $(c, 0)$ (avec $0 < c < d$), puis le long du segment vertical descendant de $(c, 0)$ jusqu'à l'intersection avec la droite d'équation $y = -\frac{ux}{\alpha}$, puis suivre cette droite vers le bas jusqu'au point d'abscisse d, puis remonter verticalement jusqu'au point $(d, 0)$; puis passer à la limite avec $d \to \infty$ et $c \to 0$.]

3.* Soit X une v.a. N(0, 1). Montrer que $\varphi_X(u) = e^{-u^2/2}$ en intégrant le long du segment horizontal de $(R, 0)$ à $(-R, 0)$, puis le long du segment vertical de $(-R, 0)$ à $(-R, -iu)$, puis horizontalement jusqu'à $(R, -iu)$, puis horizontalement jusqu'à $(R, 0)$; puis passer à la limite $(R \to \infty)$.

4.* Soit X une v.a. réelle telle que $E(X^2) < \infty$ et $E(X) = 0$. Montrer que $\sigma^2 := \mathrm{Var}\,(X) < \infty$, et que

$$\varphi_X(u) = 1 - \frac{1}{2} u^2 \sigma^2 + o(u^2)$$

quand $u \to 0$. [Rappelons qu'une fonction g est $o(t^\alpha)$ quand $t \to 0$ si $\lim_{t \to 0} \frac{|g(t)|}{t^\alpha} = 0$.]

5. Soit $X = (X_1, \ldots, X_n)$ une v.a. à valeurs dans $\mathbb{R}^n$. Montrer que

   a) $\varphi_X(u, 0, 0, \ldots, 0) = \varphi_{X_1}(u) \qquad (u \in \mathbb{R})$

   b) $\varphi_X(u, u, u, \ldots, u) = \varphi_{X_1 + \cdots + X_n}(u) \quad (u \in \mathbb{R})$

6. On note $\overline{z}$ le complexe conjugué de z, i.e. si $z = a + ib$ on a $\overline{z} = a - ib$ $(a, b \in \mathbb{R})$. Montrer que si X est une v.a., on a $\overline{\varphi_X(u)} = \varphi_X(-u)$.

7. Soit X une v.a. Montrer que $\varphi_X(u)$ est à valeurs réelles si et seulement si X a une loi symétrique, i.e. $P_X = P_{-X}$. [*Indication :* Utiliser l'exercice 6 et les théorèmes 13.3 et 14.1.]

8. Montrer que si X et Y sont des v.a. indépendantes et de même loi, alors $Z = X - Y$ a une loi symétrique. (*Indication :* On peut utiliser l'exercice 7.)

9. Soit $X_1, \ldots, X_n$ des v.a. réelles indépendantes, ayant chacune un moment d'ordre 3 fini (i.e. $E(|X_i|^3) < \infty$) et de moyenne nulle (i.e. $E(X_i) = 0$). Montrer que

$$E\left(\left(\sum_{i=1}^n X_i\right)^3\right) = \sum_{i=1}^n E(X_i^3).$$

(*Indication :* Utiliser les fonctions caractéristiques.)

10. Soit $\mu_1, \dots, \mu_n$ des probabilités sur $\mathbb{R}^n$. Soit $\lambda_j \geqslant 0$ avec $\sum_{j=1}^n \lambda_j = 1$. Posons $\nu = \sum_{j=1}^n \lambda_j \mu_j$. Montrer que ν est aussi une probabilité, et que

$$\widehat{\nu}(u) = \sum_{j=1}^n \lambda_j \widehat{\mu}_j(u).$$

11. Soit X une v.a. de loi exponentielle double (Laplace) de paramètres $\alpha = 0$, $\beta = 1$, i.e. admettant la densité

$$f_X(x) = \frac{1}{2} e^{-|x|}.$$

Montrer que $\varphi_X(u) = \frac{1}{1+u^2}$. (*Indication :* Utiliser l'exercice 10 avec pour μ_1 la loi d'une v.a. exponentielle Y, pour μ_2 la loi de $-Y$, et $\lambda_1 = \lambda_2 = \frac{1}{2}$.)

12.* (*Loi triangulaire.*) Soit X une v.a. réelle de densité $f_X(x) = (1 - |x|)1_{[-1,1]}(x)$. Montrer que $\varphi_X(u) = \frac{2(1-\cos u)}{u^2}$. (*Indication :* Soit U et V des v.a. indépendantes uniformes sur $\left[-\frac{1}{2}, \frac{1}{2}\right]$; considérer $U + V$. Observer aussi que $\left(\frac{e^{iu/2} - e^{-iu/2}}{iu}\right)^2 = \left(\frac{2\sin(u/2)}{u}\right)^2$.)

13. Soit X une v.a. positive. Sa *transformée de Mellin* est la fonction

$$T_X(\theta) = E(X^\theta)$$

pour toutes les valeurs de θ pour lesquelles l'espérance de X^θ existe.

a) Montrer que

$$T_X(\theta) = \varphi_{\log X}\left(\frac{\theta}{i}\right)$$

quand les deux membres sont bien définis.

b) Montrer que si X et Y sont indépendantes et positives, on a

$$T_{XY}(\theta) = T_X(\theta) T_Y(\theta).$$

c) Montrer que $T_{bX^a}(\theta) = b^\theta T_X(a\theta)$ pour $b > 0$ et $a\theta$ dans le domaine de définition de T_X.

14. Soit X une v.a. lognormale de paramètres (μ, σ^2). Trouver la transformée de Mellin (cf. exercice 13) $T_X(\theta)$. Utiliser cette transformée et le fait que $T_X(k) = E(X^k)$ pour calculer le k-ième moment de X pour $k = 1, 2, \dots$

15. Soit X une v.a. N(0, 1). Montrer que $E(X^{2n+1}) = 0$ et

$$E(X^{2n}) = \frac{(2n)!}{2^n n!} = (2n-1)(2n-3)\dots 3\cdot 1.$$

16.* Soit X une v.a. N(0, 1). Posons

$$M(s) = E(e^{sX}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(sx - \frac{1}{2}x^2\right) dx.$$

Montrer que $M(s) = e^{s^2/2}$. (*Indication :* Compléter le carré dans l'exponentielle.)

17.* Remplacer s par iu dans l'exercice 16 pour obtenir que $\varphi_X(u) = e^{-u^2/2}$; justifier ce calcul par la théorie du prolongement analytique des fonctions de variables complexes.

18. Soit X une v.a. $N(\mu, \sigma^2)$. Montrer que $Y = \frac{X-\mu}{\sigma}$ suit la loi N(0, 1).

19.* (Feller [12].) Soit X une v.a. de loi gamma de paramètres $\alpha > 0$ et $\beta = 1$. On peut calculer sa fonction caractéristique sans intégration dans le plan complexe : développer e^{ix} en série, et montrer que

$$\frac{1}{\Gamma(\alpha)} \sum_{n=1}^{\infty} \frac{(iu)^n}{n!} \int_0^{\infty} e^{-x} x^{n+\alpha-1} dx = \sum_{n=0}^{\infty} \frac{\Gamma(n+\alpha)}{n!\Gamma(\alpha)} (iu)^n,$$

puis que la somme de cette dernière série est $\frac{1}{(1-iu)^{\alpha}}$.

Chapitre 15

Sommes de variables aléatoires indépendantes

Une part importante des applications des probabilités découle des propriétés des sommes de variables aléatoires indépendantes. Un exemple simple apparaît en statistique : si nous répétons n fois la même expérience indépendamment et si nous appelons X_j le résultat de la j-ième expérience, la « valeur moyenne » est donnée par $\overline{x} = \frac{1}{n}\sum_{j=1}^{n} X_j$. La v.a. $\overline{x}$ est alors appelée un *estimateur* de la moyenne μ de la loi commune des X_j. La statistique étudie quand (et comment) $\overline{x}$ converge vers μ quand $n \to \infty$. Même une fois qu'on sait que $\overline{x}$ tend vers μ, on a aussi besoin de savoir quelle taille donner à n pour être raisonnablement sûr que $\overline{x}$ est proche de la vraie valeur μ, qui est en général inconnue. D'autres problèmes, plus ardus, se posent aussi naturellement : quelle est la loi de $\overline{x}$? Si on ne peut la déterminer, peut-on au moins l'approximer ? Quelle taille doit avoir n pour que cette approximation soit bonne ? Si on dispose d'informations complémentaires sur μ, comment les utiliser pour améliorer l'estimateur $\overline{x}$? Bien évidemment, avant de commencer à résoudre les plus simples de ces questions, il nous faut étudier les sommes de v.a. indépendantes.

Théorème 15.1. *Soit* X *et* Y *deux v.a. réelles. La loi* μ_Z *de* $Z = X + Y$ *est le* produit de convolution *des lois* μ_X *et* μ_Y *de* X *et* Y*, qui est défini par*

$$(\mu_X * \mu_Y)(A) = \iint 1_A(x+y)\mu_X(dx)\mu_Y(dy). \tag{15.1}$$

Preuve. Comme X et Y sont indépendantes, la loi du couple (X, Y) est $\mu_X \otimes \mu_Y$. Donc

$$E(g(X, Y)) = \iint g(x, y)\mu_X(dx)\mu_Y(dy).$$

et en particulier, avec $g(x, y) = f(x + y)$, on obtient pour toute fonction borélienne f sur $\mathbb{R}$, positive ou bornée :

$$E(f(X + Y)) = \iint f(x + y)\mu_X(dx)\mu_Y(dy), \tag{15.2}$$

Il suffit alors de prendre $f(x) = 1_A(x)$. ∎

Remarque 15.1. La formule (15.2) montre que pour $f : \mathbb{R} \to \mathbb{R}$ borélienne et $Z = X + Y$, avec X et Y indépendantes :

$$E(f(Z)) = \int f(z)(\mu_X * \mu_Y)(dz) = \iint f(x+y)\mu_X(dx)\mu_Y(dy).$$

Théorème 15.2. *Soit X et Y deux v.a. réelles indépendantes, et $Z = X + Y$. La fonction caractéristique φ_Z est le produit de φ_X et de φ_Y, i.e.,*

$$\varphi_Z(u) = \varphi_X(u)\varphi_Y(u).$$

Preuve. Prendre $f(z) = e^{iuz}$ et utiliser (15.2). ∎

Attention : si $Z = X + Y$, la propriété que $\varphi_Z(u) = \varphi_X(u)\varphi_Y(u)$ pour tout $u \in \mathbb{R}$ n'implique pas l'indépendance de X et Y.

Théorème 15.3. *Soit X et Y deux v.a. réelles indépendantes, et $Z = X + Y$.*

(a) *Si X admet la densité f_X, alors Z admet la densité f_Z donnée par*

$$f_Z(z) = \int f_X(z-y)\mu_Y(dy)$$

(b) *Si de plus Y admet la densité f_Y, alors*

$$f_Z(z) = \int f_X(z-y)f_Y(y)dy = \int f_X(x)f_Y(z-x)dx.$$

Preuve. (b) : Supposons (a) vrai. On a

$$f_Z(z) = \int f_X(z-y)\mu_Y(dy).$$

Comme de plus $\mu_Y(dy) = f_Y(y)dy$, on obtient la première égalité. La seconde s'obtient en échangeant les rôles de X et Y.

(a) : D'après le théorème 15.1 on a

$$\begin{aligned} \mu_Z(A) &= \iint 1_A(x+y)\mu_X(dx)\mu_Y(dy) \\ &= \int \left\{ \int 1_A(x+y) f_X(x)dx \right\} \mu_Y(dy). \end{aligned}$$

En posant $z = x + y$, donc « $dz = dx$ » :

$$= \int \left\{ \int 1_{\mathrm{A}}(z) f_{\mathrm{X}}(z-y) dz \right\} \mu_{\mathrm{Y}}(dy)$$

et en appliquant le théorème de Fubini :

$$= \int \left\{ \int f(z-y) \mu_{\mathrm{Y}}(dy) \right\} 1_{\mathrm{A}}(z) dz.$$

Comme A est un borélien arbitraire, on a le résultat. ■

Le résultat suivant est trivial, mais extrêmement utile.

Théorème 15.4. *Soit* X *et* Y *deux v.a. réelles indépendantes, de carré intégrable (i.e.* $\mathrm{E}(\mathrm{X}^2) < \infty$ *et* $\mathrm{E}(\mathrm{Y}^2) < \infty$*). On a alors*

$$\sigma^2_{\mathrm{X+Y}} = \sigma^2_{\mathrm{X}} + \sigma^2_{\mathrm{Y}}.$$

Preuve. À cause de l'indépendance, on a $\mathrm{E(XY)} = \mathrm{E(X)E(Y)}$, donc

$$\sigma^2_{\mathrm{X+Y}} = \mathrm{E}(\mathrm{X}^2) + 2\mathrm{E}(\mathrm{XY}) + \mathrm{E}(\mathrm{Y}^2) - (\mathrm{E}(\mathrm{X}) + \mathrm{E}(\mathrm{Y}))^2 = \sigma^2_{\mathrm{X}} + \sigma^2_{\mathrm{Y}}.$$

■

Exemples.

1. Soit $\mathrm{X}_1, \ldots, \mathrm{X}_n$ des v.a. de Bernoulli (p). Alors $\mathrm{Y} = \sum_{j=1}^n \mathrm{X}_j$ est binomiale (n, p). On a vu que

$$\mathrm{E}(\mathrm{Y}) = \mathrm{E}\left(\sum_{j=1}^n \mathrm{X}_j\right) = \sum_{j=1}^n \mathrm{E}(\mathrm{X}_j) = \sum_{j=1}^n p = np.$$

Noter que

$$\sigma^2_{\mathrm{X}_j} = \mathrm{E}(\mathrm{X}_j^2) - \mathrm{E}(\mathrm{X}_j)^2 = p - p^2 = p(1-p).$$

Donc le théorème 15.4 entraîne que

$$\sigma^2_{\mathrm{Y}} = \sum_{j=1}^n \sigma^2_{\mathrm{X}_j} = np(1-p).$$

Cette méthode de calcul de la variance est préférable à un calcul direct à partir de la loi de Y, qui nécessite de trouver la somme

$$\sigma^2_{\mathrm{Y}} = \sum_{j=0}^n (j - np)^2 \binom{n}{j} p^j (1-p)^{n-j}.$$

2. Soit X une v.a. de Poisson (λ) et Y une v.a. de Poisson (μ), indépendante de X. Alors $Z = X + Y$ suit aussi une loi de Poisson, de paramètre $(\lambda + \mu)$. Pour le voir, on écrit $\varphi_Z = \varphi_X \varphi_Y$, ce qui donne

$$\varphi_Z(u) = e^{\lambda(e^{iu}-1)} e^{\mu(e^{iu}-1)} = e^{(\lambda+\mu)(e^{iu}-1)},$$

   et on reconnait là la fonction caractéristique de la loi de Poisson de paramètre $(\lambda + \mu)$: donc Z suit la loi de Poisson grâce au théorème d'unicité des fonctions caractéristiques (théorème 14.1).

3. Soit X et Y des v.a. binomiales indépendantes de paramètres respectifs (n, p) et (m, p) (le même p pour les deux v.a.). Soit $Z = X + Y$. Alors

$$\varphi_Z = \varphi_X \varphi_Y,$$

   donc

$$\varphi_Z(u) = (pe^{iu} + (1-p))^n (pe^{iu} + (1-p))^m = (pe^{iu} + (1-p))^{n+m},$$

   qui est la fonction caractéristique d'une loi binomiale $(m + n, p)$; donc Z est binomiale $(m + n, p)$. Noter que pour ce résultat on n'a pas réellement besoin des fonctions caractéristiques : on peut simplement observer que

$$X = \sum_{j=1}^{n} U_j \qquad \text{et} \qquad Y = \sum_{j=1}^{m} V_j,$$

   et donc

$$Z = \sum_{j=1}^{n} U_j + \sum_{j=1}^{m} V_j,$$

   où les U_j et les V_j sont indépendantes et Bernoulli (p). Donc

$$Z = \sum_{j=1}^{m+n} W_j$$

   où les W_j sont aussi indépendantes et Bernoulli (p). (Les n premières W_j sont les U_j, et les m suivantes sont les V_j.)

4. Soit X et Y indépendantes et respectivement de lois $N(\mu, \sigma^2)$ et $N(\nu, \tau^2)$. Alors $Z = X + Y$ suit la loi normale $N(\mu + \nu, \sigma^2 + \tau^2)$: on écrit

$$\varphi_Z = \varphi_X \varphi_X,$$

ce qui donne

$$\varphi_Z(u) = e^{iu\mu - u^2\sigma^2/2} e^{iu\nu - u^2\tau^2/2} = e^{iu(\mu+\nu) - u^2(\sigma^2+\tau^2)/2}$$

qui est la fonction caractéristique de $N(\mu+\nu, \sigma^2+\tau^2)$, et on utilise encore le théorème 14.1.

5. Soit X et Y des v.a. indépendantes de lois gamma de paramètres respectifs (α, β) et (δ, β) (avec le même β), et $Z = X + Y$. On a $\varphi_Z = \varphi_X \varphi_Y$, et donc

$$\varphi_Z(u) = \frac{\beta^\alpha}{(\beta - iu)^\alpha} \frac{\beta^\delta}{(\beta - iu)^\delta} = \frac{\beta^{\alpha+\delta}}{(\beta - iu)^{\alpha+\delta}},$$

donc Z suit la loi gamma de paramètres $(\alpha + \delta, \beta)$.

6. Dans le chapitre 11 nous avons défini la *loi du chi-deux à p degrés de liberté* (χ^2_p), qui est aussi la loi gamma de paramètres $\frac{p}{2}$ et $\frac{1}{2}$. Nous avons observé que si Z est $N(0,1)$ alors Z^2 suit la loi χ^2_1. Si maintenant $Z_1, \ldots, Z_p$ sont des v.a. indépendantes de loi $N(0,1)$, les v.a. Z_j^2 sont aussi indépendantes, de sorte que l'exemple 5 implique que $X = \sum_{j=1}^p Z_j^2$ suit une loi gamma de paramètres $\frac{p}{2}$ et $\frac{1}{2}$, i.e. la loi χ^2_p. En d'autres termes, une v.a. X de loi du chi-deux à p degrés de liberté peut être vue comme la somme des carrés de p v.a. indépendantes de loi $N(0,1)$.

Exercices

1. Soit $X_1, \ldots, X_n$ des v.a. indépendantes telles que $E(X_j) = \mu$ et $\mathrm{Var}\,(X_j) = \sigma^2 < \infty$, $1 \leqslant j \leqslant n$. Posons

$$\overline{x} = \frac{1}{n}\sum_{j=1}^n X_j \qquad \text{et} \qquad S^2 = \frac{1}{n}\sum_{j=1}^n (X_j - \overline{x})^2.$$

($\overline{x}$ et S^2 sont aussi des v.a., connues sous le nom de « moyenne empirique » et « variance empirique » respectivement.) Montrer que

a) $E(\overline{x}) = \mu$;

b) $\mathrm{Var}\,(\overline{x}) = \frac{\sigma^2}{n}$;

c) $E(S^2) = \frac{n-1}{n}\sigma^2$.

2. Soit $X_1, \ldots, X_n$ des v.a. indépendantes de variances finies. Posons $S_n = \sum_{j=1}^n X_j$. Montrer que

$$\sigma^2_{\frac{1}{n}S_n} = \frac{1}{n^2} \sum_{j=1}^n \sigma^2_{X_j},$$

et en déduire que si $\sigma^2_{X_j} = \sigma^2$ pour $1 \leqslant j \leqslant n$, alors $\sigma^2_{\frac{1}{n}S_n} = \sigma^2/n$.

3. Montrer que si les v.a. $X_1, \ldots, X_n$ sont indépendantes de même loi, alors

$$\varphi_{S_n}(u) = (\varphi_X(u))^n,$$

où $S_n = \sum_{j=1}^n X_j$.

Les exercices 4 à 8 font intervenir la somme d'un nombre aléatoire *de v.a. indépendantes. On se donne une suite infinie* $X_1, X_2, \ldots$ *de v.a. réelles indépendantes de même loi, et une autre v.a.* N *à valeurs entières, indépendante des* X_j. *On pose aussi*

$$S_n = \sum_{j=1}^n X_j, \qquad et \qquad S_N = X_1 + X_2 + \cdots + X_N,$$

avec la convention que $S_N = 0$ *sur l'ensemble où* $N = 0$.

4. Si A est un borélien de $\mathbb{R}$, montrer que

$$P(S_N \in A \mid N = n) = P(S_n \in A).$$

5. Si $E(N) < \infty$ et $E(|X_j|) < \infty$, montrer que

$$E(S_N) = \sum_{n=0}^\infty E(S_n) P(N = n).$$

(*Indication :* Montrer d'abord que

$$E(S_N) = E\left(\sum_{n=0}^\infty S_n 1_{\{N=n\}}\right) = \sum_{n=0}^\infty E\left(S_n 1_{\{N=n\}}\right),$$

en justifiant la seconde égalité.)

6. Si $E(N) < \infty$ et $E(|X_j|) < \infty$, montrer que $E(S_N) = E(N)E(X_j)$. (*Indication :* Utiliser l'exercice 5.)

7. Si $E(N) < \infty$ et $E(|X_j|) < \infty$, montrer que

$$\varphi_{S_N}(u) = E\big((\varphi_{X_j}(u))^N\big).$$

(*Indication :* Montrer d'abord que $\varphi_{S_N}(u) = \sum_{n=1}^{\infty} E(e^{iuS_n} 1_{\{N=n\}})$.)

8. Démontrer le résultat de l'exercice 6 en utilisant l'exercice 7 et les fonctions caractéristiques.

9. Soit X et Y des v.a. réelles indépendantes. Supposons que X et X + Y ont même loi. Montrer que Y = 0 p.s.

10. Soit f et g des fonctions boréliennes sur $\mathbb{R}$, telles que

$$\int_{-\infty}^{\infty} |f(x)|dx < \infty \quad \text{et} \quad \int_{-\infty}^{\infty} |g(x)|dx < \infty.$$

Montrer que

a) $(f * g)(x) = \int_{-\infty}^{\infty} f(x-y)g(y)dy$ existe pour presque tout x (relativement à la mesure de Lebesgue) ;

b) $(f*g)(x) = (g*f)(x)$ ($f*g$ est appelé le *produit de convolution* de f et g) ;

c) Si l'une des deux fonctions f ou g au moins est continue, alors $f*g$ est continue.

11. Soit X et Y des v.a. réelles indépendantes de même loi. Supposons de plus que X + Y et X − Y soient indépendantes. Montrer que $\varphi_X(2u) = \varphi_X(u)^3 \varphi_X(-u)$.

12.* Soit X et Y comme dans l'exercice 11, et supposons que $E(X) = 0$ et $E(X^2) = 1$. Montrer que X suit la loi $N(0,1)$. (*Indication :* Montrer qu'il existe $a > 0$ tel que $\varphi(u) \neq 0$ pour tout u avec $|u| \leqslant a$. Soit $\psi(u) = \frac{\varphi(u)}{\varphi(-u)}$ pour $|u| \leqslant a$. Montrer que $\psi(u) = \{\psi(u/2^n)\}^{2^n}$, puis que $\{\psi(u/2^n)\}^{2^n} \to 1$ quand $n \to \infty$ (cf. exercice 4 du chapitre 14), puis que $\varphi(t) = \{\varphi(t/2^n)\}^{4^n}$, et faire $n \to \infty$.)

13. Soit $X_1, X_2, \ldots, X_n$ des v.a. indépendantes de loi $N(\mu, \sigma^2)$. Soit $\overline{x} = \frac{1}{n}\sum_{j=1}^{n} X_j$ et $Y_j = X_j - \overline{x}$. Trouver la fonction caractéristique de $(\overline{x}, Y_1, \ldots, Y_n)$. En déduire que si $S^2 = \frac{1}{n}\sum_{j=1}^{n} Y_j^2$ les v.a. $\overline{x}$ et S^2 sont indépendantes.

14. Montrer que $|1 - e^{ix}|^2 = 2(1-\cos x) \leqslant x^2$ pour tout $x \in \mathbb{R}$. Montrer que pour toute v.a. réelle X on a $|1 - \varphi_X(u)| \leqslant E(|uX|)$.

15. Soit $A = [-\frac{1}{u}, \frac{1}{u}]$. Montrer que si μ_X est la loi de X, on a

$$\int_A x^2 \mu_X(dx) \leqslant \frac{24}{11u^2}\{1 - \mathrm{Re}\, \varphi_X(u)\}.$$

(*Indication :* $1 - \cos x \geqslant 0$ et $1 - \cos x \geqslant \frac{1}{2}x^2 - \frac{1}{24}x^4$ pour tout $x \in \mathbb{R}$.)

16. Si φ est une fonction caractéristique, montrer que $|\varphi|^2$ en est une aussi (*Indication :* prendre deux v.a. indépendantes X et Y de fonction caractéristique φ, et considérer $Z = X - Y$.)

17. Soit $X_1, \ldots, X_\alpha$ de v.a. indépendantes de loi exponentielle de paramètre $\beta > 0$. Montrer que $Y = \sum_{i=1}^{\alpha} X_i$ suit une loi gamma de paramètres α et β.

Chapitre 16

Variables aléatoires gaussiennes

Rappelons que la loi normale $N(\mu, \sigma^2)$, où $\mu \in \mathbb{R}$ et $\sigma^2 > 0$, est la probabilité sur $\mathbb{R}$ de densité

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}. \tag{16.1}$$

Il est commode d'étendre la classe des lois normales de façon à inclure le cas où $\sigma^2 = 0$, de la manière suivante : on note $N(\mu, 0)$ la loi égale à la masse de Dirac au point μ, i.e. la loi d'une v.a. qui est p.s. égale à la constante μ. Cette loi n'admet pas de densité, aussi parle-t-on parfois de loi normale « dégénérée ». La loi $N(0, 1)$ s'appelle la *loi normale réduite*, ou *standard*.

Si X est une v.a. de loi $N(\mu, \sigma^2)$, sa fonction caractéristique φ_X est

$$\varphi_X(u) = e^{iu\mu - \frac{\sigma^2 u^2}{2}}. \tag{16.2}$$

Quand $\sigma^2 > 0$ cela a fait l'objet de l'exemple 13.6, et quand $\sigma^2 = 0$ c'est un résultat trivial. Rappelons aussi qu'on a alors

$$E(X) = \mu, \qquad \mathrm{Var}\,(X) = \sigma^2. \tag{16.3}$$

À première vue il peut sembler étrange d'appeler des v.a. ayant une densité aussi compliquée des variables *normales*. La raison en remonte au début du XVIII[e] siècle, quand les premières versions du théorème-limite central apparurent dans les livres de Jacob Bernoulli (1713) et A. de Moivre (1718). Ces versions primitives du théorème-limite central ont été généralisées par P. Laplace et surtout C. F. Gauss : ce sont les travaux de ce dernier qui ont valu à ces lois d'être aussi appelées *lois gaussiennes*, et à leur auteur (et à la fonction représentative de la densité normale, dite « courbe de Gauss ») d'apparaître sur l'ancien billet de 10 marks. La version de Gauss du théorème-limite central peut être paraphrasée en écrivant que les sommes de variables i.i.d. (pour « Indépendantes et Identiquement Distribuées », c'est-à-dire de même loi) sont approximativement gaussiennes. C'est un résultat très profond, puisqu'il affirme qu'on n'a besoin de connaître que très peu de choses sur la loi des v.a. dont on fait la somme pour conclure que la loi de la somme est approximativement gaussienne. Finalement, soulignons que plus tard Paul Lévy a déterminé les hypothèses minimales pour que le théorème-limite central soit valide. Ce théorème joue un rôle absolument central en statistique, et c'est ce rôle central qui lui a valu son nom et qui explique pourquoi les variables gaussiennes ou normales sont si importantes et présentes partout en probabilité. Ce théorème sera traité dans le chapitre 21 ; ici, nous en établissons les bases, en étudiant les v.a. gaussiennes.

Pour une v.a. réelle X la définition $\mathcal{L}(\mathrm{X}) = \mathrm{N}(\mu, \sigma^2)$ est claire : c'est une v.a. de densité donnée par (16.1) si $\sigma^2 > 0$, c'est $\mathrm{X} \equiv \mu$ si $\sigma^2 = 0$. Pour une v.a. à valeurs dans $\mathbb{R}^n$ la définition est plus subtile ; la raison en est qu'on veut ainsi décrire la classe des variables qui peuvent être obtenues comme limites dans le théorème-limite central, et cette classe est plus compliquée dans $\mathbb{R}^n$ quand $n \geqslant 2$.

Définition 16.1. *Une variable aléatoire* $\mathrm{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_n)$ *à valeurs dans* $\mathbb{R}^n$ *(ou « vecteur aléatoire ») est dite* gaussienne, *ou* normale multivariée, *si toute combinaison linéaire des composantes, soit* $\sum_{j=1}^n a_j \mathrm{X}_j$ *avec des* $a_j \in \mathbb{R}$ *quelconques, suit une loi normale uni-dimensionnelle (éventuellement dégénérée, par exemple si on prend* $a_j = 0$ *pour tout* j*).*

Les fonctions caractéristiques sont très utiles dans ce contexte.

Théorème 16.1. X *est une variable gaussienne* n*-dimensionnelle si et seulement si sa fonction caractéristique est de la forme*

$$\varphi_{\mathrm{X}}(u) = \exp\{i\langle u, \mu\rangle - \frac{1}{2}\langle u, \mathrm{Q}u\rangle\} \tag{16.4}$$

où $\mu \in \mathbb{R}^n$ *et où* Q *est une matrice* $n \times n$ *symétrique. Dans ce cas le vecteur* μ *est le vecteur moyenne de* X *(i.e.* $\mu_j = \mathrm{E}(\mathrm{X}_j)$ *pour tout* j*), et* Q *est sa matrice de covariance, et est donc semi-définie positive.*

Preuve (Condition suffisante). Supposons qu'on ait (16.4). Soit

$$\mathrm{Y} = \sum_{j=1}^{n} a_j \mathrm{X}_j = \langle a, \mathrm{X}\rangle$$

une combinaison linéaire des composantes de X. Pour tout $v \in \mathbb{R}$ on a

$$\varphi_{\mathrm{Y}}(v) = \varphi_{\mathrm{X}}(va) = \exp\left\{iv\langle a, \mu\rangle - \frac{v^2}{2}\langle a, \mathrm{Q}a\rangle\right\}$$

Remarquer que $\langle a, Qa\rangle \geqslant 0$ puisque $|\varphi_Y(v)| \leqslant 1$ pour tout v. Grâce à (16.2) on voit alors que $\varphi_Y(v)$ est la fonction caractéristique de la loi $N(\langle a, \mu\rangle, \langle a, Qa\rangle)$, donc par le théorème 14.1 on obtient la normalité de Y : cela montre la condition suffisante.

De plus, un calcul des dérivées $\frac{\partial \varphi_X}{\partial u_j}(0)$ et $\frac{\partial^2 \varphi_X}{\partial u_j \partial u_k}(0)$ et une double utilisation du théorème 13.2 montre que $\mu_j = E(X_j)$ et que $E(X_j X_k) = \mu_j \mu_k + Q_{j,k}$, de sorte que μ est le vecteur moyenne de X et Q est sa matrice de covariance.

(Condition nécessaire.) Supposons le vecteur aléatoire X gaussien, et posons de nouveau

$$Y = \sum_{j=1}^{n} a_j X_j = \langle a, X\rangle$$

Soit μ le vecteur moyenne et Q la matrice de covariance de X. On a alors par linéarité

$$E(Y) = \langle a, \mu\rangle$$

tandis que le théorème 12.4 donne

$$\sigma^2(Y) = \langle a, Qa\rangle.$$

Comme Y est normale par hypothèse, une nouvelle application de (16.2) donne

$$\varphi_Y(v) = \exp\left\{iv\langle a, \mu\rangle - \frac{v^2}{2}\langle a, Qa\rangle\right\}.$$

Donc

$$\varphi_Y(1) = \varphi_{\langle a,X\rangle}(1) = E\big(\exp(i\langle a, X\rangle)\big) = \varphi_X(a),$$

et on a (16.4). ■

Notation. Pour X comme dans le théorème précédent, on note $N(\mu, Q)$ sa loi. Elle dépend de deux « paramètres » : le *vecteur moyenne* μ et la *matrice de covariance* Q.

Exemple 1.

Si les X_j pour $j = 1, \ldots, n$ sont des v.a. réelles *indépendantes* de lois $N(\mu_j, \sigma_j^2)$, le vecteur $X = (X_1, \ldots, X_n)$ est gaussien : il suffit de remarquer que

$$\varphi_X(u_1, \ldots, u_n) = \prod_{j=1}^{n} \varphi_{X_j}(u_j)$$

par le corollaire 14.1. donc

$$\begin{aligned}
\varphi_X(u) &= \prod_{j=1}^{n} e^{iu_j\mu_j - u_j^2\sigma_j^2/2} \\
&= \exp\left(\sum_{j=1}^{n} iu_j\mu_j - \frac{1}{2}\sum_{j=1}^{n} u_j^2\sigma_j^2\right) \\
&= e^{i\langle u,\mu\rangle - \frac{1}{2}\langle u,Qu\rangle}
\end{aligned}$$

où $\mu = (\mu_1, \dots, \mu_n)$ et Q est la matrice diagonale

$$\begin{pmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{pmatrix},$$

et on conclut par le théorème 16.1. Cela peut aussi se voir directement : pour tous réels a_j les v.a. a_jX_j sont normales et indépendantes, donc $\sum_{j=1}^n a_jX_j$ est normale (exemple 4 du chapitre 15), et le vecteur X est gaussien.

Une forme de réciproque de cet exemple est vraie :

Corollaire 16.1. *Soit $X = (X_1, \dots, X_n)$ une v.a. gaussienne n-dimensionnelle. Ses composantes X_j sont indépendantes si et seulement si la matrice de covariance Q est diagonale.*

Preuve. La nécessité provient du théorème 12.3. Supposons inversement Q diagonale de la forme

$$Q = \begin{pmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{pmatrix}.$$

Dans ce cas l'équation (16.4) implique que φ_X se factorise :

$$\varphi_X(u) = \prod_{j=1}^{n} \varphi_{X_j}(u_j),$$

où

$$\varphi_{X_j}(u_j) = \exp\left\{iu_j\mu_j - \frac{1}{2}u_j^2\sigma_j^2\right\}.$$

Le corollaire 14.1 entraîne alors l'indépendance des X_j. ∎

Le théorème suivant montre que *tous* les vecteurs aléatoires gaussiens peuvent s'obtenir comme transformées linéaires de vecteurs gaussiens dont les composantes sont indépendantes.

Théorème 16.2. *Soit* X *un vecteur aléatoire n-dimensionnel gaussien de vecteur moyenne* μ. *Il existe alors des v.a. réelles indépendantes* $Y_1, \ldots, Y_n$ *de lois*

$$\mathcal{L}(Y_j) = N(0, \lambda_j) \qquad \textit{avec} \quad \lambda_j \geqslant 0,$$

et une matrice $n \times n$ *orthogonale* A, *telles que* $X = \mu + AY$.

Commentaire important. Certains des λ_j ci-dessus peuvent être nuls, auquel cas $Y_j = 0$. Donc le nombre de v.a. normales Y_j nécessaires dans le théorème 16.2 peut être strictement plus petit que n, même si chaque X_j est non constante. En fait le nombre minimal de v.a. Y_j est égal au rang de la matrice de covariance Q de X.

Preuve du théorème 16.2. Comme Q est symétrique semi-définie positive, il existe une matrice orthogonale A et une matrice diagonale Λ à termes positifs, telles que $Q = A\Lambda A^*$. (Rappelons qu'une *matrice orthogonale* est une matrice A dont les vecteurs lignes (resp. colonnes) sont orthogonaux et de norme 1 : dans ce cas la matrice transposée A^* de A est aussi son inverse.)

Posons

$$Y = A^*(X - \mu).$$

Comme X est gaussien par hypothèse, le vecteur Y est aussi gaussien puisque toute combinaison linéaire des composantes de Y est une combinaison linéaire des composantes de X plus une constante, donc est une v.a. réelle normale. De plus la matrice de covariance de Y est $A^*QA = \Lambda$, la matrice diagonale cherchée. Comme $X = \mu + AY$ (parce que $A^{*-1} = A$), on a prouvé le résultat. ∎

Corollaire 16.2. *Un vecteur aléatoire gaussien n-dimensionnel admet une densité sur* $\mathbb{R}^n$ *si et seulement si sa matrice de covariance* Q *est non dégénérée (i.e., il n'y a pas de vecteur non nul* a *de* $\mathbb{R}^n$ *tel que* $Qa = 0$, *ou de manière équivalente* $\det(Q) \neq 0$*).*

Preuve. Par le théorème précédent il existe des v.a. indépendantes $Y_1, \ldots, Y_n$ de lois $N(0, \lambda_j)$, avec $Q = A\Lambda A^*$ pour une matrice orthogonale A. Si $\det(Q) \neq 0$, on doit avoir $\lambda_j > 0$ pour tout j $(1 \leqslant j \leqslant n)$, parce que $\det(Q) = \det(\Lambda) = \prod_{i=1}^n \lambda_i$. Comme $\lambda_j > 0$ et $\mathcal{L}(Y_j) = N(0, \lambda_j)$, nous savons que le vecteur Y admet la densité

$$f_Y(y) = \prod_{j=1}^n \frac{1}{\sqrt{2\pi\lambda_j}} e^{-y_j^2/2\lambda_j},$$

et comme $X = \mu + AY$, on déduit du théorème 12.7 que X admet la densité

$$f_X(x) = \frac{1}{2\pi^{n/2}} \frac{1}{\sqrt{\det Q}} e^{-\frac{1}{2}\langle (x-\mu), Q^{-1}(x-\mu)\rangle}. \tag{16.5}$$

Supposons à l'inverse Q degénérée : il existe $a \in \mathbb{R}^n$ avec $a \neq 0$ et $Qa = 0$. La v.a. $Z = \langle a, X\rangle$ a la variance $\langle a, Qa\rangle = 0$, donc elle est p.s. égale à sa moyenne $\langle a, \mu\rangle$. Donc $P(X \in H) = 1$, où H est l'hyperplan affine orthogonal à a et contenant le vecteur μ, i.e. $H = \{x \in \mathbb{R}^n ; \langle x - \mu, a\rangle = 0\}$. Comme la dimension de H est $n - 1$, la mesure de Lebesgue n-dimensionnelle de H vérifie $m_n(H)=0$ (cf. exercice 1). Si X avait une densité, on aurait la propriété

$$1 = P(X \in H) = \int_H f(x)dx = \int_H f(x_1, \ldots, x_n)dx_1 \ldots dx_n. \tag{16.6}$$

Cependant

$$\int 1_H(x_1, \ldots, x_n)dx_1 \ldots dx_n = m_n(H) = 0$$

donc (16.6) ne peut pas être vraie et X n'admet pas de densité. ■

Commentaire. Ce corollaire montre que quand $n \geq 2$ il existe des vecteurs aléatoires gaussiens non constants, mais sans densité (si $n = 1$ une v.a. normale est soit constante, soit avec densité). Il serait tentant de définir les v.a. gaussiennes multivariées comme étant les v.a. admettant une densité de la forme (16.5), mais une telle définition serait insuffisante pour couvrir toutes les limites qu'on peut obtenir dans le théorème-limite central du chapitre 21.

Une propriété élémentaire mais importante des v.a. gaussiennes est la suivante :

Théorème 16.3. *Soit X un vecteur gaussien n-dimensionnel, et Y un vecteur gaussien m-dimensionnel indépendant de X. La v.a. $Z = (X, Y)$ à valeurs dans $\mathbb{R}^{n+m}$ est aussi gaussienne.*

Preuve. On a à cause de l'indépendance de X et Y, et par une extension immédiate du corollaire 14.1 :

$$\varphi_Z(u) = \varphi_X(w)\varphi_Y(v), \qquad u = (w, v); \quad w \in \mathbb{R}^n, \quad v \in \mathbb{R}^m.$$

Le théorème 16.1 implique, avec des notations évidentes :

$$\begin{aligned}\varphi_Z(u) &= \exp\left\{i\langle w, \mu^X\rangle - \frac{1}{2}\langle w, Q^X w\rangle\right\}\exp\left\{i\langle v, \mu^Y\rangle - \frac{1}{2}\langle v, Q^Y v\rangle\right\}\\ &= \exp\left\{i\langle (w, v), (\mu^X, \mu^Y)\rangle - \frac{1}{2}\langle u, Qu\rangle\right\},\end{aligned}$$

où

$$Q = \begin{pmatrix} Q^X & 0 \\ 0 & Q^Y \end{pmatrix}.$$

Il suffit d'appliquer de nouveau le théorème 16.1 pour obtenir le résultat. ■

On dit que deux v.a. réelles X et Y de carré intégrable sont *non corrélées* si $\mathrm{Cov}(X, Y) = 0$. Comme

$$\mathrm{Cov}(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y),$$

cela revient à dire que $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$. Cette propriété est vraie si X et Y sont indépendantes (théorème 12.3). La réciproque est *fausse* en général (deux v.a. peuvent être non corrélées sans être indépendantes), mais elle est *vraie* si le couple (X, Y) est gaussien. Plus généralement, on a le résultat suivant, qui est une reformulation partielle du corollaire 16.1 (pour l'obtenir, si $(X_1, \ldots, X_n)$ est un vecteur gaussien chaque couple (X_j, X_k) est aussi un vecteur gaussien bidimensionnel, auquel on peut appliquer ce corollaire) :

Théorème 16.4. *Soit X une vecteur aléatoire gaussien n-dimensionnel, et soit $j \neq k$. Les composantes X_j et X_k sont indépendantes si et seulement si elles sont non corrélées.*

Un modèle fréquemment utilisé en économie, en biologie, en ingénierie, etc., est celui de la *régression linéaire*[1], que nous allons décrire maintenant. On se donne des v.a. réelles Y_i, $1 \leqslant i \leqslant n$, de la forme

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \tag{16.7}$$

1. Le terme « linéaire » décrit la dépendance linéaire en les paramètres α et β et non en les x_i.

où les α, β et x_i sont des constantes, et où les ε_i sont des v.a. réelles. Typiquement, on peut penser qu'on mesure les quantités $\alpha + \beta x_i$ en faisant une erreur de mesure aléatoire ε_i. À cause du théorème-limite central on suppose souvent que le vecteur aléatoire $(\varepsilon_i)_{1 \leqslant i \leqslant n}$ suit une loi normale multivariée. En suivant par exemple Berger et Casella [7] on appelle x_i la *variable prédictrice* (rappelons que x_i n'est pas aléatoire), et Y_i s'appelle la « réponse »[2].

Supposons que $E(\varepsilon_i) = 0$ pour tout i. En prenant les espérances dans (16.7) on obtient

$$E(Y_i) = \alpha + \beta x_i. \tag{16.8}$$

Typiquement on cherche à déterminer la nature de la relation linéaire entre les Y_i et les x_i, qui serait évidente s'il n'y avait pas les perturbations induites par les ε_i. C'est-à-dire que nous cherchons à déterminer α et β en observant les Y_i, sachant que les x_i sont connus. Pour commencer, nous excluons le cas où tous les x_i sont égaux parce qu'alors nous pouvons au mieux déterminer la constante $\alpha + \beta x_1$ sans pouvoir différencier α et β. En d'autres termes, nous supposons que $\sum_{i=1}^{n}(x_i - \overline{x})^2 > 0$, où $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

On peut traiter ces modèles dans un cadre tout à fait général (voir par exemple le chapitre 12 de [7]), mais nous nous limiterons ici au cas le plus important pour les applications, à savoir celui où les « erreurs » sont normales.

Nous allons donc en fait supposer que les v.a. $(\varepsilon_i)_{1 \leqslant i \leqslant n}$ sont indépendantes et de même loi $N(0, \sigma^2)$. Des *estimateurs* U pour α et V pour β sont des v.a. qui dépendent des valeurs observées $Y_1, \ldots, Y_n$ et aussi si l'on veut des valeurs connues $x_1, \ldots, x_n$ et σ^2, mais *pas* des quantités inconnues α et β. Parmi tous les estimateurs possibles, les plus simples sont les *estimateurs linéaires*, qui sont de la forme

$$U = u_0 + \sum_{i=1}^{n} u_i Y_i, \qquad V = v_0 + \sum_{i=1}^{n} v_i Y_i \tag{16.9}$$

pour des constantes $u_0, \ldots, u_n$ et $v_0, \ldots, v_n$. Les estimateurs U et V sont dits *sans biais* s'ils vérifient $E(U) = \alpha$ et $E(V) = \beta$. Comme $E(\varepsilon_i) = 0$, c'est le cas si et seulement si

2. Parfois x_i est appelée la « variable indépendante » et Y_i la « variable dépendante » ; cette terminologie, cause de confusion avec les v.a. indépendantes, ne sera pas utilisée ici.

$$\alpha = \mathrm{E}(\mathrm{U}) = u_0 + \alpha\left(\sum_{i=1}^{n} u_i\right) + \beta\left(\sum_{i=1}^{n} u_i x_i\right) \tag{16.10}$$

$$\beta = \mathrm{E}(\mathrm{V}) = v_0 + \alpha\left(\sum_{i=1}^{n} v_i\right) + \beta\left(\sum_{i=1}^{n} v_i x_i\right). \tag{16.11}$$

Ces équations doivent être satisfaites pour toutes les valeurs possibles de α et β, ce qui est vrai si et seulement si

$$u_0 = 0, \qquad \sum_{i=1}^{n} u_i = 1 \quad \text{et} \quad \sum_{i=1}^{n} u_i x_i = 0. \tag{16.12}$$

$$v_0 = 0, \qquad \sum_{i=1}^{n} v_i = 0 \quad \text{et} \quad \sum_{i=1}^{n} v_i x_i = 1. \tag{16.13}$$

Parmi les estimateurs U (resp. V) de la forme (16.9) et qui satisfont (16.12) (resp. (16.13)), comment en choisir des « bons » ? Une méthode standard consiste à minimiser l'erreur quadratique : i.e. U_1 et V_1 seront considérés comme « meilleurs » que U_2 et V_2 si

$$\mathrm{E}((\mathrm{U}_1-\alpha)^2) \leqslant \mathrm{E}((\mathrm{U}_2-\alpha)^2), \qquad \mathrm{E}((\mathrm{V}_1-\beta)^2) \leqslant \mathrm{E}((\mathrm{V}_2-\beta)^2). \tag{16.14}$$

C'est alors un exercice classique de minimisation sous contraintes de vérifier (ce qui se trouve dans tous les cours de statistique, voir par exemple [7, p. 557-564]) que le choix suivant des coefficients u_i et v_i :

$$v_i = \frac{x_i - \overline{x}}{\sum_{j=1}^{n}(x_j - \overline{x})^2}, \qquad u_i = \frac{1}{n} - \overline{x} v_i \tag{16.15}$$

minimise l'erreur quadratique sous les contraintes de non biais (16.12) et (16.13), et donc donne les meilleurs estimateurs parmi les estimateurs linéaires sans biais. À cause de (16.15) ces estimateurs optimaux ont la forme suivante, où $\overline{\mathrm{Y}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Y}_i$:

$$\mathrm{B} = \sum_{i=1}^{n} \frac{x_i - \overline{x}}{\sum_{j=1}^{n}(x_j - \overline{x})^2}\mathrm{Y}_i, \qquad \mathrm{A} = \overline{\mathrm{Y}} - \overline{x}\mathrm{B}. \tag{16.16}$$

On peut maintenant déterminer la loi des estimateurs A et B dans le cas gaussien. C'est une propriété remarquable du cas gaussien que le couple (A, B) est alors gaussien aussi : la « gaussianité » (si l'on peut dire) est donc préservée, ce qui est extrêmement utile en statistique.

Théorème 16.5. *Supposons les v.a. $(\varepsilon_i)_{1\leqslant i\leqslant n}$ indépendantes et de loi $\mathrm{N}(0,\sigma^2)$, et soit*

$$\mathrm{Y}_i = \alpha + \beta x_i + \varepsilon_i.$$

Les estimateurs A pour α et B pour β donnés par (16.16) forment une v.a. 2-dimensionnelle gaussienne, de moyennes $\mathrm{E}(\mathrm{A}) = \alpha$ et $\mathrm{E}(\mathrm{B}) = \beta$ et de matrice de covariance donnée par

$$\begin{aligned}
\mathrm{Var}\,(\mathrm{A}) &= \frac{\sigma^2}{n\sum_{j=1}^n (x_j-\overline{x})^2}\sum_{j=1}^n x_j^2,\\
\mathrm{Var}\,(\mathrm{B}) &= \frac{\sigma^2}{\sum_{j=1}^n (x_j-\overline{x})^2},\\
\mathrm{Cov}(\mathrm{A},\mathrm{B}) &= \frac{-\sigma^2\overline{x}}{\sum_{j=1}^n (x_j-\overline{x})^2},
\end{aligned}$$

où $\overline{x} = \frac{1}{n}\sum_{j=1}^n x_j$.

Preuve. Comme les ε_i sont indépendantes et normales, elles forment un vecteur gaussien par l'exemple 1. Comme A et B égalent une constante plus une combinaison linéaire des ε_i, le couple (A,B) est aussi gaussien par la définition même des v.a. normales multivariées. Par construction elles sont de moyennes respectives α et β.

Pour toute v.a. U on a $\mathrm{Var}\,(u\mathrm{U}+v) = u^2\mathrm{Var}\,(\mathrm{U})$, de sorte que

$$\begin{aligned}
\mathrm{Var}\,(\mathrm{B}) &= \mathrm{Var}\left(\sum_{i=1}^n v_i\mathrm{Y}_i\right)\\
&= \mathrm{Var}\left(\sum_{i=1}^n v_i\varepsilon_i\right)\\
&= \sum_{i=1}^n v_i^2\mathrm{Var}\,(\varepsilon_i) \qquad\qquad \text{(par l'indépendance des } \varepsilon_i)\\
&= \sigma^2\sum_{i=1}^n\left(\frac{x_i-\overline{x}}{\sum_{j=1}^n (x_j-\overline{x})^2}\right)^2 \qquad \text{(car } \mathrm{Var}\,(\varepsilon_i)=1)\\
&= \frac{\sigma^2}{\sum_{i=1}^n (x_i-\overline{x})^2}.
\end{aligned}$$

De même on a

$$\begin{aligned}
\mathrm{Var}(\mathrm{A}) &= \sum_{i=1}^{n} v_i^2 \,\mathrm{Var}(\varepsilon_i) \\
&= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} + \overline{x}^2 v_i^2 - 2\frac{\overline{x} v_i}{n} \right) \\
&= \sigma^2 \left(\frac{1}{n} + \overline{x}^2 \sum_{i=1}^{n} v_i^2 \right) \\
&= \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}. \\
\mathrm{Cov}(\mathrm{A},\mathrm{B}) &= \mathrm{Cov}\left(\sum_{i=1}^{n} u_i \varepsilon_i, \sum_{i=1}^{n} v_i \varepsilon_i \right) \\
&= \sum_{i=1}^{n} u_i v_i \,\mathrm{Var}(\varepsilon_i) \\
&= \sigma^2 \sum_{i=1}^{n} \left(\frac{v_i}{n} - \overline{x} v_i^2 \right) = -\sigma^2 \overline{x} \sum_{i=1}^{n} v_i^2 \\
&= -\overline{x}\,\mathrm{Var}(\mathrm{B})
\end{aligned}$$

■

Terminons ce chapitre par un exemple qui sert aussi d'avertissement : X peut être une v.a. à valeurs dans $\mathbb{R}^n$ dont toutes les composantes sont normales, mais qui n'est pas un vecteur gaussien. Donc, *la propriété d'être un vecteur gaussien est plus forte que le fait que chaque composante soit gaussienne.*

Exemple 2. Soit Y une v.a. N(0, 1), et posons pour un $a > 0$

$$\mathrm{Z} = \mathrm{Y}1_{\{|\mathrm{Y}| \leqslant a\}} - \mathrm{Y}1_{\{|\mathrm{Y}| > a\}}.$$

La loi de Z est aussi N(0, 1) (voir l'exercice 2), mais $\mathrm{Y}+\mathrm{Z} = 2\mathrm{Y}1_{\{|\mathrm{Y}| \leqslant a\}}$ n'est pas une v.a. normale puisque (par exemple) $\mathrm{P}(\mathrm{Y}+\mathrm{Z} > 2a) = 0$ et Y + Z n'est pas p.s. égale à une constante. Donc X = (Y, Z) est une v.a. à valeurs dans $\mathbb{R}^2$ qui n'est pas gaussienne, bien que chacune de ses deux composantes le soient.

Cela vaut la peine de souligner encore un certain nombre de propriétés qu'ont les vecteurs gaussiens, et qui ne sont pas vraies en général pour les vecteurs aléatoires non gaussiens :

1. Les composantes sont indépendantes si et seulement si elles sont non corrélées.

2. Les composantes sont normales, donc appartiennent à la même famille de lois que la variable vectorielle.
3. Un vecteur gaussien X de loi $N(\mu, Q)$ avec Q inversible peut être transformé linéairement en un vecteur de loi $N(0, I)$ (où I est la matrice identité de même dimension que X) (cf. exercice 6) : les transformations linéaires préservent la famille de lois.
4. La densité existe si et seulement si la matrice de covariance est inversible, ce qui donne un critère simple d'existence pour la densité
5. Les lois conditionnelles des vecteurs gaussiens sont encore des lois gaussiennes (voir l'exercice 10).

Ces six propriétés montrent une stabilité remarquable de la famille des lois gaussiennes. Il existe encore bien d'autres propriétés spécifiques de ces lois, que nous n'évoquons pas ici.

Il est toutefois intéressant de remarquer que les lois normales ne se rencontrent pas réellement dans la nature, alors qu'elles interviennent via un passage à la limite (par le théorème-limite central) et constituent donc une approximation de la réalité, souvent même une excellente approximation. Quand on dit par exemple que la taille des hommes de 20 ans en France est distribuée selon une loi normale de moyenne μ et de variance σ^2, on veut en fait dire que les tailles sont approximativement distribuées selon cette loi normale : en effet si le modèle normal était rigoureusement exact, il y aurait une probabilité strictement positive de trouver un homme plus grand que la tour Eiffel, ou de taille négative, ce qui est évidemment absurde. Cependant ces probabilités, bien que strictement positives, sont si petites qu'elles en sont négligeables : pour tous les calculs concrets l'approximation normale est excellente, alors que la « vraie » loi est évidemment inconnue.

Exercices

1. Soit $a \in \mathbb{R}^n$ avec $a \neq 0$, et $\mu \in \mathbb{R}^n$. Soit H l'hyperplan de $\mathbb{R}^n$ défini par
$$\mathrm{H} = \{x \in \mathbb{R}^n : \langle x - \mu, a \rangle = 0\}.$$
Montrer que $m_n(\mathrm{H}) = 0$ où m_n est la mesure de Lebesgue sur $\mathbb{R}^n$, et en déduire que
$$\int_{\mathrm{H}} f(x)dx = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) 1_{\mathrm{H}}(x_1, \dots, x_n) dx_1 \dots dx_n = 0$$
pour toute fonction borélienne f sur $\mathbb{R}^n$.

2. Soit Y de loi $N(0,1)$ et $a > 0$. Posons
$$Z = \begin{cases} Y & \text{si } |Y| \leqslant a, \\ -Y & \text{si } |Y| > a. \end{cases}$$
Montrer que Z est aussi de loi $N(0,1)$.
3. Soit X une v.a. $N(0,1)$ et soit Z indépendante de X avec $P(Z = 1) = P(Z = -1) = \frac{1}{2}$. Soit $Y = ZX$. Montrer que Y est de loi $N(0,1)$, mais que (X,Y) n'est pas gaussien.
4. Soit (X,Y) un vecteur gaussien bidimensionnel. On suppose que les variances de X et Y, soit σ_X^2 et σ_Y^2, sont non nulles, et on note ρ le coefficient de corrélation de X et Y, et μ_X et μ_Y leurs moyennes. Montrer que si $-1 < \rho < 1$ la densité de (X,Y) existe et égale
$$f_{(X,Y)}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{\frac{-1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - \frac{2\rho(x-\mu_1)(y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right\}.$$
Montrer que si $\rho = -1$ ou $\rho = 1$, la v.a. (X,Y) n'admet pas de densité.
5. Soit $\rho \in [-1,1]$, $\mu_1, \mu_2 \in \mathbb{R}$ et $\sigma_1 > 0$, $\sigma_2 > 0$. Construire des v.a. X_1 et X_2 normales de moyennes μ_1 et μ_2, de variances σ_1^2 et σ_2^2, et de coefficient de corrélation ρ. (*Indication :* Soit Y_1 et Y_2 des v.a. indépendantes $N(0,1)$ et poser $U_1 = Y_1$ et $U_2 = \rho Y_1 + \sqrt{1-\rho^2}Y_2$, puis $X_j = \mu_j + \sigma_j U_j$ $(j = 1,2)$.)
6. Supposons X gaussien $N(\mu, Q)$ sur $\mathbb{R}^n$, avec $\det(Q) > 0$. Montrer qu'il existe une matrice B telle que $Y = B(X - \mu)$ soit de loi $N(0, I)$, où I est la matrice identité $n \times n$. *(Cela montre qu'un vecteur gaussien non dégénéré peut se transformer linéairement en une vecteur gaussien « standard ».)*
7. Soit $X = (X_1, \dots, X_n)$ un vecteur gaussien, et
$$Y = \sum_{j=1}^{n} a_j X_j,$$
Montrer que Y est $N(\mu, \sigma^2)$ avec
$$\mu = \sum_{j=1}^{n} a_j E(X_j)$$

et

$$\sigma^2 = \sum_{j=1}^{n} a_j^2 \mathrm{Var}\,(X_j) + 2\sum_{j<k} a_j a_k \,\mathrm{Cov}(X_j, X_k).$$

8. Soit (X, Y) bidimensionnelle de loi $N(\mu, Q)$, où

$$Q = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

et supposons que $\det(Q) > 0$, donc (X, Y) admet une densité. Montrer que la densité conditionnelle $f_{X=x}$ est celle de la loi $N\Big(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X), \sigma_Y^2(1-\rho^2)\Big)$. (Cf. le théorème 12.2 ou l'exercice 3 du chapitre 12.)

9. Soit X une v.a. bidimensionnelle de loi $N(\mu, Q)$ avec $\mu = (1, 1)$ et $Q = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$. Trouver la loi conditionnelle de $Y = X_1 + X_2$ sachant $Z = X_1 - X_2 = 0$.

$$\left[\text{Rép. : } f_{Z=0}(y) = \frac{1}{\sqrt{2\pi}\sqrt{\frac{20}{3}}} \exp\left\{-\frac{1}{2}\frac{(y-2)^2}{\frac{20}{3}}\right\}.\right]$$

10. Soit $X = (X_1, \ldots, X_n)$ de loi $N(\mu, Q)$ avec $\det(Q) > 0$. Montrer que la loi conditionnelle de $(X_1, \ldots, X_m)$ sachant $(X_{m+1} = x_{m+1}, \ldots, X_n = x_n)$ (avec $1 \leqslant m < n$) est encore une loi gaussienne. [Cela généralise l'exercice 8.]

11. (Gut, 1995.) Soit (X, Y) une v.a. de densité sur $\mathbb{R}^2$:

$$f_{(X,Y)}(x, y) = c\, e^{-(1+x^2)(1+y^2)},$$

où c est choisi de manière à ce que f soit effectivement une densité. Montrer que le couple (X, Y) n'est pas gaussien, mais que les densités conditionnelles $f_{X=x}$ et $f_{Y=y}$ sont des densités gaussiennes (cela montre que la réciproque de l'exercice 10 est fausse).

12. Soit (X, Y) une v.a. bidimensionnelle gaussienne de moyenne nulle, et soit ρ le coefficient de corrélation de X et Y. Montrer que si $|\rho| < 1$, la variable $Z = \frac{X}{Y}$ suit une loi de Cauchy de paramètres $\alpha = \rho\frac{\sigma_X}{\sigma_Y}$ et $\beta = \frac{\sigma_X}{\sigma_Y}\sqrt{1-\rho^2}$. (Remarque : ce résultat a déjà été établi lorsque X et Y sont indépendantes).

13.* Soit (X, Y) une v.a. bidimensionnelle gaussienne de moyenne nulle, et soit ρ le coefficient de corrélation de X et Y. Soit β tel que

$$\cos\beta = \rho \qquad (0 \leqslant \beta \leqslant \pi).$$

Montrer que

$$P(XY < 0) = \frac{\beta}{\pi}.$$

(*Indication :* Se rappeler que, d'après l'exercice 12, si $Z = \frac{X}{Y}$ et $z = \rho\frac{\sigma_X}{\sigma_Y}$, on a

$$F_Z(z) = \frac{1}{2} + \frac{1}{\pi}\operatorname{Arctan}\left(\frac{z\sigma_Y - \rho\sigma_X}{\sigma_X\sqrt{1-\rho^2}}\right).$$

Soit $\alpha = \operatorname{Arcsin}\rho$ $(-\frac{\pi}{2} \leqslant \alpha \leqslant \frac{\pi}{2})$; montrer d'abord que $P(XY < 0) = \frac{1}{2} - \frac{\alpha}{\pi}$, en utilisant que $\operatorname{Arctan}\frac{\rho}{\sqrt{1-\rho^2}} = \operatorname{Arcsin}\rho$.)

14. Soit (X, Y), α et ρ comme dans l'exercice 13. Montrer que

$$P(X > 0, Y > 0) = P(X < 0, Y < 0) = \frac{1}{4} + \frac{\alpha}{2\pi};$$
$$P(X > 0, Y < 0) = P(X < 0, Y > 0) = \frac{1}{4} - \frac{\alpha}{2\pi}.$$

15.* Soit (X, Y) une v.a. bidimensionnelle de densité

$$f_{(X,Y)}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2} - \frac{2\rho xy}{\sigma_X\sigma_Y} + \frac{y^2}{\sigma_Y^2}\right)}.$$

Montrer que

a) $E(XY) = \rho\sigma_X\sigma_Y$

b) $E(X^2Y^2) = E(X^2)E(Y^2) + 2(E(XY))^2$

c) $E(|XY|) = \frac{2\sigma_X\sigma_Y}{\pi}(\cos\alpha + \alpha\sin\alpha)$ où $\alpha = \operatorname{Arcsin}\rho$ $(-\frac{\pi}{2} \leqslant \alpha \leqslant \frac{\pi}{2})$ (cf. exercice 13).

16. Soit (X, Y) une v.a. gaussienne bidimensionnelle avec $\sigma_X^2 = \sigma_Y^2 > 0$, et soit ρ le coefficient de corrélation de X et Y. Montrer que X et $Y - \rho X$ sont indépendantes.

17. Soit X de loi $N(\mu, Q)$ sur $\mathbb{R}^n$, avec $\det(Q) > 0$. Montrer que

$$(X-\mu)^* Q^{-1}(X-\mu) \quad \text{suit la loi } \chi_n^2.$$

18. Soit $X_1, \dots, X_n$ des v.a. indépendantes de loi $N(0, \sigma^2)$, et posons

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} X_j \ , \quad S^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{x})^2 .$$

(Rappelons que d'après l'exercice 13 du chapitre 15 les v.a. $\overline{x}$ et S^2 sont indépendantes.) Montrer que

$$\sum_{j=1}^{n} X_j^2 = \sum_{j=1}^{n}(X_j - \overline{x})^2 + n\overline{x}^2$$

et en déduire que $(n-1)S^2/\sigma^2$ admet la loi χ^2_{n-1} et que $n\overline{x}^2/\sigma^2$ admet la loi χ^2_1.

19. Soit $\varepsilon_1, \dots, \varepsilon_n$ des v.a. indépendantes de loi $N(0, \sigma^2)$ et soit $Y_i = \alpha + \beta x_i + \varepsilon_i$. Supposons aussi que les x_i ne sont pas tous égaux, et posons $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. On définit les *résidus de régression* par

$$\widehat{\varepsilon}_i = Y_i - A - Bx_i,$$

où A et B sont donnés par (16.16).

a) Montrer que $E(\widehat{\varepsilon}_i) = 0$.

b)* Montrer que

$$\mathrm{Var}\,(\widehat{\varepsilon}_i) = \sigma^2 \left(\frac{n-1}{n} - \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right).$$

20. On se place dans la situation de l'exercice précédent. Supposons σ inconnu, et posons

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varepsilon}_i^2 .$$

Montrer que $E(\widehat{\sigma}^2) = \frac{n-2}{n}\sigma^2$. (On a $E(\widehat{\sigma}^2) \neq \sigma^2$, et $\widehat{\sigma}^2$ est appelé un estimateur *biaisé* de σ^2 ; un estimateur non biaisé serait $S^2 = \frac{n}{n-2}\widehat{\sigma}^2$.)

21. Toujours dans la situation des deux exercices précédents, montrer que les v.a. (A,B) et $\widehat{\sigma}^2$ sont indépendantes.

Chapitre 17

Convergence des variables aléatoires

Dans les cours élémentaires de mathématiques on introduit la convergence des suites de fonctions : si $f_n : \mathbb{R} \to \mathbb{R}$, alors $\lim_{n\to\infty} f_n = f$ si $\lim_{n\to\infty} f_n(x) = f(x)$ pour tout $x \in \mathbb{R}$, et on dit alors que la suite (f_n) *converge simplement* vers f. Une v.a. est bien sûr une fonction ($X : \Omega \to \mathbb{R}$ pour un espace « abstrait » Ω), et nous avons donc la même notion : une suite $X_n : \Omega \to \mathbb{R}$ *converge simplement* vers X si $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ pour tout $\omega \in \Omega$. Cette définition naturelle est, de manière surprenante, à peu près inutile en probabilités : l'exemple suivant explique pourquoi.

Exemple 1. Soit X_n une suite de v.a. qui sont i.i.d. (indépendantes et identiquement distribuées, i.e. de même loi), avec $P(X_n = 1) = p$ et $P(X_n = 0) = 1 - p$. Par exemple on peut imaginer qu'elles donnent les résultats successifs quand on lance (une infinité de fois) une pièce de monnaie, $X_n = 1$ (resp. $X_n = 0$) traduisant le fait que le n-ième tirage donne face (resp. pile) ; si $p \neq 1/2$ cela veut dire que la pièce est truquée. Lorsque n est grand, on s'attend à ce que la proportion de faces soit à peu près égale à p (c'est l'essence de « l'approche par les fréquences » en probabilités). Mathématiquement, on voudrait que

$$\lim_{n\to\infty} \frac{X_1(\omega) + \cdots + X_n(\omega)}{n} = p \qquad \text{pour tout } \omega \in \Omega.$$

Cela est tout simplement faux : par exemple si $\omega_0 = \{f, f, f, \ldots\}$ est la suite ne contenant que des faces, on a

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X_j(\omega_0) = 0.$$

Plus généralement $\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X_j(\omega) = 0$ pour tout ω dans l'ensemble $A = \{\omega : \text{il n'y a qu'un nombre fini de faces}\}$. Plus généralement encore, on peut trouver des ω pour lesquels la fréquence converge vers n'importe quel nombre fixé dans $[0, 1]$, et d'autres pour lesquels la fréquence ne converge pas. D'ailleurs si on remplace p par un autre nombre $q \in [0, 1]$, on change la probabilité, mais on ne change ni Ω, ni les fonctions $\frac{1}{n}(X_1 + \cdots + X_n)$, alors qu'on souhaiterait que ces fonctions convergent vers p dans le premier cas, et vers q dans le second.

Bien entendu l'événement A est plutôt invraisemblable, et on peut en fait montrer (voir exercice 14) qu'il vérifie $P(A) = 0$. Ce qu'on verra plus tard (voir la « loi des grands nombres » au chapitre 20) est que

$$P\left(\left\{\omega : \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} X_j(\omega) = p\right\}\right) = 1.$$

Ce type de convergence, pour lequel on n'a pas convergence pour *tout* ω, mais seulement pour *presque tout* ω, est typiquement ce qui arrive pour des variables aléatoires.

Avertissement. Dans tout ce chapitre les v.a. qui interviennent sont à valeurs réelles (sauf dans la dernière remarque) et sont toutes définies sur le même espace de probabilité $(\Omega, \mathcal{A}, P)$.

Rappelons d'abord, sous la forme d'une définition, la notion de convergence p.s. d'une suite de v.a., qui a déjà été utilisée dans le théorème 9.1 par exemple.

Définition 17.1. *On dit qu'une suite* $(X_n)_{n\geqslant 1}$ *de v.a.* converge presque sûrement *vers une v.a.* X *si l'ensemble* N *des* ω *tels que la suite numérique* $(X_n(\omega))$ *ne converge pas vers* $X(\omega)$ *vérifie* $P(N) = 0$. *Rappelons que l'ensemble* N *est alors dit « négligeable ».*

Noter qu'on a aussi

$$N^c = \Lambda = \left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} \quad \text{vérifie } P(\Lambda) = 1.$$

On note la convergence presque sûre ainsi :

$$\lim_{n\to\infty} X_n = X \text{ p.s.} \qquad \text{ou} \qquad X_n \to X \text{ p.s.}$$

Il apparaît que même la convergence p.s. est trop forte dans certaines situations, et nous introduisons ci-dessous deux autres modes de convergence.

Définition 17.2. *On dit qu'une suite* $(X_n)_{n\geqslant 1}$ *de v.a.* converge dans L^p *(où* $p \in [1, \infty[$*) vers une v.a.* X *si les* X_n *et* X *sont dans* L^p, *et si on a*

$$\lim_{n\to\infty} E(|X_n - X|^p) = 0.$$

On dit aussi que X_n *converge en moyenne d'ordre* p *vers* X, *et on écrit*

$$X_n \xrightarrow{L^p} X.$$

Les cas les plus importants sont $p = 1$ (convergence « en moyenne ») et $p = 2$ (convergence « en moyenne quadratique »). Quand $p = 1$, on a $\left|E(X_n - X)\right| \leqslant E(|X_n - X|)$ et $\left|E(|X_n|) - E(|X|)\right| \leqslant E(|X_n - X|)$ parce que $\left||x| - |y|\right| \leqslant |x - y|$. Donc

$$X_n \xrightarrow{L^1} X \quad \text{implique} \quad E(X_n) \to E(X) \quad \text{et} \quad E(|X_n|) \to E(|X|). \qquad (17.1)$$

De même, quand $X_n \xrightarrow{L^p} X$ pour $p \in]1, \infty[$, on peut montrer que $E(|X_n|^p) \to E(|X|^p)$: voir l'exercice 15 pour le cas $p = 2$.

Définition 17.3. *On dit qu'une suite* $(X_n)_{n \geqslant 1}$ *de v.a.* converge en probabilité *vers une v.a.* X *si pour tout* $\varepsilon > 0$ *on a*

$$\lim_{n\to\infty} P(\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}) = 0,$$

ou plus simplement : $P(|X_n - X| > \varepsilon) \to 0$ *quand* $n \to \infty$. *On écrit alors*

$$X_n \xrightarrow{P} X.$$

Une autre manière d'écrire les choses est la suivante : X_n converge en probabilité vers X lorsque, pour tous $\varepsilon > 0$ et $\delta > 0$ il existe $N = N(\delta, \varepsilon)$ tel que

$$n \geqslant N \qquad \Rightarrow \qquad P(|X_n - X| > \varepsilon) < \delta.$$

Avant d'établir les rapports entre les différents modes de convergence on va donner un critère simple de convergence en probabilité.

Théorème 17.1. *On a* $X_n \xrightarrow{P} X$ *si et seulement si*

$$\lim_{n\to\infty} E\left(\frac{|X_n - X|}{1 + |X_n - X|}\right) = 0.$$

Preuve. On ne nuit pas à la généralité en supposant que $X = 0$, de sorte qu'il nous suffit de montrer que $X_n \xrightarrow{P} 0$ équivaut à $\lim_{n\to\infty} E\left(\frac{|X_n|}{1+|X_n|}\right) = 0$. Supposons d'abord que $X_n \xrightarrow{P} 0$. Pour tout $\varepsilon > 0$, on a $\lim_{n\to\infty} P(|X_n| > \varepsilon) = 0$. Mais

$$\frac{|X_n|}{1 + |X_n|} \leqslant \frac{|X_n|}{1 + |X_n|} 1_{\{|X_n| > \varepsilon\}} + \varepsilon 1_{\{|X_n| \leqslant \varepsilon\}} \leqslant 1_{\{|X_n| > \varepsilon\}} + \varepsilon,$$

donc

$$\mathrm{E}\left(\frac{|\mathrm{X}_n|}{1+|\mathrm{X}_n|}\right) \leqslant \mathrm{E}\left(1_{\{|\mathrm{X}_n|>\varepsilon\}}\right) + \varepsilon = \mathrm{P}(|\mathrm{X}_n| > \varepsilon) + \varepsilon.$$

En prenant les limites on arrive à

$$\limsup_{n\to\infty} \mathrm{E}\left(\frac{|\mathrm{X}_n|}{1+|\mathrm{X}_n|}\right) \leqslant \varepsilon\,;$$

comme ε est arbitrairement petit il vient $\lim_{n\to\infty} \mathrm{E}(\frac{|\mathrm{X}_n|}{1+|\mathrm{X}_n|}) = 0$.

Supposons à l'inverse que $\lim_{n\to\infty} \mathrm{E}(\frac{|\mathrm{X}_n|}{1+|\mathrm{X}_n|}) = 0$. La fonction $f(x) = \frac{x}{1+x}$ est strictement croissante, donc

$$\frac{\varepsilon}{1+\varepsilon} 1_{\{|\mathrm{X}_n|>\varepsilon\}} \leqslant \frac{|\mathrm{X}_n|}{1+|\mathrm{X}_n|} 1_{\{|\mathrm{X}_n|>\varepsilon\}} \leqslant \frac{|\mathrm{X}_n|}{1+|\mathrm{X}_n|}.$$

En prenant les espérances, puis les limites, on arrive à

$$\frac{\varepsilon}{1+\varepsilon} \lim_{n\to\infty} \mathrm{P}(|\mathrm{X}_n| > \varepsilon) \leqslant \lim_{n\to\infty} \mathrm{E}\left(\frac{|\mathrm{X}_n|}{1+|\mathrm{X}_n|}\right) = 0.$$

Comme $\varepsilon > 0$ est fixé, on conclut que $\lim_{n\to\infty} \mathrm{P}(|\mathrm{X}_n| > \varepsilon) = 0$. ■

Remarque. Ce théorème nous dit que $\mathrm{X}_n \xrightarrow{\mathrm{P}} \mathrm{X}$ si et seulement si $\mathrm{E}(f(|\mathrm{X}_n - \mathrm{X}|)) \to 0$ pour la fonction $f(x) = \frac{|x|}{1+|x|}$. Un examen soigneux de sa preuve montre que la même équivalence reste vraie pour toute fonction f sur $[0,\infty[$ qui est bornée, strictement croissante, continue et nulle en 0. Par exemple $\mathrm{X}_n \xrightarrow{\mathrm{P}} \mathrm{X}$ équivaut à $\mathrm{E}(|\mathrm{X}_n - \mathrm{X}| \wedge 1) \to 0$, et aussi à $\mathrm{E}(\mathrm{Arctan}\,|\mathrm{X}_n - \mathrm{X}|) \to 0$.

Le théorème suivant nous indique que la convergence en probabilité est le mode de convergence le plus faible introduit jusqu'à présent.

Théorème 17.2. *Soit $(\mathrm{X}_n)_{n\geqslant 1}$ une suite de v.a.*

(a) *Si $\mathrm{X}_n \xrightarrow{\mathrm{L}^p} \mathrm{X}$, on a $\mathrm{X}_n \xrightarrow{\mathrm{P}} \mathrm{X}$.*

(b) *Si $\mathrm{X}_n \to \mathrm{X}$ p.s., on a $\mathrm{X}_n \xrightarrow{\mathrm{P}} \mathrm{X}$.*

Preuve. (a) En se rappelant que pour tout événement A on a $\mathrm{P}(\mathrm{A}) = \mathrm{E}(1_\mathrm{A})$, où 1_A est la fonction indicatrice de A, on voit que

$$\mathrm{P}(|\mathrm{X}_n - \mathrm{X}| > \varepsilon) = \mathrm{E}\left(1_{\{|\mathrm{X}_n - \mathrm{X}|>\varepsilon\}}\right).$$

Mais $\frac{|X_n - X|^p}{\varepsilon^p} > 1$ sur l'événement $\{|X_n - X| > \varepsilon\}$, donc

$$\begin{aligned}
P(|X_n - X| > \varepsilon) &\leqslant E\left(\frac{|X_n - X|^p}{\varepsilon^p} 1_{\{|X_n - X| > \varepsilon\}}\right) \\
&= \frac{1}{\varepsilon^p} E\left(|X_n - X|^p 1_{\{|X_n - X| > \varepsilon\}}\right),
\end{aligned}$$

et comme $|X_n - X|^p \geqslant 0$ partout, on peut simplement omettre la fonction indicatrice, ce qui donne

$$P(|X_n - X| > \varepsilon) \leqslant \frac{1}{\varepsilon^p} E(|X_n - X|^p).$$

Cette quantité tend vers 0 quand $n \to \infty$ (pour $\varepsilon > 0$ fixé), d'où le résultat.

(b) On a $\frac{|X_n - X|}{1+|X_n - X|} \leqslant 1$, donc

$$\lim_{n\to\infty} E\left(\frac{|X_n - X|}{1 + |X_n - X|}\right) = E\left(\lim_{n\to\infty} \frac{|X_n - X|}{1 + |X_n - X|}\right) = E(0) = 0$$

grâce au théorème de convergence dominée de Lebesgue 9.1(f). Il nous suffit alors d'appliquer le théorème 17.1. ■

La réciproque du théorème 17.2 est fausse : il existe cependant deux « réciproques » partielles. La plus délicate concerne la convergence p.s., et s'énonce comme suit :

Théorème 17.3. *Supposons que* $X_n \xrightarrow{P} X$. *Il existe alors une sous-suite* n_k *de la suite des entiers, telle que* $X_{n_k} \to X$ *p.s.*

Preuve. On a $\lim_{n\to\infty} E\left(\frac{|X_n - X|}{1+|X_n - X|}\right) = 0$ par le théorème 17.1. On peut donc trouver une suite croissante d'entiers n_k telle que $E\left(\frac{|X_{n_k} - X|}{1+|X_{n_k} - X|}\right) < \frac{1}{2^k}$. On a alors $\sum_{k=1}^{\infty} E\left(\frac{|X_{n_k} - X|}{1+|X_{n_k} - X|}\right) < \infty$, et le théorème 9.2 implique que $\sum_{k=1}^{\infty} \frac{|X_{n_k} - X|}{1+|X_{n_k} - X|} < \infty$ p.s. ; puisque le terme général d'une série convergente tend vers 0, on en déduit que

$$\lim_{n\to n} |X_{n_k} - X| = 0 \quad \text{p.s.}$$ ■

Remarque 17.1. Ce résultat peut aussi être démontré de manière assez simple en utilisant le lemme de Borel-Cantelli (théorème 10.5).

Exemple 2. La convergence en probabilité n'implique pas la convergence presque sûre. Prenons par exemple $\Omega =]0,1]$, avec la tribu borélienne $\mathcal{A}$ et la probabilité uniforme P. Si A_n est un sous-intervalle de Ω de longueur a_n et si $X_n = 1_{A_n}$, on a $P(|X_n| > \varepsilon) = a_n$ pour $\varepsilon \in]0,1[$; donc, dès que $a_n \to 0$, on a $X_n \xrightarrow{P} 0$. Soit alors $X_{n,j}$ la fonction indicatrice de l'intervalle $]\frac{j-1}{n}, \frac{j}{n}]$, pour $1 \leqslant j \leqslant n$ et $n \geqslant 1$. On peut réordonner la « double » suite $X_{n,j}$ selon l'ordre lexicographique (choisir d'abord n, puis j), ce qui conduit à une nouvelle suite Y_n, dont les premiers termes sont :

$$\begin{array}{cccccccc} X_{1,1} & X_{2,1} & X_{2,2} & X_{3,1} & X_{3,2} & X_{3,3} & X_{4,1} & \dots \\ Y_1 & Y_2 & Y_3 & Y_4 & Y_5 & Y_6 & Y_7 & \dots \end{array}$$

Pour tout ω et tout $n \geqslant 2$ il existe un j tel que $X_{n,j}(\omega) = 1$ et un autre tel que $X_{n,j}(\omega) = 0$: donc $\limsup_{m\to\infty} Y_m(\omega) = 1$ et $\liminf_{m\to\infty} Y_m(\omega) = 0$, et la suite Y_n ne converge pour aucune valeur de ω. Cependant Y_n est l'indicatrice d'un intervalle dont la longueur tend vers 0, donc $Y_n \xrightarrow{P} 0$.

Théorème 17.4. *Supposons que $X_n \xrightarrow{P} X$ et que $|X_n| \leqslant Y$ pour tout n, où Y est une v.a. appartenant à L^p pour un $p \in [1,\infty[$. On a alors $X \in L^p$ et $X_n \xrightarrow{L^p} X$.*

Preuve. Comme $E(|X_n|^p) \leqslant E(Y^p) < \infty$, on a $X_n \in L^p$. Pour tout $\varepsilon > 0$ on a aussi

$$\begin{aligned}\{|X_n| > Y + \varepsilon\} &\subset \{|X| > |X_n| + \varepsilon\} \\ &\subset \{|X| - |X_n| > \varepsilon\} \\ &\subset \{|X - X_n| > \varepsilon\},\end{aligned}$$

donc

$$P(|X| > Y + \varepsilon) \leqslant P(|X - X_n| > \varepsilon),$$

et comme ceci est vrai pour chaque n, il vient

$$P(|X| > Y + \varepsilon) \leqslant \lim_{n\to\infty} P(|X - X_n| > \varepsilon) = 0,$$

par hypothèse. Ceci étant vrai pour tout $\varepsilon > 0$, on en déduit que

$$P(|X| > Y) \leqslant \lim_{m\to\infty} P\Big(|X| > Y + \frac{1}{m}\Big) = 0,$$

qui entraîne $|X| \leqslant Y$ p.s., et donc $X \in L^p$.

Supposons maintenant que X_n ne converge pas vers X dans L^p. Il existe alors une suite d'entiers strictement croissante n_k et un $\varepsilon > 0$

tels que $E(|X_{n_k} - X|^p) \geqslant \varepsilon$ pour tout k. La sous-suite X_{n_k} converge évidemment vers X en probabilité, donc par le théorème 17.3 elle admet une « sous-sous-suite » $X_{n_{k_j}}$ qui converge p.s. vers X. Donc les v.a. $X_{n_{k_j}} - X$ convergent p.s. vers 0, tout en restant majorées en valeur absolue par 2Y : une application du théorème de Lebesgue donne alors que $E(|X_{n_{k_j}} - X|^p) \to 0$, en contradiction avec le fait que $E(|X_{n_k} - X|^p) \geqslant \varepsilon$ pour tout k. ∎

Le théorème suivant est facile, mais utile.

Théorème 17.5. *Soit f une fonction continue.*

(a) *Si $X_n \to X$ p.s., alors $f(X_n) \to f(X)$ p.s.*

(b) *Si $X_n \xrightarrow{P} X$, alors $f(X_n) \xrightarrow{P} f(X)$.*

Preuve. (a) Soit N l'ensemble introduit dans la définition 17.1. On a $P(N) = 0$ par hypothèse. Si $\omega \notin N$, il vient

$$\lim_{n\to\infty} f(X_n(\omega)) = f\left(\lim_{n\to\infty} X_n(\omega)\right) = f(X(\omega)),$$

où la première égalité provient de la continuité de f : le résultat en découle.

(b) Pour chaque $k > 0$ on a :

$$\{|f(X_n) - f(X)| > \varepsilon\} \subset \{|f(X_n) - f(X)| > \varepsilon, |X| \leqslant k\} \cup \{|X| > k\}. \tag{17.2}$$

Comme f est continue, elle est uniformément continue sur chaque intervalle borné. Donc, pour ε et k fixés il existe $\delta > 0$ tel que $|f(x) - f(y)| \leqslant \varepsilon$ si $|x - y| \leqslant \delta$ et $x \in [-k, k]$. Par suite

$$\begin{aligned}\{|f(X_n)-f(X)| > \varepsilon, |X| \leqslant k\} &\subset \{|X_n - X| > \delta, |X| \leqslant k\} \\ &\subset \{|X_n - X| > \delta\}.\end{aligned}$$

En combinant ceci avec (17.2) il vient

$$\{|f(X_n) - f(X)| > \varepsilon\} \subset \{|X_n - X| > \delta\} \cup \{|X| > k\}. \tag{17.3}$$

En utilisant la sous-additivité ($P(A \cup B) \leqslant P(A) + P(B)$) on déduit de (17.3) que

$$P\left(|f(X_n) - f(X)| > \varepsilon\right) \leqslant P(|X_n - X| > \delta) + P(|X| > k).$$

Mais $\{|X| > k\}$ tend vers l'ensemble vide lorsque $k \to \infty$, donc $\lim_{k\to\infty} P(|X| > k) = 0$. Pour $\gamma > 0$ arbitraire on peut alors choisir

k de sorte que $P(|X| > k) < \gamma$. Une fois k fixé on choisit δ pour avoir (17.3), et par suite

$$\limsup_{n\to\infty} P\left(|f(X_n) - f(X)| > \varepsilon\right) \leqslant \lim_{n\to\infty} P\left(|X_n - X| > \delta\right) + \gamma = \gamma.$$

Comme $\gamma > 0$ est arbitraire, on a le résultat. ∎

Remarque 17.2. Nous avons supposé les X_n et X réelles. Les mêmes définitions et les mêmes résultats sont valables si les X_n et X prennent leurs valeurs dans le même espace $\mathbb{R}^d$, à condition de comprendre la notation $|x|$ comme la norme euclidienne du vecteur $x \in \mathbb{R}^d$. De plus, dire que X_n converge vers X pour l'un des modes de convergence introduits ci-dessus revient exactement à dire que la suite des j-ièmes composantes des X_n converge vers la composante correspondante de X, pour tout $j = 1, \ldots, d$.

Exercices

1. Soit les $X_{n,j}$ comme dans l'exemple 2, et $Z_{n,j} = n^{\frac{1}{p}} X_{n,j}$. On note maintenant Y_n la suite obtenue en réordonnant les $Z_{n,j}$ (au lieu des $X_{n,j}$), comme dans l'exemple 2 également. Montrer que $Y_n \to 0$ en probabilité mais pas dans L^p, bien que chaque Y_n appartienne à L^p.
2. Montrer que le théorème 17.5(b) est faux en général, lorsqu'on ne suppose pas f continue. (*Indication :* Prendre $f(x) = 1_{\{0\}}(x)$ et les X_n tendant vers 0 en probabilité.)
3. Soit X_n des v.a. indépendantes avec $P(X_n = 1) = P(X_n = -1) = \frac{1}{2}$. Montrer que
$$\frac{1}{n}\sum_{j=1}^{n} X_j \xrightarrow{P} 0.$$
(*Indication :* Soit $S_n = \sum_{j=1}^{n} X_j$; utiliser l'inégalité de Bienaymé-Chebyshev pour $P(|S_n| > n\varepsilon)$.)
4. Soit X_n et S_n comme dans l'exercice 3. Montrer que $\frac{1}{n^2}S_{n^2} \to 0$ p.s. (*Indication :* Montrer $\sum_{n=1}^{\infty} P\left(\frac{1}{n^2}|S_{n^2}| > \varepsilon\right) < \infty$ et utiliser le lemme de Borel-Cantelli.)
5. Si on a $|X_n| \leqslant Y$ p.s. pour tout n, montrer que $\sup_n |X_n| \leqslant Y$ p.s. également.

6. Soit $X_n \xrightarrow{P} X$. Montrer que les fonctions caractéristiques φ_{X_n} convergent simplement vers φ_X.
7. Soit $X_1, \dots, X_n$ des v.a. indépendantes de loi de Cauchy de paramètres $\alpha = 0$ et $\beta = 1$ (donc de densité $f(x) = \frac{1}{\pi(1+x^2)}$). Montrer que $\frac{1}{n}\sum_{j=1}^n X_j$ suit aussi une loi de Cauchy. (*Indication :* Utiliser les fonctions caractéristiques ; voir l'exercice 1 du chapitre 13.)
8. Sous les mêmes hypothèses que dans l'exercice précédent, montrer qu'il n'existe aucune constante γ telle que $\frac{1}{n}\sum_{j=1}^n X_j \xrightarrow{P} \gamma$. (*Indication :* Utiliser l'exercice 7.) En déduire qu'il n'existe pas de constante γ telle que $\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^n X_j = \gamma$ p.s.
9. Si les v.a. $(X_n)_{n\geqslant 1}$ sont de moyennes nulles et de variances $\sigma^2_{X_n}$ tendant vers 0 quand $n \to \infty$, montrer que X_n converge vers 0 dans L^2 et en probabilité.
10. Soit X_j des v.a. indépendantes, de même loi, de moyenne nulle et de variance finie. Montrer que si $S_n = \sum_{j=1}^n X_j$, alors $\frac{1}{n}S_n$ tend vers 0 dans L^2 et en probabilité.
11.* Supposons que $X_n \to X$ p.s., les v.a. X_n et X étant réelles. Soit $Y = \sup_n |X_n|$. Montrer que $Y < \infty$ p.s.
12.* Supposons que $X_n \to X$ p.s., les v.a. X_n et X étant réelles, et soit $Y = \sup_n |X_n - X|$. Montrer que $Y < \infty$ p.s. (voir l'exercice 11), et après avoir défini une nouvelle mesure de probabilité Q par

$$Q(A) = \frac{1}{c}E\Big(1_A \frac{1}{1+Y}\Big), \qquad \text{où } c = E\Big(\frac{1}{1+Y}\Big),$$

montrer que X_n tend vers X dans L^1 pour la probabilité Q.
13. Soit A l'événement de l'exemple 1. Montrer que $P(A) = 0$. (*Indication :* Soit

$$A_n = \{ \text{ on a Face au } n\text{-ième tirage } \}.$$

Montrer que $\sum_{n=1}^\infty P(A_n) = \infty$ et utiliser le lemme de Borel-Cantelli.))
14. Soit X_n et X des v.a. de L^2, telles que X_n tende vers X dans L^2. Montrer que $E(X_n^2) \to E(X^2)$ (*Indication :* utiliser que $|x^2 - y^2| = (x-y)^2 + 2|y||x-y|$ et l'inégalité de Cauchy-Schwarz).
15.* (Un autre *théorème de convergence dominée.*) Soit $(X_n)_{n\geqslant 1}$ des v.a. satisfaisant $X_n \xrightarrow{P} X$. Supposons que $|X_n(\omega)| \leqslant C$ pour une certaine constante C et tout ω. Montrer que $\lim_{n\to\infty} E(|X_n - X|) = 0$. (*Indication :* Montrer d'abord que $P(|X| \leqslant C) = 1$.)

Chapitre 18

Convergence en loi

Dans le chapitre 17 nous avons défini trois types de convergence pour les variables aléatoires : presque sûre, dans L^p et en probabilité. Ces trois types de convergence, bien que différant de la convergence usuelle des fonctions en analyse, sont néanmoins de même nature et peuvent être conçues comme des variantes de la convergence simple habituelle. Il existe un autre mode de convergence, de nature bien différente, mais aussi très utile en probabilité : il s'agit de la convergence « en loi », appelée aussi « convergence étroite », et parfois « convergence faible ». Comme ce dernier nom l'indique, il s'agit d'un mode faible de convergence : plus la convergence est « faible », plus il est facile pour des v.a. de converger ! Ce qui est inhabituel pour ce type de convergence, c'est que des v.a. X_n peuvent converger vers X sans que les valeurs prises par X_n se rapprochent effectivement de celles prises par X en un même point ω, et même, ces variables peuvent être définies sur des espace de probabilité différents. En fait, il s'agit d'une convergence des *lois* des v.a. X_n vers la loi de X, et on a donc ainsi un mode de convergence *radicalement différent* de ce qui a été vu au chapitre 17.

Comme nous allons traiter de convergence de lois, nous commençons par examiner simplement les lois de probabilité sur $\mathbb{R}^d$ pour un d fixé, $d \geqslant 1$.

Définition 18.1. *Soit μ_n et μ des probabilités sur $\mathbb{R}^d$. On dit que la suite μ_n* converge étroitement *vers μ si $\int f(x)\mu_n(dx)$ converge vers $\int f(x)\mu(dx)$ pour toute fonction f réelle sur $\mathbb{R}^d$ qui est* continue et bornée.

À première vue il semble qu'il y ait une erreur typographique : on est habitué à considérer la propriété

$$\lim_{n\to\infty} \int f_n(x)\mu(dx) = \int f(x)\mu(dx);$$

mais ici la fonction f est fixée et ce sont les mesures qui varient. Noter aussi qu'on ne considère pas la convergence des intégrales pour toute fonction borélienne bornée, mais que nous imposons à f d'être en plus *continue*.

Définition 18.2. *Soit X_n et X des v.a. à valeurs dans $\mathbb{R}^d$. On dit que X_n* converge en loi vers X *si les lois P_{X_n} convergent étroitement vers P_X, et on écrit $X_n \xrightarrow{\mathcal{L}} X$.*

Une remarque très importante est à faire tout de suite : dans cette définition, les v.a. X_n et X peuvent être définies sur des espaces distincts, puisque seules leurs lois sont en cause. Il arrive même qu'une suite X_n converge vers une limite X qui ne peut pas exister sur les espaces sur lesquels sont définies les X_n, parce que ceux-ci sont « trop petits » : par exemple X_n est une variable binomiale de taille n, convenablement normalisée, et la limite X est normale ; l'espace naturel sur lequel X_n est définie contient $n+1$ points, et sur un tel espace toutes les v.a. sont discrètes. La convergence en loi permet donc une sorte de convergence pour des v.a. pour lesquelles toute autre forme de convergence serait impossible.

Théorème 18.1. *Soit X_n et X des v.a. à valeurs dans $\mathbb{R}^d$. On a $X_n \xrightarrow{\mathcal{L}} X$ si et seulement si*

$$\lim_{n\to\infty} E(f(X_n)) = E(f(X)),$$

pour toute fonction réelle continue et bornée sur $\mathbb{R}^d$.

Preuve. Il suffit de combiner les définitions 18.1 et 18.2 et d'observer que

$$E(f(X_n)) = \int f(x) P_{X_n}(dx), \qquad E(f(X)) = \int f(x) P_X(dx). \qquad \blacksquare$$

Pour avoir la convergence p.s. ou dans L^p ou en probabilité les v.a. X_n et X doivent être définies sur le même espace. Donc *a priori* la convergence en loi n'est pas comparable aux autres. On a cependant le résultat suivant :

Théorème 18.2. *Soit X_n et X des v.a., toutes définies sur le même espace de probabilité $(\Omega, \mathcal{A}, P)$. Si X_n converge vers X en probabilité, alors X_n converge aussi vers X en loi.*

Preuve. Soit f une fonction réelle bornée et continue sur $\mathbb{R}^d$. Le théorème 17.5 implique que $f(X_n)$ converge vers $f(X)$ en probabilité également. Comme f est bornée, cette convergence a aussi lieu dans L^1 (théorème 17.4). Donc $\lim_{n\to\infty} E(f(X_n)) = E(f(X))$, et le théorème 18.1 donne le résultat. $\blacksquare$

Ce théorème admet une réciproque *très partielle* :

Théorème 18.3. *Soit* X_n *et* X *des v.a. définies sur le même espace de probabilité* $(\Omega, \mathcal{A}, P)$. *Si* X_n *converge en loi vers* X *et si de plus* X *est p.s. égale à une constante, alors* X_n *converge vers* X *en probabilité.*

Preuve. Supposons que $X = a$ p.s. La fonction $f(x) = \frac{|x-a|}{1+|x-a|}$ étant bornée et continue, on a $\lim_{n\to\infty} E(\frac{|X_n-a|}{1+|X_n-a|}) = 0$, et donc $X_n \xrightarrow{P} X$ en vertu du théorème 17.1. ■

Il est tentant de penser que si $X_n \xrightarrow{\mathcal{L}} X$, alors $P(X_n \in A)$ converge vers $P(X \in A)$ pour tout ensemble borélien A. *Ce n'est pas vrai en général.* On a $P(X_n \in A) \to P(X \in A)$ pour certains ensembles A, mais ceux-ci sont tout à fait particuliers. Cela est en rapport avec la convergence des fonctions de répartition, dans le cas de v.a. réelles : supposons que $X_n \xrightarrow{\mathcal{L}} X$; dire que $F_n(x) = P(X_n \leqslant x)$ converge vers $F(x) = P(X \leqslant x)$ revient à dire que $P(X_n \in]-\infty, x]) \to P(X \in]-\infty, x])$, et cette propriété n'est pas toujours vraie !

Supposons donc les X_n et X à valeurs réelles, de fonctions de répartition F_n et F respectivement. Le théorème suivant est plutôt délicat et peut être omis : on remarquera que ce théorème est beaucoup plus simple dans le cas où la fonction de répartition F de la variable limite X est *continue*, ce qui suffit pour la plupart des applications. Nous donnons cependant la preuve dans le cas général. Pour comprendre l'énoncé, il convient de rappeler que la fonction F est croissante et continue à droite ; elle admet donc partout des limites à gauche, i.e. $F(x-) = \lim_{y\to x, y<x} F(y)$ existe pour tout x (voir l'exercice 4).

Théorème 18.4. *Soit* X_n *et* X *des v.a. réelles.*

(a) *Si* $X_n \xrightarrow{\mathcal{L}} X$, *alors* $\lim_{n\to\infty} F_n(x) = F(x)$ *pour tout* x *dans le sous-ensemble (partout dense) de* $\mathbb{R}$ *défini par* $D = \{x : F(x-) = F(x)\}$. D *est appelé l'ensemble des points de continuité de la fonction* F.

(b) *Supposons que* $\lim_{n\to\infty} F_n(x) = F(x)$ *pour tout* x *dans un sous-ensemble partout dense de* $\mathbb{R}$. *Alors* $X_n \xrightarrow{\mathcal{L}} X$.

Preuve de (a). Supposons que $X_n \xrightarrow{\mathcal{L}} X$, et soit $D = \{x : F(x-) = F(x)\}$. L'ensemble D est partout dense dans $\mathbb{R}$ puisque son complémentaire est au plus dénombrable (voir les exercices 4 et 5).

Fixons $x \in \mathbb{R}$. Pour chaque entier $p \geqslant 1$ on introduit les fonctions suivantes (dépendant de x), bornées et continues et vérifiant $g_p \leqslant 1_{]-\infty, x]} \leqslant f_p$:

$$f_p(y) = \begin{cases} 1 & \text{si } y \leqslant x \\ p(x-y)+1 & \text{si } x < y < x + \dfrac{1}{p} \\ 0 & \text{si } x + \dfrac{1}{p} \leqslant y \end{cases}$$

$$g_p(y) = \begin{cases} 1 & \text{si } y \leqslant x - \dfrac{1}{p} \\ p(x-y) & \text{si } x - \dfrac{1}{p} < y < x \\ 0 & \text{si } x \leqslant y. \end{cases}$$

On a alors pour chaque p :

$$\lim_{n\to\infty} \mathrm{E}(f_p(\mathrm{X}_n)) = \mathrm{E}(f_p(\mathrm{X})), \qquad \lim_{n\to\infty} \mathrm{E}(g_p(\mathrm{X}_n)) = \mathrm{E}(g_p(\mathrm{X})).$$

$$\mathrm{E}(g_p(\mathrm{X}_n)) \leqslant \mathrm{F}_n(x) \leqslant \mathrm{E}(f_p(\mathrm{X}_n)),$$

et donc

$$\mathrm{E}(g_p(\mathrm{X})) \leqslant \liminf_{n\to\infty} \mathrm{F}_n(x) \leqslant \limsup_{n\to\infty} \mathrm{F}_n(x) \leqslant \mathrm{E}(f_p(\mathrm{X})) \tag{18.1}$$

pour chaque $p \geqslant 1$. Mais $\lim_{p\to\infty} f_p(y) = 1_{]-\infty,x]}(y)$ et $\lim_{p\to\infty} g_p(y) = 1_{]-\infty,x[}(y)$, donc le théorème de convergence dominée de Lebesgue entraîne que

$$\begin{aligned} \lim_{p\to\infty} \mathrm{E}(f_p(\mathrm{X})) &= \mathrm{E}\left(1_{]-\infty,x]}(\mathrm{X})\right) \\ &= \mathrm{P}(\mathrm{X} \leqslant x) = \mathrm{F}(x), \end{aligned}$$

et de même $\lim_{p\to\infty} \mathrm{E}(g_p(\mathrm{X})) = \mathrm{F}(x-)$ (la limite à gauche au point x). En combinant ces résultats avec (18.1) on obtient

$$\mathrm{F}(x-) \leqslant \liminf_n \mathrm{F}_n(x) \leqslant \limsup_n \mathrm{F}_n(x) \leqslant \mathrm{F}(x). \tag{18.2}$$

Si alors $x \in \mathrm{D}$ on a $\mathrm{F}(x-) = \mathrm{F}(x)$, donc $\mathrm{F}_n(x) \to \mathrm{F}(x)$.

Preuve de (b). Maintenant nous supposons que $\lim_{n\to\infty} \mathrm{F}_n(x) = \mathrm{F}(x)$ pour tout $x \in \Delta$, où Δ est un sous-ensemble partout dense de $\mathbb{R}$. Soit f une fonction bornée et continue sur $\mathbb{R}$, et soit $\varepsilon > 0$. Soit aussi $r, s \in \Delta$ tels que

$$\mathrm{P}(\mathrm{X} \notin]r,s]) = 1 - \mathrm{F}(s) + \mathrm{F}(r) \leqslant \varepsilon.$$

(On peut trouver de tels r et s car $\lim_{x\to-\infty} \mathrm{F}(x) = 0$ et $\lim_{x\to+\infty} \mathrm{F}(x) = 1$, et puisque Δ est partout dense.) Comme F_n converge vers F sur Δ par hypothèse, il existe N_1 tel que pour $n \geqslant \mathrm{N}_1$,

$$\mathrm{P}(\mathrm{X}_n \notin \,]r,s]) = 1 - \mathrm{F}_n(s) + \mathrm{F}_n(r) \leqslant 2\varepsilon. \tag{18.3}$$

Comme $[r,s]$ est un intervalle fermé (compact) et f est continue, elle est uniformément continue sur $[r,s]$: il existe donc un nombre *fini* de points $r = r_0 < r_1 < \cdots < r_k = s$ appartenant à Δ (utiliser encore le fait que Δ est partout dense) et tels que

$$|f(x) - f(r_j)| \leqslant \varepsilon \quad \text{si} \quad r_{j-1} \leqslant x \leqslant r_j.$$

Il s'ensuit que la fonction

$$g(x) = \sum_{j=1}^{k} f(r_j) 1_{]r_{j-1},r_j]}(x) \tag{18.4}$$

vérifie $|f(x) - g(x)| \leqslant \varepsilon$ si $x \in \,]r,s]$. Donc si $\alpha = \sup_x |f(x)|$, nous obtenons

$$|\mathrm{E}(f(\mathrm{X}_n)) - \mathrm{E}(g(\mathrm{X}_n))| \leqslant \alpha \mathrm{P}(\mathrm{X}_n \notin \,]r,s]) + \varepsilon, \tag{18.5}$$

et la même équation avec X au lieu de X_n.

D'après (18.4) on a

$$\mathrm{E}(g(\mathrm{X}_n)) = \sum_{j=1}^{k} f(r_j)\{\mathrm{F}_n(r_j) - \mathrm{F}_n(r_{j-1})\},$$

et

$$\mathrm{E}(g(\mathrm{X})) = \sum_{j=1}^{k} f(r_j)\{\mathrm{F}(r_j) - \mathrm{F}(r_{j-1})\}.$$

Comme les r_j sont tous dans Δ on a $\lim_{n\to\infty} \mathrm{F}_n(r_j) = \mathrm{F}(r_j)$ pour chaque j. Les r_j étant en nombre fini, on peut trouver N_2 tel que

$$n \geqslant \mathrm{N}_2 \qquad \Rightarrow \qquad |\mathrm{E}(g(\mathrm{X}_n)) - \mathrm{E}(g(\mathrm{X}))| \leqslant \varepsilon. \tag{18.6}$$

Combinons alors (18.5) pour X_n et X avec (18.6) : si $n \geqslant \max(\mathrm{N}_1, \mathrm{N}_2)$, il vient

$$\begin{aligned}
&|\mathrm{E}(f(\mathrm{X}_n)) - \mathrm{E}(f(\mathrm{X}))| \\
&\quad \leqslant |\mathrm{E}(f(\mathrm{X}_n)) - \mathrm{E}(g(\mathrm{X}_n))| + |\mathrm{E}(g(\mathrm{X}_n)) - \mathrm{E}(g(\mathrm{X}))| \\
&\qquad\qquad\qquad\qquad + |\mathrm{E}(g(\mathrm{X})) - \mathrm{E}(f(\mathrm{X}))| \\
&\quad \leqslant (\alpha \mathrm{P}(\mathrm{X}_n \notin \,]r,s]) + \varepsilon) + \varepsilon + (\alpha \mathrm{P}(\mathrm{X} \notin \,]r,s]) + \varepsilon) \\
&\quad \leqslant (2\alpha\varepsilon + \varepsilon) + \varepsilon + (\alpha\varepsilon + \varepsilon) \\
&\quad \leqslant 3\alpha\varepsilon + 3\varepsilon.
\end{aligned}$$

Comme $\varepsilon > 0$ est arbitrairement petit on a $\lim_{n\to\infty} \mathrm{E}(f(\mathrm{X}_n)) = \mathrm{E}(f(\mathrm{X}))$, d'où le résultat (voir le théorème 18.1). ■

Exemples.

1. Soit $(\mu_n)_{n\geqslant 1}$ une suite de probabilités sur $\mathbb{R}$ qui sont toutes des masses de Dirac : i.e., pour chaque n il existe un point α_n tel que $\mu_n(\{\alpha_n\}) = 1$ et $\mu_n(\{\alpha_n\}^c) = \mu_n(\mathbb{R} \setminus \{\alpha_n\}) = 0$. La suite μ_n converge étroitement vers une limite μ si et seulement si α_n converge vers un réel α, et dans ce cas μ est la masse de Dirac au point α. Pour le voir, on peut bien sûr utiliser la définition 18.1 directement ; il est également instructif d'appliquer le théorème 18.4 aux fonctions de répartition $\mathrm{F}_n(x) = 1_{[\alpha_n,\infty[}(x)$ des μ_n : la suite $\mathrm{F}_n(x)$ converge vers une limite $\mathrm{F}(x)$ pour tout x dans un ensemble partout dense si et seulement si α_n converge vers une limite α, et dans ce cas $\mathrm{F}(x) = 1_{[\alpha,\infty[}(x)$. Noter que dans ce cas, l'ensemble des x pour lesquels $\mathrm{F}_n(x) \to \mathrm{F}(x)$ est $\mathbb{R}$ tout entier si on a $\alpha_n \leqslant \alpha$ pour tout n assez grand, et est $\mathbb{R}\setminus\{\alpha\}$ sinon.
2. Soit
$$\mathrm{F}_n(x) = \begin{cases} 0 & \text{si } x \leqslant -\dfrac{1}{n} \\ \dfrac{1}{2} + \dfrac{n}{2}x & \text{si } -\dfrac{1}{n} < x < \dfrac{1}{n} \\ 1 & \text{si } x \geqslant -\dfrac{1}{n}. \end{cases}$$
Alors
$$\lim_{n\to\infty} \mathrm{F}_n(x) = \mathrm{F}(x) = 1_{[0,\infty[}(x)$$
pour tout $x \neq 0$; donc l'ensemble D du théorème 18.4 est $\mathrm{D} = \mathbb{R} \setminus \{0\}$. Ainsi des v.a. X_n de fonctions de répartition F_n vérifient $\mathrm{X}_n \xrightarrow{\mathcal{L}} \mathrm{X}$, où X est la v.a. constante égale à 0. On a donc montré qu'une suite de v.a. uniformément distribuées sur les intervalles $[-\frac{1}{n}, \frac{1}{n}]$ converge en loi vers 0.
3. Soit $(\mathrm{X}_n)_{n\geqslant 1}$ et X des v.a. réelles admettant les densités f_n et f. La fonction de répartition
$$\mathrm{F}(x) = \int_{-\infty}^{x} f(u)du$$
est continue et $\mathrm{F}(x-) = \mathrm{F}(x)$ pour tout x. Supposons que $f_n(x) \leqslant g(x)$ pour tous n et x, que $\int_{-\infty}^{\infty} g(x)dx < \infty$, et que $\lim_{n\to\infty} f_n = f$

p.p. Alors $F_n(x)$ converge vers $F(x)$ pour tout x par le théorème de convergence dominée, et $X_n \xrightarrow{\mathcal{L}} X$.

Pour traiter cet exemple on peut aussi écrire

$$\begin{aligned}\lim_{n\to\infty} \int h(x) P_{X_n}(dx) &= \lim_{n\to\infty} \int h(x) f_n(x) dx \\ &= \int h(x) \lim_n f(x) dx \\ &= \int h(x) f(x) dx = \int h(x) P_X(dx)\end{aligned}$$

pour toute fonction bornée continue h (encore grâce au théorème de convergence dominée). Cette démonstration marche aussi lorsque les v.a. X_n et X sont à valeurs dans $\mathbb{R}^d$; de plus on obtient une forme de convergence un peu plus forte que la convergence en loi, puisque la propriété $\int h\, dP_{X_n} \to \int h\, dP_X$ est vraie non seulement pour les fonctions h continues bornées, mais pour toutes les fonctions boréliennes bornées.

L'exemple précédent admet l'extension suivante, qui peut sembler un peu surprenante *a priori* : on peut échanger limite et intégrale dans un cas où les fonctions ne sont pas « dominées » par une fonction intégrable : cela est dû au fait que les densités f et f_n sont positives et d'intégrale 1.

Théorème 18.5. *Soit $(X_n)_{n\geqslant 1}$ et X des v.a. à valeurs dans $\mathbb{R}^d$ et admettant les densités f_n et f. Si la suite f_n converge simplement, ou même presque partout (relativement à la mesure de Lebesgue sur $\mathbb{R}^d$), alors $X_n \xrightarrow{\mathcal{L}} X$.*

Preuve. Soit h une fonction borélienne bornée sur $\mathbb{R}^d$, et soit $\alpha = \sup_x |h(x)|$. Posons $h_1(x) = h(x) + \alpha$ et $h_2(x) = \alpha - h(x)$. Ces deux fonctions sont positives, donc les fonctions $h_1 f_n$ et $h_2 f_n$ également ; de plus les suites $h_i f_n$ pour $i = 1, 2$ convergent p.p. vers $h_i f$. Le lemme de Fatou (théorème 9.1) donne alors

$$\begin{aligned}E(h_i(X)) = \int f(x) h_i(x)\, dx &\leqslant \liminf_{n\to\infty} \int f_n(x) h_i(x)\, dx \\ &= \liminf_{n\to\infty} E(h_i(X_n)). \quad (18.7)\end{aligned}$$

Noter que $E(h(X_n)) = E(h_1(X_n)) - \alpha$ et $E(h(X_n)) = \alpha - E(h_2(X_n))$, et les mêmes égalités sont vraies avec X au lieu de X_n. Comme $\liminf(x_n) = -\limsup(-x_n)$, en appliquant (18.7) pour $i = 1$ et $i = 2$ on obtient

$$E(h(X)) \leqslant \liminf_{n\to\infty} E(h(X_n)),$$
$$E(h(X)) \geqslant \limsup_{n\to\infty} E(h(X_n)).$$

Donc $E(h(X_n)) \to E(h(X))$, et le résultat est démontré. ■

Le théorème suivant est connu sous le nom de « principe de Helly ». C'est un résultat difficile, mais nous en aurons besoin pour étudier les rapports entre convergence en loi et convergence des fonctions caractéristiques. La propriété (18.8) est souvent appelée *tension de la suite de probabilités* μ_n.

Théorème 18.6. *Soit* $(\mu_n)_{n\geqslant 1}$ *une suite de probabilités sur* $\mathbb{R}$. *Supposons que*

$$\lim_{m\to\infty} \sup_n \mu_n([-m,m]^c) = 0. \tag{18.8}$$

Il existe alors une suite (n_k) *d'entiers croissant vers l'infini telle que la sous-suite de probabilités* $(\mu_{n_k})_{k\geqslant 1}$ *converge étroitement.*

Preuve. Soit $F_n(x) = \mu_n(]-\infty,x])$. Pour chaque x on a $0 \leqslant F_n(x) \leqslant 1$, donc la suite de réels $(F_n(x))_{n\geqslant 1}$ est bornée. D'après le théorème de Bolzano-Weierstrass il existe alors une suite d'entiers (n_k) (dépendant de x, bien sûr) tendant vers l'infini, telle que la suite $(F_{n_k}(x))_{k\geqslant 1}$ converge.

Pour montrer la convergence étroite nous devons montrer la convergence des fonctions de répartition sur un ensemble partout dense, mais qui peut être dénombrable. Nous allons donc nous restreindre à l'ensemble $\mathbb{Q}$ des rationnels, qui est dénombrable et partout dense dans $\mathbb{R}$. Soit $r_1, r_2, \ldots, r_j, \ldots$ une énumération de tous les rationnels. Pour r_1, il existe une sous-suite $(n_{1,k})$ de la suite de (n) de tous les entiers, telle que la limite suivante existe :

$$G(r_1) = \lim_{k\to\infty} F_{n_{1,k}}(r_1).$$

Pour r_2, il existe une sous-suite $(n_{2,k})$ de la suite $(n_{1,k})$ telle que la limite suivante existe :

$$G(r_2) = \lim_{k\to\infty} F_{n_{2,k}}(r_2).$$

On continue de la même manière, par récurrence sur j : pour r_j on peut trouver une sous-suite $(n_{j,k})$ de la suite $(n_{j-1,k})$ telle que la limite suivante existe :

$$G(r_j) = \lim_{k\to\infty} F_{n_{j,k}}(r_j).$$

On construit alors une sous-suite unique par le « procédé diagonal », en posant $n_k = n_{k,k}$. Pour chaque j on a

$$G(r_j) = \lim_{k\to\infty} F_{n_k}(r_j),$$

puisque (n_k) est une sous-suite de la suite $(n_{j,k})$ dès que $k \geqslant j$.

Ensuite, on pose

$$F(x) = \inf_{\substack{y \in \mathbb{Q} \\ y > x}} G(y). \tag{18.9}$$

Comme la fonction G définie sur $\mathbb{Q}$ est croissante, la fonction F (définie, quant à elle, sur $\mathbb{R}$) est aussi croissante, et elle est continue à droite par construction.

Soit $\varepsilon > 0$. Par hypothèse il existe m tel que

$$F_n([-m, m]^c) \leqslant \varepsilon$$

pour tous les n. Donc

$$F_n(x) \leqslant \varepsilon \text{ si } x < -m, \qquad F_n(x) \geqslant 1 - \varepsilon \text{ si } x > m;$$

On a évidemment les mêmes inégalités pour $G(x)$ (pour x rationnel), donc aussi

$$\left.\begin{array}{ll} F(x) \leqslant \varepsilon & \text{si } x < -m \\ F(x) \geqslant 1 - \varepsilon & \text{si } x \geqslant m. \end{array}\right\} \tag{18.10}$$

Comme $0 \leqslant F \leqslant 1$ et comme F est continue à droite, la propriété (18.10) suffit à entraîner que F est la fonction de répartition d'une probabilité μ sur $\mathbb{R}$.

Finalement, supposons que x vérifie $F(x-) = F(x)$. Pour $\varepsilon > 0$ il existe $y, z \in \mathbb{Q}$ avec $y < x < z$

$$F(x) - \varepsilon \leqslant G(y) \leqslant F(x) \leqslant G(z) \leqslant F(x) + \varepsilon.$$

Donc pour k suffisamment grand,

$$F(x) - 2\varepsilon \leqslant F_{n_k}(y) \leqslant F_{n_k}(x) \leqslant F_{n_k}(z) \leqslant F(x) + 2\varepsilon. \tag{18.11}$$

Les inégalités (18.11) donnent

$$\begin{aligned} F(x) - 2\varepsilon \leqslant F(y) &\leqslant \liminf_{k\to\infty} F_{n_k}(x) \\ &\leqslant \limsup_{k\to\infty} F_{n_k}(x) \leqslant F(z) \leqslant F(x) + 2\varepsilon \end{aligned}$$

et comme $\varepsilon > 0$ est arbitrairement petit on en déduit qu'en fait $F_{n_k}(x) \to F(x)$: par suite μ_{n_k} converge étroitement vers μ par le théorème 18.4. ■

Remarque 18.1. Ce théorème admet une version multi-dimensionnelle (la preuve est analogue, en un peu plus compliqué) : si les μ_n sont des probabilités sur $\mathbb{R}^d$, il suffit de remplacer la condition (18.8) par

$$\lim_{m\to\infty} \sup_n \mu_n(\{x \in \mathbb{R}^d : |x| > m\}) = 0. \tag{18.12}$$

Il est parfois utile de savoir que pour vérifier que μ_n converge étroitement vers μ, il n'est pas nécessaire de montrer que $\int f \, d\mu_n \to \int f \, d\mu$ pour *toutes* les fonctions bornées continues f. Nous exprimons ce résultat en termes de convergence en loi de v.a. :

Théorème 18.7. *Soit* X_n *et* X *des v.a. à valeurs dans* $\mathbb{R}^d$*. Pour que* $X_n \xrightarrow{\mathcal{L}} X$ *il faut et il suffit que* $E(g(X_n)) \to E(g(X))$ *pour toute fonction* f *bornée et lipschitzienne.*

Preuve. Une fonction g est lipschitzienne s'il existe une constante k telle que $|g(x) - g(y)| \leqslant k\|x - y\|$ pour tous x, y. La condition nécessaire étant triviale, nous ne prouvons que la condition suffisante.

Soit f une fonction continue bornée, et $\alpha = \sup_x |f(x)|$. Supposons qu'il existe des fonctions lipschitziennes bornées g_i vérifiant $-\alpha \leqslant g_i \leqslant g_{i+1} \leqslant f$ et $\lim_{i\to\infty} g_i(x) = f(x)$. On a alors

$$\liminf_{n\to\infty} E(f(X_n)) \geqslant \lim_{n\to\infty} E(g_i(X_n)) = E(g_i(X)),$$

pour chaque i fixé. Le théorème de convergence monotone appliqué à la suite $g_i(X) + \alpha$ et à sa limite $f(X) + \alpha$ entraîne que

$$\lim_{i\to\infty} E(g_i(X)) = E(f(X)).$$

Donc

$$\liminf_{n\to\infty} E(f(X_n)) \geqslant E(f(X)). \tag{18.13}$$

Le même argument appliqué à la fonction $-f$ donne

$$\limsup_{n\to\infty} E(f(X_n)) \leqslant E(f(X)), \tag{18.14}$$

et en combinant (18.13) et (18.14) on obtient

$$\lim_{n\to\infty} E(f(X_n)) = E(f(X)),$$

ce qui montre la convergence en loi cherchée.

Il nous reste donc à construire les fonctions g_i. Pour cela, on va construire des fonctions lipschitziennes h_j telles que $\sup_k h_k(x) = f(x)$ et $h_k(x) \geqslant -\alpha$, et il nous suffira de poser $g_i(x) = \max\{h_1(x), \dots, h_i(x)\}$.

En remplaçant f par $\tilde{f} = f + \alpha$ si nécessaire, on peut supposer sans perte de généralité que la fonction continue et bornée f est aussi positive. Pour tout borélien A de $\mathbb{R}^d$ définissons la fonction « distance à A » par

$$d_A(x) = \inf\{\|x - y\| ; y \in A\}.$$

Pour tout rationnel $r \geqslant 0$ et tout entier m, posons $A_r = \{y : f(y) \leqslant r\}$ et

$$h_{m,r}(x) = r \wedge (m\, d_{A_r}(x)).$$

On a $|d_A(x) - d_A(y)| \leqslant \|x - y\|$, donc $|h_{m,r}(x) - h_{m,r}(y)| \leqslant m\|x - y\|$, et $h_{m,r}$ est lipschitzienne. Comme $0 \leqslant h_{m,r}(x) \leqslant r$ et $h_{m,r}(x) = 0$ si $f(x) \leqslant r$, on a $0 \leqslant h_{m,r}(x) \leqslant f(x)$.

Choisissons un point $x \in \mathbb{R}^d$ et un $\varepsilon > 0$, puis un rationnel $r > 0$ tel que $f(x) - \varepsilon < r < f(x)$. Comme f est continue, on a $f(y) > r$ pour tout y dans un voisinage de x. Donc $d_{A_r}(x) > 0$, et $h_{m,r}(x) = r > f(x) - \varepsilon$ pour m assez grand. Comme l'ensemble des rationnels est dénombrable, la famille $\{h_{m,r} ; m \in \mathbb{N}, r \in \mathbb{Q}_+\}$ est dénombrable. Si $\{h_i\}_{i \geqslant 1}$ en constitue une énumération, on a que $\sup_i h_i(x) \geqslant f(x)$. Mais $h_i \leqslant f$ pour tout i, donc en fait $\sup_i h_i(x) = f(x)$, et la preuve est terminée. ■

Corollaire 18.1. *Soit* X_n *et* X *des v.a. à valeurs dans* $\mathbb{R}^d$. *Pour que* $X_n \xrightarrow{\mathcal{L}} X$ *il faut et il suffit que* $E(g(X_n)) \to E(g(X))$ *pour toute fonction* g *bornée et uniformément continue.*

Preuve. Comme toute fonction lipschitzienne est uniformément continue, il suffit d'appliquer le théorème 18.7. ■

Remarque 18.2. Dans le théorème 18.7 nous avons réduit la classe des « fonctions test » pour que des v.a. à valeurs dans $\mathbb{R}^d$ convergent en loi. On peut se demander s'il est possible de la réduire encore. La réponse est *oui* : elle peut être réduite, par exemple, à l'ensemble des fonctions C^∞ à support compact (i.e. les fonctions indéfiniment différentiables et nulles en dehors d'un ensemble borné, ou compact) ; voir les exercices 19 à 22.

Une conséquence du théorème 18.7 est :

Théorème 18.8. (Théorème de Slutsky.) *Soit* $(X_n)_{n \geqslant 1}$ *et* $(Y_n)_{n \geqslant 1}$ *deux suites de v.a. à valeurs dans* $\mathbb{R}^d$. *Supposons que* $X_n \xrightarrow{\mathcal{L}} X$ *et que* $\|X_n - Y_n\| \to 0$ *en probabilité. Alors* $Y_n \xrightarrow{\mathcal{L}} X$.

Preuve. Grâce au théorème 18.7 il suffit de montrer que $\lim_n \mathrm{E}(f(\mathrm{Y}_n)) = \mathrm{E}(f(\mathrm{X}))$ pour toute fonction bornée lipschitzienne f. On a alors $|f(x) - f(y)| \leqslant k\|x-y\|$ et $|f(x)| \leqslant k$ pour une certaine constante k. Par suite si $\varepsilon > 0$ on a

$$\begin{aligned} |\mathrm{E}(f(\mathrm{X}_n) - f(\mathrm{Y}_n))| &\leqslant \mathrm{E}(|f(\mathrm{X}_n) - f(\mathrm{Y}_n)|) \\ &\leqslant k\varepsilon + \mathrm{E}(|f(\mathrm{X}_n) - f(\mathrm{Y}_n)| 1_{\{\|\mathrm{X}_n - \mathrm{Y}_n\| > \varepsilon\}}) \\ &\leqslant k\varepsilon + 2k\mathrm{P}(\|\mathrm{X}_n - \mathrm{Y}_n\| > \varepsilon). \end{aligned}$$

Mais $\lim_{n\to\infty} \mathrm{P}(\|\mathrm{X}_n - \mathrm{Y}_n\| > \varepsilon) = 0$, et comme $\varepsilon > 0$ est arbitrairement petit on en déduit que $\lim_{n\to\infty} |\mathrm{E}(f(\mathrm{X}_n) - f(\mathrm{Y}_n))| = 0$. Par suite

$$\lim_{n\to\infty} \mathrm{E}(f(\mathrm{Y}_n)) = \lim_{n\to\infty} \mathrm{E}(f(\mathrm{X}_n)) = \mathrm{E}(f(\mathrm{X})),$$

et le théorème est prouvé. ■

Terminons ce chapitre par l'étude de la convergence en loi des v.a. discrètes (ne prenant qu'un nombre fini ou dénombrable de valeurs), comme les v.a. binomiales, de Poisson, hypergéométriques, etc. Comme l'espace d'état, disons E, est au plus dénombrable, on peut considérer que *toute* fonction sur E est continue : cela revient à munir E de la topologie discrète. (*Attention* : si l'espace E est contenu de manière naturelle dans $\mathbb{R}$, la topologie usuelle de $\mathbb{R}$ induit sur E la topologie discrète si et seulement si le minimum de $|x-y|$ pour $x, y \in \mathrm{E} \cap [-m, m]$ est strictement positif pour tout $m > 0$; c'est le cas par exemple si $\mathrm{E} = \mathbb{N}$ ou $\mathrm{E} = \mathbb{Z}$, ou si E est une suite de réels tendant vers l'infini ; ce n'est pas vrai si E est une suite de réels tendant vers une limite finie.) Le théorème suivant donne une caractérisation simple de la convergence en loi dans le cas discret, et peut se comparer au théorème 18.5.

Théorème 18.9. *Soit* X_n *et* X *des v.a. à valeurs dans un espace* E *fini ou dénombrable. Alors* $\mathrm{X}_n \xrightarrow{\mathcal{L}} \mathrm{X}$ *si et seulement si on a*

$$\lim_{n\to\infty} \mathrm{P}(\mathrm{X}_n = j) = \mathrm{P}(\mathrm{X} = j) \tag{18.15}$$

pour tout $j \in \mathrm{E}$.

Preuve. Supposons d'abord que $\mathrm{X}_n \xrightarrow{\mathcal{L}} \mathrm{X}$. On a

$$\lim_{n\to\infty} \mathrm{E}(f(\mathrm{X}_n)) = \mathrm{E}(f(\mathrm{X}))$$

pour toute fonction continue bornée f sur E. Chaque fonction indicatrice $1_{\{j\}}$ étant continue et bornée, on a (18.15).

Supposons inversement (18.15). Soit f une fonction bornée et $\alpha = \sup_{j\in \mathrm{E}} |f(j)|$. Soit $\varepsilon > 0$. Comme

$$\sum_{j\in \mathrm{E}} \mathrm{P}(\mathrm{X} = j) = 1$$

est une série convergente, il existe un sous-ensemble *fini* Λ de E tel que

$$\sum_{j\in \Lambda} \mathrm{P}(\mathrm{X} = j) \geqslant 1 - \varepsilon.$$

(18.15) entraîne alors que pour n assez grand,

$$\sum_{j\in \Lambda} \mathrm{P}(\mathrm{X}_n = j) \geqslant 1 - 2\varepsilon.$$

On a

$$\mathrm{E}(f(\mathrm{X})) = \sum_{j\in \mathrm{E}} f(j)\mathrm{P}(\mathrm{X} = j),$$

donc pour n assez grand il vient

$$\left.\begin{aligned} &\left|\mathrm{E}(f(\mathrm{X})) - \textstyle\sum_{j\in\Lambda} f(j)\mathrm{P}(\mathrm{X} = j)\right| \leqslant \alpha\varepsilon \\ &\left|\mathrm{E}(f(\mathrm{X}_n)) - \textstyle\sum_{j\in\Lambda} f(j)\mathrm{P}(\mathrm{X}_n = j)\right| \leqslant 2\alpha\varepsilon. \end{aligned}\right\} \tag{18.16}$$

Finalement on observe que, comme Λ est fini,

$$\lim_{n\to\infty} \sum_{j\in\Lambda} f(j)\mathrm{P}(\mathrm{X}_n = j) = \sum_{j\in\Lambda} f(j)\mathrm{P}(\mathrm{X} = j). \tag{18.17}$$

On déduit alors de (18.16) et de (18.17) que

$$\limsup_{n\to\infty} |\mathrm{E}(f(\mathrm{X}_n)) - \mathrm{E}(f(\mathrm{X}))| \leqslant 3\alpha\varepsilon.$$

Comme $\varepsilon > 0$ est arbitrairement petit, il vient

$$\lim_{n\to\infty} \mathrm{E}(f(\mathrm{X}_n)) = \mathrm{E}(f(\mathrm{X}))$$

et on en déduit que $\mathrm{X}_n \xrightarrow{\mathcal{L}} \mathrm{X}$ ■

Exemples (suite).

4. Si μ_λ désigne la loi de Poisson de paramètre λ, on a
$$\mu_\lambda(j) = e^{-\lambda}\frac{\lambda^j}{j!},$$
et donc si $\lambda_n \to \lambda$, on a $\mu_{\lambda_n}(j) \to \mu_\lambda(j)$ pour tout $j \in \mathbb{N}$. Par suite μ_{λ_n} converge étroitement vers μ_λ.
5. Si μ_p désigne la loi binomiale de paramètres (n, p) et si $p_k \to p$, exactement comme dans l'exemple 4 on vérifie que μ_{p_k} converge étroitement vers μ_p.
6. Soit $\mu_{n,p}$ la loi binomiale (n, p). Considérons la suite μ_{n,p_n} où $\lim_{n\to\infty} np_n = \lambda > 0$. En reprenant l'exercice 1 du chapitre 4, on vérifie que la suite μ_{n,p_n} converge étroitement vers la loi de Poisson de paramètre λ.

Exercices

1. Montrer que si $X_n \xrightarrow{L^p} X$ ($p \geqslant 1$), alors $X_n \xrightarrow{\mathcal{L}} X$.
2. Soit $\alpha \in \mathbb{R}^d$ et $\varepsilon > 0$. Construire une fonction continue $f : \mathbb{R}^d \to \mathbb{R}$ telle que $0 \leqslant f \leqslant 1$, que $f(\alpha) = 0$, et que $f(x) = 1$ si $|x - \alpha| \geqslant \varepsilon$. (*Indication :* Résoudre d'abord l'exercice quand $d = 1$, puis généraliser au cas $d \geqslant 2$.)
3. Soit X une v.a. réelle de fonction de répartition F. Montrer que $F(x-) = F(x)$ si et seulement si $P(X = x) = 0$.
4.* Soit $g : \mathbb{R} \to [0, 1]$ croissante et continue à droite. Montrer que g a des limites à gauche en chaque point, et que l'ensemble $\Lambda = \{x : g(x-) \neq g(x)\}$ est au plus dénombrable (*Indication :* Montrer d'abord qu'il n'y a qu'un nombre fini (éventuellement nul) de points tels que $g(x) - g(x-) > \frac{1}{k}$).
5.* Soit F une fonction de répartition, et $D = \{x : F(x-) = F(x)\}$. Montrer que D est partout dense dans $\mathbb{R}$. (*Indication :* Utiliser l'exercice 4.)
6. Soit X_n une v.a. uniformément distribuée sur $[-n, n]$. Dans quel(s) sens la suite X_n converge-t-elle ? [*Rép.* : Aucun.]
7. Soit f_n des densités de probabilité sur $\mathbb{R}$, et supposons que $\lim_{n\to\infty} f_n(x) = e^{-x}1_{(x>0)}$. Si les X_n sont des v.a. de densités f_n, que peut-on dire de la convergence des X_n ? [*Rép.* : $X_n \xrightarrow{\mathcal{L}} X$, où X est exponentielle de paramètre 1.]

8. Soit $(X_n)_{n\geqslant 1}$ des v.a. indépendantes de loi de Cauchy de paramètres $\alpha = 0$ et $\beta = 1$. Soit $Y_n = \frac{X_1+\cdots+X_n}{n}$. Montrer que Y_n converge en loi et trouver la limite. Est-ce que les Y_n convergent en probabilité ?
9. Soit $(X_n)_{n\geqslant 1}$ une suite de v.a. réelles, telle que $\sup_n E((X^n)^2) < \infty$, et μ_n la loi de X_n. Montrer que la suite (μ_n) est tendue (i.e. vérifie (18.8)). (*Indication :* utiliser l'inégalité de Bienaymé-Chebyshev.)
10.* Soit X_n, X et Y des v.a. réelles, toutes définies sur le même espace de probabilité $(\Omega, \mathcal{A}, P)$. Supposons que

$$\lim_{n\to\infty} E(f(X_n)g(Y)) = E(f(X)g(Y))$$

pour toutes fonctions bornées f et g, avec f continue et g borélienne. Montrer que la suite bidimensionnelle (X_n, Y) converge en loi vers (X, Y). Si de plus $X = h(Y)$ pour une fonction borélienne h, montrer que $X_n \xrightarrow{P} X$.
11. Soit μ_α la loi de Pareto de paramètre α. Si $\alpha_n \to \alpha > 0$, montrer que μ_{α_n} converge étroitement vers μ_α.
12. Soit μ_α la loi géométrique de paramètre α. Si $\alpha_n \to \alpha > 0$, montrer que μ_{α_n} converge étroitement vers μ_α.
13. Soit $\mu_{(N,b,n)}$ une loi hypergéométrique. On fixe n et on fait tendre N et b vers l'infini, de sorte que $p = \frac{b}{N}$ reste contant. Montrer que $\mu_{(N,b,n)}$ converge étroitement vers la loi binomiale de paramètres (n, p). [Comparer avec l'exercice 8 du chapitre 5.]
14. (Théorème de Slutsky.) Soit X_n des v.a. convergeant en loi vers X, et Y_n des v.a. convergeant en probabilité vers une constante c, toutes ces v.a. étant définies sur le même espace. Montrer (a) que $X_nY_n \xrightarrow{\mathcal{L}} cX$, et (b) que $\frac{X_n}{Y_n} \xrightarrow{\mathcal{L}} \frac{X}{c}$ si $c \neq 0$.
15. Soit $(X_n)_{n\geqslant 1}$ et $(Y_n)_{n\geqslant 1}$ des v.a. réelles définies sur le même espace. Si $X_n \xrightarrow{\mathcal{L}} X$ et si $Y_n \xrightarrow{P} 0$, montrer que $X_n + Y_n \xrightarrow{\mathcal{L}} X$.
16. Soit $(X_n)_{n\geqslant 1}$ des v.a. réelles de fonctions de répartition F_n, avec $X_n \xrightarrow{\mathcal{L}} X$. Soit $p > 0$. Montrer que pour tout $N > 0$ on a

$$\int_{-N}^{N} |x|^p F(dx) \leqslant \limsup_{n\to\infty} \int_{-N}^{N} |x|^p F_n(dx) < \infty.$$

17.* Soit $(X_n)_{n\geqslant 1}$ et X des v.a. réelles de fonctions de répartition F_n et F vérifiant pour un $r > 0$:

$$\lim_{n\to\infty}\int_{-\infty}^{\infty}|F_n(x)-F(x)|^r dx = 0.$$

Montrer que $X_n \xrightarrow{\mathcal{L}} X$. (*Indication :* Supposer qu'il existe un point de continuité y de F tel que $\lim_{n\to\infty} F_n(y) \neq F(y)$. Il existe alors $\varepsilon > 0$ et une sous-suite $(n_k)_{k\geqslant 1}$ tels que $|F_{n_k}(y) - F(y)| > \varepsilon$ pour tout k. Montrer qu'alors $|F_{n_k}(x) - F(x)| > \frac{\varepsilon}{2}$ pour $x \in [y_1, y[$ ou pour $x \in]y, y_2]$, pour des réels y_1, y_2 adéquats. Utiliser ceci pour obtenir une contradiction.)

18.* Soit $(F_n)_{n\geqslant 1}$ une suite de fonctions de répartition convergeant simplement vers une fonction de répartition F *continue*. Montrer que la convergence de F_n vers F est uniforme sur $\mathbb{R}$. (*Indication :* Commencer par montrer qu'il existe des points $x_1 < \cdots < x_m$ tels que $F(x_1) < \varepsilon$, $F(x_{j+1}) - F(x_j) < \varepsilon$, et $1 - F(x_m) < \varepsilon$. Puis montrer qu'il existe N tel que si $n > N$ on ait $|F_n(x_j) - F(x_j)| < \varepsilon$, $1 \leqslant j \leqslant m$.)

19. Soit f une fonction uniformément continue sur $\mathbb{R}$, et X et Y deux v.a. réelles. Supposons que $|f(x) - f(y)| < \varepsilon$ si $|x - y| < \delta$. Montrer que

$$|E(f(X)) - E(f(X+Y))| \leqslant \varepsilon + 2\sup_x |f(x)| P(|Y| \geqslant \delta).$$

20.* (Pollard [20].) Soit $(X_n)_{n\geqslant 1}$, X et Y des v.a. réelles, toutes définies sur le même espace, et supposons que $X_n + \sigma Y$ converge en loi vers $X + \sigma Y$ pour tout $\sigma > 0$. Montrer que X_n converge en loi vers X. (*Indication :* Utiliser l'exercice 19.)

21. (Pollard [20].) Soit X et Y des v.a. réelles indépendantes, avec Y de loi N(0, 1). Soit f une fonction continue bornée. Montrer que

$$E(f(X + \sigma Y)) = E(f_\sigma(X))$$

où

$$f_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} f(z) e^{-\frac{1}{2}|z-x|^2/\sigma^2} dz.$$

Montrer que f_σ est bornée et $\mathcal{C}^\infty$.

22. Soit $(X_n)_{n\geqslant 1}$ et X des v.a. réelles. Montrer que X_n converge en loi vers X si et seulement si $E(f(X_n)) \to E(f(X))$ pour toute f bornée et $\mathcal{C}^\infty$. (*Indication :* Utiliser les exercices 20 et 21.)

Chapitre 19

Convergence en loi et fonctions caractéristiques

La convergence en loi est d'une certaine manière le cœur des probabilités et de la statistique ; on va voir qu'il existe des rapports très étroits entre convergence en loi et convergence des fonctions caractéristiques, ce qui est l'une des raisons pour lesquelles les fonctions caractéristiques sont aussi utiles.

Théorème 19.1. (Théorème de continuité de P. Lévy.) *Soit* $(\mu_n)_{n\geqslant 1}$ *une suite de probabilités sur* $\mathbb{R}^d$, *et* $(\widehat{\mu}_n)_{n\geqslant 1}$ *la suite de leurs transformées de Fourier (ou fonctions caractéristiques).*

(a) *Si* μ_n *converge étroitement vers une probabilité* μ, *alors* $\widehat{\mu}_n$ *converge simplement vers la transformée de Fourier* $\widehat{\mu}$ *de* μ.

(b) *Si* $\widehat{\mu}_n(u)$ *converge simplement vers une fonction* f, *et si de plus cette fonction est* continue au point 0, *alors il existe une probabilité* μ *sur* $\mathbb{R}^d$ *dont la transformée de Fourier est* $\widehat{\mu} = f$, *et* μ_n *converge étroitement vers* μ.

Preuve. (a) Supposons que μ_n converge étroitement vers μ. Comme la fonction $x \mapsto e^{iux}$ est continue et bornée,

$$\widehat{\mu}_n(u) = \int e^{iux}\,\mu_n(dx)$$

converge vers

$$\widehat{\mu}(u) = \int e^{iux}\mu(dx)$$

(cette fonction est à valeurs complexes, mais on peut prendre séparément les parties réelle et imaginaire.)

(b) Bien que le théorème soit énoncé pour $\mathbb{R}^d$ nous allons le démontrer pour $\mathbb{R}$ seulement. Supposons que $\lim_{n\to\infty} \widehat{\mu}_n(u) = f(u)$ existe pour tout u. On va montrer tout d'abord la *tension* (voir p. 164) de la suite μ_n. D'après le théorème de Fubini (théorème 10.3, et plus précisément exercice 10 du chapitre 14) on a

$$\begin{aligned}\int_{-\alpha}^{\alpha} \widehat{\mu}_n(u)du &= \int_{-\alpha}^{\alpha} \left\{\int_{-\infty}^{\infty} e^{iux}\mu_n(dx)\right\} du\\ &= \int_{-\infty}^{\infty} \left\{\int_{-\alpha}^{\alpha} e^{iux} du\right\} \mu_n(dx);\end{aligned}$$

et comme $e^{iux} = \cos ux + i \sin ux$,

$$= \int_{-\infty}^{\infty} \left\{\int_{-\alpha}^{\alpha} (\cos ux + i \sin ux)\, du\right\} \mu_n(dx).$$

$\sin ux$ étant une fonction impaire, l'intégrale de la partie imaginaire sur l'intervalle symétrique $[-\alpha, \alpha]$ est nulle, donc

$$= \int_{-\infty}^{\infty} \frac{2}{x} \sin \alpha x \ \mu_n(dx).$$

Comme $\int_{-\alpha}^{\alpha} 1 \ du = 2\alpha$, il vient

$$\begin{aligned}\frac{1}{\alpha}\int_{-\alpha}^{\alpha} (1 - \widehat{\mu}_n(u))du &= 2 - \int_{-\infty}^{\infty} \frac{2}{\alpha x} \sin \alpha x \ \mu_n(dx)\\ &= 2\int_{-\infty}^{\infty} \left(1 - \frac{\sin \alpha x}{\alpha x}\right) \mu_n(dx).\end{aligned}$$

Comme $2(1 - \frac{\sin v}{v})$ est plus grand que 1 si $|v| \geqslant 2$ et toujours positif, l'expression ci-dessus est

$$\begin{aligned}&\geqslant \int_{-\infty}^{\infty} 1_{[-2,2]^c}(\alpha x)\mu_n(dx)\\ &= \int 1_{[-2/\alpha, 2/\alpha]^c}(x)\mu_n(dx)\\ &= \mu_n\left(\left[\frac{-2}{\alpha}, \frac{2}{\alpha}\right]^c\right).\end{aligned}$$

Soit $\beta = \frac{2}{\alpha}$. On obtient alors l'estimée suivante :

$$\mu_n([-\beta, \beta]^c) \leqslant \frac{\beta}{2}\int_{-2/\beta}^{2/\beta} (1 - \widehat{\mu}_n(u))du. \tag{19.1}$$

Soit $\varepsilon > 0$. Comme *par hypothèse f est continue en* 0, il existe $\alpha > 0$ tel que $|1 - f(u)| \leqslant \varepsilon/4$ si $|u| \leqslant 2/\alpha$ (car $\widehat{\mu}_n(0) = 1$ pour tout n, donc aussi $f(0) = 1$). Donc

$$\left|\frac{\alpha}{2}\int_{-2/\alpha}^{2/\alpha} (1 - f(u))du\right| \leqslant \frac{\alpha}{2}\int_{-2/\alpha}^{2/\alpha} \frac{\varepsilon}{4}du = \frac{\varepsilon}{2}. \tag{19.2}$$

Comme les $\widehat{\mu}_n(u)$ sont des fonctions caractéristiques, elles vérifient $|\widehat{\mu}_n(u)| \leqslant 1$ et le théorème de convergence dominée implique que

$$\lim_{n\to\infty} \int_{-2/\alpha}^{2/\alpha} (1 - \widehat{\mu}_n(u))du = \int_{-2/\alpha}^{2/\alpha} (1 - f(u))du.$$

Il existe donc N tel que pour tout $n \geqslant \mathrm{N}$ on ait

$$\left| \int_{-2/\alpha}^{2/\alpha} (1 - \widehat{\mu}_n(u))du - \int_{-2/\alpha}^{2/\alpha} (1 - f(u))du \right| \leqslant \frac{\varepsilon}{\alpha},$$

d'où $\frac{\alpha}{2} \int_{-2/\alpha}^{2/\alpha} (1 - \widehat{\mu}_n(u))du \leqslant \varepsilon$. En utilisant enfin (19.1) on arrive à $\mu_n([-\alpha, \alpha]^c) \leqslant \varepsilon$ pour tout $n \geqslant \mathrm{N}$.

Il n'y a qu'un nombre fini d'entiers $n < \mathrm{N}$, et pour chacun d'entre eux il existe un α_n tel que $\mu_n([-\alpha_n, \alpha_n]^c) \leqslant \varepsilon$. Il reste à poser $a = \max(\alpha_1, \ldots, \alpha_n; \alpha)$ pour obtenir

$$\mu_n([-a, a]^c) \leqslant \varepsilon, \qquad \text{pour tout } n.$$

Comme $\varepsilon > 0$ est arbitraire, on a donc montré que les μ_n vérifient (18.8).

On peut maintenant appliquer le théorème 18.6 de façon à obtenir une sous-suite $(n_k)_{k\geqslant 1}$ telle que μ_{n_k} converge étroitement vers une certaine probabilité limite μ. La partie (a) entraîne alors que

$$\lim_{k\to\infty} \widehat{\mu}_{n_k}(u) = \widehat{\mu}(u)$$

pour tout u, donc $f = \widehat{\mu}$ et f est la transformée de Fourier d'une probabilité.

Il reste à montrer que la suite $(\mu_n)_{n\geqslant 1}$ elle-même (et pas seulement une sous-suite) converge étroitement vers μ. Notons F_n et F les fonctions de répartition de μ_n et de μ. Soit aussi D l'ensemble des points de continuité de F, i.e

$$\mathrm{D} = \{x : \mathrm{F}(x-) = \mathrm{F}(x)\}.$$

Supposons alors que μ_n ne converge pas étroitement vers μ. D'après le théorème 18.4 il existe au moins un point $x \in \mathrm{D}$ et une sous-suite $(n_k)_{k\geqslant 1}$ tels que $\beta = \lim_{k\to\infty} \mathrm{F}_{n_k}(x)$ existe et soit différente de $\mathrm{F}(x)$. D'après le théorème 18.6 il existe une sous-suite de la sous-suite (n_k), soit $(n_{k_j})_{j\geqslant 1}$, telle que $(\mu_{n_{k_j}})_{j\geqslant 1}$ converge étroitement vers une limite ν. Exactement comme plus haut, on obtient que

$$\lim_{j\to\infty} \widehat{\mu}_{n_{k_j}}(u) = \widehat{\nu}(u),$$

et comme $\lim_n \widehat{\mu}_n = f$ on en déduit que $\widehat{\nu} = f$. Comme on a vu que $f = \widehat{\mu}$, en vertu du théorème d'unicité 14.1 on a $\mu = \nu$. Mais alors le théorème 18.4 entraîne que $\beta = \lim_{j\to\infty} F_{n_{k_j}}(x) = F(x)$, puisque x appartient à l'ensemble des points de continuité de μ, ce qui fournit une contradiction. ■

Remarque 19.1. On pourrait en fait montrer plus : si les μ_n convergent étroitement vers μ, alors les fonctions $\widehat{\mu}_n$ convergent vers $\widehat{\mu}$ *uniformément sur tout compact de* $\mathbb{R}^d$.

Exemple. Soit $(X_n)_{n\geq 1}$ une suite de v.a. de Poisson de paramètres $\lambda_n = n$. Alors, si $Z_n = \frac{1}{\sqrt{n}}(X_n - n)$, on a

$$Z_n \xrightarrow{\mathcal{L}} Z, \quad \text{où } Z \text{ est de loi } N(0,1).$$

Pour le voir, on remarque que

$$\begin{aligned} E\left(e^{iuZ_n}\right) &= E\left(e^{iu\left(\frac{1}{\sqrt{n}}(X_n - n)\right)}\right) \\ &= e^{-iu\sqrt{n}} E\left(e^{i\frac{u}{\sqrt{n}}X_n}\right) \\ &= e^{-iu\sqrt{n}} e^{n\left(e^{iu/\sqrt{n}}-1\right)} \end{aligned}$$

(cf. exemple 13.3).

En faisant un développement de Taylor pour e^z, on voit que

$$\begin{aligned} E\left(e^{iuZ_n}\right) &= e^{-iu\sqrt{n}} e^{n\left(i\frac{u}{\sqrt{n}} - \frac{u^2}{2n} - \frac{iu^3}{6n^{3/2}} + \dots\right)} \\ &= e^{-iu\sqrt{n}+iu\sqrt{n}} e^{-u^2/2} e^{-\frac{h(u,n)}{\sqrt{n}}} \\ &= e^{-u^2/2} e^{-\frac{h(u,n)}{\sqrt{n}}} \end{aligned}$$

où $h(u,n)$ reste borné en n pour tout u et donc $\lim_{n\to\infty} \frac{h(u,n)}{\sqrt{n}} = 0$. Par suite

$$\lim_{n\to\infty} \varphi_{Z_n}(u) = e^{-u^2/2},$$

et comme $e^{-u^2/2}$ est la fonction caractéristique de la loi $N(0,1)$ (exemple 13.5), on obtient notre assertion en appliquant le théorème 19.1(b).

Exercices

1. Soit $(X_n)_{n\geqslant 1}$ des v.a. de lois $N(\mu_n, \sigma_n^2)$. Montrer que si $\mu_n \to \mu \in \mathbb{R}$ et $\sigma_n^2 \to \sigma^2 \geqslant 0$, alors $X_n \xrightarrow{\mathcal{L}} X$, où X est de loi $N(\mu, \sigma^2)$.

2. Soit $(X_n)_{n\geqslant 1}$ des v.a. de loi $N(\mu_n, \sigma_n^2)$, et supposons que $X_n \xrightarrow{\mathcal{L}} X$. Montrer que les suites μ_n et σ_n^2 ont des limites $\mu \in \mathbb{R}$ et $\sigma^2 \geqslant 0$, et que X est $N(\mu, \sigma^2)$ (*Indication :* φ_{X_n} et φ_X étant les fonctions caractéristiques, écrire $\varphi_{X_n} = e^{iu\mu - \frac{u^2\sigma_n^2}{2}}$ et utiliser le théorème de Lévy pour obtenir que $\varphi_X(u) = e^{iu\mu - \frac{u^2\sigma^2}{2}}$ pour un $\mu \in \mathbb{R}$ et un $\sigma^2 \geqslant 0$.)

3. Soit $(X_n)_{n\geqslant 1}$ et $(Y_n)_{n\geqslant 1}$ des suites de v.a. à valeurs dans $\mathbb{R}^d$, définies sur le même espace de probabilité. On suppose d'une part que X_n et Y_n sont indépendantes pour chaque n et que X est indépendante de Y, d'autre part que $X_n \xrightarrow{\mathcal{L}} X$ et $Y_n \xrightarrow{\mathcal{L}} Y$. Montrer que $X_n + Y_n \xrightarrow{\mathcal{L}} X + Y$.

Chapitre 20

La loi des grands nombres

Nous présentons dans ce chapitre l'un des résultats fondamentaux de la théorie des probabilités, celui qui en particulier « justifie » l'approche des probabilités par les fréquences, ou dans un tout autre domaine l'utilisation des méthodes de simulation de v.a. (méthodes de Monte-Carlo) pour l'intégration numérique (voir l'exemple 2).

Soit $(X_n)_{n\geqslant 1}$ une suite de v.a. réelles, définies sur le même espace de probabilité, et posons $S_n = \sum_{j=1}^{n} X_j$. Une assertion qui affirme que la suite $\frac{1}{n}S_n$, c'est-à-dire la suite des moyennes des n premières variables X_j, converge en un certain sens vers une limite s'appelle une *loi des grands nombres*. Il existe de nombreux résultats de ce type, par exemple les « théorèmes ergodiques » dans L^2 ou le théorème ergodique ponctuel, qui sont des formes de la loi des grands nombres (cf. le théorème 20.3 ci-dessous). La convergence peut être en probabilité, dans L^p, ou presque sûre ; dans ce dernier cas, on parle de *loi forte des grands nombres.*

Théorème 20.1. (Loi forte des grands nombres.) *Soit $(X_n)_{n\geqslant 1}$ des v.a. indépendantes et de même loi (« i.i.d. »), de carré intégrable, et posons*

$$\mu = E(X_j), \quad \textit{et} \quad \sigma^2 = \sigma^2_{X_j} < \infty.$$

Soit $S_n = \sum_{j=1}^{n} X_j$. On a alors

$$\lim_{n\to\infty} \frac{S_n}{n} = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} X_j = \mu \qquad \textit{p.s. et dans } L^2.$$

Remarque 20.1. On écrit μ et σ^2 puisque, les v.a. X_n ayant même loi, elles ont mêmes moments. On déduit de ce qui précède la convergence $\frac{S_n}{n} \to \mu$ en probabilité, puisque les convergences dans L^2 et p.s. impliquent chacune la convergence en probabilité. En fait la convergence en probabilité est très facile à obtenir dans ce contexte (en utilisant l'inégalité de Bienaymé-Chebyshev), et cette propriété est appelée la loi faible des grands nombres. La preuve du théorème 20.1 est plus simple lorsqu'on suppose que les X_n sont dans L^4. À l'inverse, un résultat plus fort, ne nécessitant que l'existence de la moyenne μ, est énoncé au théorème 20.2 et prouvé dans le chapitre 27.

Preuve du théorème 20.1. On peut supposer sans perdre de généralité que $\mu = E(X_j) = 0$: sinon, en effet, on peut remplacer X_j par $Z_j = X_j - \mu$ et obtenir $\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} Z_j = 0$, soit

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu) = \lim_{n\to\infty}\left(\frac{1}{n}\sum X_j\right) - \mu = 0,$$

ce qui donne le résultat.

On suppose donc dans la suite que $\mu = 0$. Rappelons que $S_n = \sum_{j=1}^{n} X_j$ et posons $Y_n = \frac{S_n}{n}$. On a $E(Y_n) = \frac{1}{n}\sum_{j=1}^{n} E(X_j) = 0$. De plus $E(Y_n^2) = \frac{1}{n^2}\sum_{1\leqslant j,k\leqslant n} E(X_j X_k)$. Cependant si $j \neq k$ on a

$$E(X_j X_k) = E(X_j)E(X_k) = 0$$

puisque X_j et X_k sont indépendants, donc

$$\begin{aligned} E(Y_n^2) &= \frac{1}{n^2}\sum_{j=1}^{n} E(X_j^2) \qquad (20.1)\\ &= \frac{1}{n^2}\sum_{j=1}^{n}\sigma^2 = \frac{1}{n^2}(n\sigma^2)\\ &= \frac{\sigma^2}{n} \end{aligned}$$

et par suite $\lim E(Y_n^2) = 0$, ce qui montre la convergence de Y_n vers 0 dans L^2.

La convergence dans L^2 impliquant la convergence en probabilité, on sait qu'on peut extraire de la suite (Y_n) une sous-suite qui converge p.s. vers 0. Cependant, cela ne suffit pas : nous voulons que *la suite* Y_n *elle-même* converge p.s. Pour le montrer nous exhibons d'abord une sous-suite particulière qui converge p.s., puis nous traitons les termes qui se trouvent entre deux éléments successifs de la sous-suite.

On a $E(Y_n^2) = \frac{\sigma^2}{n}$, donc

$$\sum_{n=1}^{\infty} E(Y_{n^2}^2) = \sum_{n=1}^{\infty}\frac{\sigma^2}{n^2} < \infty;$$

par suite, d'après le théorème 9.2 on a $\sum_{n=1}^{\infty} Y_{n^2}^2 < \infty$ p.s., donc le terme général de cette série est p.s. convergent vers 0 :

$$\lim_{n\to\infty} Y_{n^2} = 0 \quad \text{p.s.} \qquad (20.2)$$

Soit ensuite $n \in \mathbb{N}$, et $p(n)$ l'unique entier tel que

$$p(n)^2 \leqslant n < (p(n)+1)^2.$$

On a

$$\mathrm{Y}_n - \frac{p(n)^2}{n}\mathrm{Y}_{p(n)^2} = \frac{1}{n}\sum_{j=p(n)^2+1}^{n} \mathrm{X}_j,$$

et comme pour (20.1) nous obtenons

$$\begin{aligned} \mathrm{E}\left(\left(\mathrm{Y}_n - \frac{p(n)^2}{n}\mathrm{Y}_{p(n)^2}\right)^2\right) &= \frac{n-p(n)^2}{n^2}\sigma^2 \\ &\leqslant \frac{2p(n)+1}{n^2}\sigma^2, \\ &\leqslant \frac{2\sqrt{n}+1}{n^2}\sigma^2 \leqslant \frac{3}{n^{\frac{3}{2}}}\sigma^2 \end{aligned}$$

en utilisant le fait que $p(n) \leqslant \sqrt{n}$.

Appliquons maintenant le meme argument que précédemment ; on a

$$\sum_{n=1}^{\infty} \mathrm{E}\left(\left(\mathrm{Y}_n - \frac{p(n)^2}{n}\mathrm{Y}_{p(n)^2}\right)^2\right) \leqslant \sum_{n=1}^{\infty} \frac{3\sigma^2}{n^{\frac{3}{2}}} < \infty,$$

de sorte que

$$\sum_{n=1}^{\infty}\left(\mathrm{Y}_n - \frac{p(n)^2}{n}\mathrm{Y}_{p(n)^2}\right)^2 < \infty \quad \text{p.s.},$$

et *a fortiori*

$$\lim_{n\to\infty}\left\{\mathrm{Y}_n - \frac{p(n)^2}{n}\mathrm{Y}_{p(n)^2}\right\} = 0 \quad \text{p.s.}$$

Mais comme $\lim_{n\to\infty} \mathrm{Y}_{p(n)^2} = 0$ p.s. par (20.2) et $\frac{p(n)^2}{n} \to 1$, on déduit que $\lim_{n\to\infty} \mathrm{Y}_n = 0$ p.s. également, et le théorème est démontré. ■

Nous donnons ci-dessous deux autres versions de la loi des grands nombres.

Théorème 20.2. (Loi des grands nombres de Kolmogorov.) *Soit* (X_j) *des v.a. réelles i.i.d., et soit* $\mu \in \mathbb{R}$. *Posons* $\mathrm{S}_n = \sum_{j=1}^{n} \mathrm{X}_j$. *La suite* $\frac{\mathrm{S}_n}{n}$ *converge p.s. vers* μ *si et seulement si* $\mathrm{E}(\mathrm{X}_j) = \mu$ *(et si en particulier les* X_j *admettent une moyenne). Dans ce cas, la convergence a aussi lieu dans* L^1.

Remarque 20.2. Ce théorème nous donne les hypothèses minimales assurant la validité de la loi des grands nombres pour les sommes de v.a. indépendantes, à savoir que les X_j soient dans L^1. Une manière élégante de prouver le théorème 20.2 consiste à utiliser le théorème de convergence des martingales inverses (théorème 27.5).

Soit $(\Omega, \mathcal{A}, P)$ un espace de probabilité, et $T : \Omega \to \Omega$ une application injective mesurable (de $(\Omega, \mathcal{A})$ dans lui-même). On suppose aussi que T *préserve la mesure*, i.e. $P(T^{-1}(A)) = P(A)$ (ou si on veut : P est sa propre image par T). On pose $T^2(\omega) = T(T(\omega))$, et on définit de manière analogue les puissances successives de T. Une partie A de Ω est dite *invariante par* T si $T^{-1}(A) = A$, ou de manière équivalente si $1_A = 1_A \circ T$.

Théorème 20.3. (Loi des grands nombres ergodique.) *Soit* T *une transformation injective préservant la mesure et telle que les ensembles mesurables invariants soient de probabilité* 0 *ou* 1. *Si* X *est une v.a. intégrable, alors*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X(T^j(\omega)) = E(X)$$

p.s. et dans L^1.

Ce théorème est une conséquence du « théorème ergodique ponctuel », ou de Birkhoff. Il remplace l'hypothèse d'indépendance par l'hypothèse d'ergodicité (l'ergodicité est le fait que les ensembles invariants soient de probabilité 0 ou 1 : noter que les v.a. $X \circ T^j$ ne sont pas indépendantes, mais elles ont toutes même loi, puisque T préserve la mesure). Il s'appelle aussi *loi forte des grands nombres pour les suites stationnaires*.

Exemple 1. Dans l'exemple 17.1 on considérait une suite $(X_j)_{j\geqslant 1}$ de v.a. indépendantes de loi de Bernoulli, avec $P(X_j = 1) = p$ et $P(X_j = 0) = q = 1 - p$. Alors $S_n = \sum_{j=1}^n X_j$ est le nombre de « succès » en n tirages, et S_n/n est la fréquence des succès. La loi forte des grands nombres nous dit que

$$\lim_{n\to\infty} \frac{S_n}{n} = E(X_1) = p \qquad \text{p.s.} \tag{20.3}$$

Cela fournit donc la justification souhaitée à l'approche des probabilités par les fréquences : les axiomes des probabilités, justifiés de manière intuitive par le résultat postulé (20.3), permettent en fait de montrer que ce résultat est mathématiquement correct.

Exemple 2. Nous présentons ci-dessous un exemple très simple d'une technique connue sous le nom de *méthode d'approximation de Monte-Carlo.* Soit f une fonction borélienne sur $[0,1]$, intégrable par rapport à la mesure de Lebesgue. Il est souvent impossible de trouver une expression analytique pour l'intégrale $\mathrm{I} = \int_0^1 f(x)dx$. Pour calculer le nombre I on peut opérer ainsi : on prend une suite de v.a. $(\mathrm{U}_j)_{j\geqslant 1}$ indépendantes et uniformément distribuées sur l'intervalle $[0,1]$, et on pose $\mathrm{I}_n = \frac{1}{n}\sum_{j=1}^n f(\mathrm{U}_j)$; il suit du théorème 20.2 que

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^n f(\mathrm{U}_j) = \mathrm{E}(f(\mathrm{U}_j)) = \int_0^1 f(x)dx,$$

p.s. et dans L^1. Si on simule sur un ordinateur la suite $(\mathrm{U}_j)_{j\geqslant 1}$ en utilisant un générateur de nombres au hasard, on obtient ainsi une approximation I_n de l'intégrale I.

Bien entendu, il ne s'agit là que d'une méthode parmi beaucoup pour calculer $\int_0^1 f(x)dx$, et certainement pas la meilleure ni la plus rapide pour une intégrale sur $[0,1]$. On verra par exemple, à l'occasion du théorème-limite central, que la « vitesse de convergence » de cette méthode est en $1/\sqrt{n}$, alors qu'une approximation de $\int_0^1 f(x)dx$ par la somme de Riemann $\frac{1}{n}\sum_{i=1}^n f(i/n)$ induit une erreur de l'ordre de $1/n$ dès que f est lipschitzienne, et que la « méthode des trapèzes » avec un pas $1/n$ conduit à une erreur de l'ordre de $1/n^2$ si f est deux fois dérivable, etc. En revanche si f est simplement borélienne la méthode de Monte-Carlo devient compétitive ; et si on remplace $[0,1]$ par, disons, un cube de l'espace $\mathbb{R}^d$, cette même méthode marche encore, toujours avec la vitesse $1/\sqrt{n}$, alors que les méthodes d'approximation analytiques conduisent à une vitesse dépendant de la dimension d et sont en fait impossibles à mettre en œuvre pratiquement dès que d dépasse 4 ou 5.

Exemple 3. ([10, p. 120]) Soit Ω le cercle de rayon $r = \frac{1}{2\pi}$. Soit $\mathcal{A}$ la tribu borélienne sur ce cercle, et P la mesure de Lebesgue (de ce point de vue, le cercle n'est autre que le segment $[0,1[$!). Soit α un nombre irrationnel et T la rotation sur Ω d'angle α. On vérifie aisément que T est injective, préserve la mesure, et que les boréliens du cercle invariants par T sont de probabilité 0 ou 1 (c'est là que l'irrationalité de α intervient). Donc le théorème 20.3 implique que si X est une fonction intégrable sur Ω,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^n \mathrm{X}(x + j\alpha) = \int_0^1 \mathrm{X}(y)dy$$

pour P-presque tout x.

Exercices

1.* (*Une loi faible des grands nombres.*) Soit (X_j) une suite de v.a. réelles telle que $\sup_j E(X_j^2) = c < \infty$ et $E(X_jX_k) = 0$ si $j \neq k$. Soit $S_n = \sum_{j=1}^n X_j$. Montrer que

   a) $P(|\frac{1}{n}S_n| \geqslant \varepsilon) \leqslant \frac{c}{n\varepsilon^2}$ pour tout $\varepsilon > 0$;

   b) $\lim_{n\to\infty} \frac{1}{n}S_n = 0$ dans L^2 et en probabilité.

   (*Commentaire :* L'hypothèse usuelle d'indépendance a été considérablement affaiblie ici.)

2. Soit $(Y_j)_{j\geqslant 1}$ une suite de v.a. indépendantes binomiales de paramètres $(1, p)$, sur le même espace de probabilité. Soit $X_n = \sum_{j=1}^n Y_j$. Montrer que X_j est binomiale de paramètre j, p et que $\frac{X_j}{j}$ converge p.s. vers p.

3. Soit $(X_j)_{j\geqslant 1}$ des variables i.i.d. réelles intégrables, et $Y_j = e^{X_j}$. Montrer que

$$\left(\prod_{j=1}^n Y_j\right)^{\frac{1}{n}}$$

   converge p.s. vers une constante α. [*Rép. :* $\alpha = e^{E(X_1)}$.]

4. Soit $(X_j)_{j\geqslant 1}$ des variables i.i.d. réelles intégrables de moyenne μ. Soit $(Y_j)_{j\geqslant 1}$ également i.i.d. réelles intégrables, de moyenne ν. Montrer que

$$\lim_{n\to\infty} \frac{1}{\sum_{j=1}^n Y_j} \sum_{j=1}^n X_j = \frac{\mu}{\nu} \qquad \text{p.s.}$$

5. Soit $(X_j)_{j\geqslant 1}$ des variables i.i.d. réelles intégrables, et supposons que pour une certaine constante ν la suite $\frac{1}{\sqrt{n}}\sum_{j=1}^n (X_j - \nu)$ converge en loi vers une limite Z. Montrer que

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n X_j = \nu \qquad \text{p.s.}$$

   (*Indication :* Si $Z_n = \frac{1}{\sqrt{n}} \sum_{j=1}^n (X_j - \nu)$, montrer d'abord que $\frac{1}{\sqrt{n}}Z_n$ converge en loi vers 0.)

6. Soit $(X_j)_{j\geqslant 1}$ des variables i.i.d. réelles appartenant à L^p pour un $p \in [1, \infty[$. Montrer que

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X_j^p = E(X^p) \qquad \text{p.s.}$$

7. Soit $(X_j)_{j\geqslant 1}$ des v.a. indépendantes $N(1,3)$. Montrer que

$$\lim_{n\to\infty} \frac{X_1 + X_2 + \cdots + X_n}{X_1^2 + X_2^2 + \cdots + X_n^2} = \frac{1}{4} \qquad \text{p.s.}$$

8. Soit $(X_j)_{j\geqslant 1}$ des variables i.i.d. réelles intégrables de moyennes μ et de variance σ^2. Montrer que

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 = \sigma^2 \qquad \text{p.s.}$$

9. Soit $(X_j)_{j\geqslant 1}$ des variables i.i.d. à valeurs entières et intégrables. Posons $S_n = \sum_{j=1}^{n} X_j$. Montrer que si $E(X_j) > 0$, alors

$$\lim_{n\to\infty} S_n = \infty \qquad \text{p.s.}$$

Chapitre 21

Le théorème-limite central

Le théorème-limite central est l'un des résultats les plus impressionnants des probabilités. À partir d'hypothèses très faibles, il donne des résultats extrêmement précis. Il joue aussi un rôle de premier plan en statistique. L'idée générale est simple. Soit $X_1, \ldots, X_j, \ldots$ des variables aléatoires i.i.d. de variance finie σ^2 et de moyenne notée μ, et soit $S_n = \sum_{j=1}^n X_j$. Alors, si n est grand, $\mathcal{L}(S_n) \approx N(n\mu, n\sigma^2)$. Le point clé dans ce genre de résultat est que, à part l'existence d'une variance, *absolument aucune hypothèse n'est requise*. Dons, si une certaine v.a. est en fait la somme de « beaucoup » de variables indépendantes de variances finies, alors sa loi est approximativement normale ; en statistique par exemple, il est alors possible d'utiliser des procédures permettant d'estimer μ et σ^2, et on connaît alors essentiellement tout !

Théorème 21.1. (Théorème-limite central.) *Soit* $(X_j)_{j\geq 1}$ *une suite i.i.d. de variables aléatoires réelles avec* $E(X_j) = \mu$ *et* $\mathrm{Var}(X_j) = \sigma^2 \in]0, \infty[$. *Soit*

$$S_n = \sum_{j=1}^n X_j \quad \textit{et} \quad Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}\,.$$

Alors, les v.a. Y_n *convergent en loi vers une variable* $N(0, 1)$.

Noter que si $\sigma^2 = 0$, alors $X_j = \mu$ p.s. pour tout j, donc $\frac{S_n}{n} = \mu$ p.s.

Preuve. Soit φ la fonction caractéristique des v.a. $X_j - \mu$ (qui ont toutes même loi par hypothèse), et $Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$. Comme les X_j sont indépendantes, le théorème 15.2 entraîne que

$$\begin{aligned}
\varphi_{Y_n}(u) &= \varphi_{\frac{1}{\sigma\sqrt{n}}\sum_{j=1}^n (X_j - \mu)}(u) && (21.1)\\
&= \varphi_{\sum_{j=1}^n (X_j-\mu)}\left(\frac{u}{\sigma\sqrt{n}}\right)\\
&= \prod_{j=1}^n \varphi\left(\frac{u}{\sigma\sqrt{n}}\right)\\
&= \left(\varphi\left(\frac{u}{\sigma\sqrt{n}}\right)\right)^n .
\end{aligned}$$

Ensuite, le fait que $E(X_j - \mu) = 0$ et $E((X_j - \mu)^2) = \sigma^2 < \infty$ entraîne par le théorème 13.2 que φ est deux fois continûment dérivable et que

$$\varphi'(u) = iE\left((X_j - \mu)e^{iu(X_j-\mu)}\right),$$
$$\varphi''(u) = -E\left((X_j - \mu)^2 e^{iu(X_j-\mu)}\right).$$

Donc $\varphi'(0) = 0$ et $\varphi''(0) = -\sigma^2$. Si on fait le développement de Taylor de la fonction φ à l'ordre 2 au voisinage du point 0, on obtient

$$\varphi(u) = 1 + 0 - \frac{\sigma^2 u^2}{2} + u^2 h(u) \tag{21.2}$$

où $h(u) \to 0$ quand $u \to 0$. Rappelons que d'après (21.1),

$$\begin{aligned}\varphi_{Y_n}(u) &= \left(\varphi\left(\frac{u}{\sigma\sqrt{n}}\right)\right)^n \\ &= e^{n \log \varphi\left(\frac{u}{\sigma\sqrt{n}}\right)} \\ &= e^{n \log\left(1 - \frac{u^2}{2n} + \frac{u^2}{n\sigma^2} h\left(\frac{u}{\sigma\sqrt{n}}\right)\right)},\end{aligned}$$

où « log » désigne la valeur principale du logarithme complexe (valant 0 au point 1 et continue dans le cercle complexe de centre 1 et de rayon 1/2). En prenant les limites quand $n \to \infty$ on voit que

$$\lim_{n\to\infty} \varphi_{Y_n}(u) = e^{-u^2/2}.$$

Le théorème de Lévy 19.1 implique alors que Y_n converge en loi vers Z, où $\varphi_Z(u) = e^{-u^2/2}$; d'après l'exemple 13.5 on sait que Z suit la loi $N(0, 1)$ (rappelons que la fonction caractéristique caractérise la loi). ∎

Examinons à présent les rapports entre la loi des grands nombres et le théorème-limite central. Soit $(X_j)_{j\geqslant 1}$, S_n, μ et σ^2 comme dans le théorème précédent. La loi des grands nombres nous dit que

$$\lim_{n\to\infty} \frac{S_n}{n} = \mu \qquad \text{p.s. et dans } L^2, \tag{21.3}$$

La limite de S_n/n est donc μ, mais la question suivante se pose alors naturellement : à quelle vitesse cette convergence a-t-elle lieu ? Réécrivons (21.3) ainsi :

$$\lim_{n\to\infty} \left|\frac{S_n}{n} - \mu\right| = 0 \qquad \text{p.s. et dans } L^2. \tag{21.4}$$

On peut alors se poser la question de savoir s'il existe un réel $\alpha > 0$ tel que, pour une certaine constante $c \neq 0$,

$$\lim_{n\to\infty} n^{\alpha} \left| \frac{\mathrm{S}_n}{n} - \mu \right| = c \qquad \text{p.s.}$$

Il n'existe en fait aucun α ayant cette propriété ; les v.a. $n^{\alpha}(\frac{\mathrm{S}_n}{n} - \mu)$ ne peuvent converger vers une constante non nulle ou vers une variable aléatoire non p.s. nulle, ni presque sûrement, ni même en probabilité. Cependant le théorème-limite central nous dit que si on prend $\alpha = \frac{1}{2}$, alors $\sqrt{n}(\frac{\mathrm{S}_n}{n} - \mu)$ converge en loi vers une v.a. $\mathrm{N}(0, \sigma^2)$. En ce sens, la vitesse de convergence pour la loi des grands nombres est $\sqrt{n}$.

On peut légèrement affaiblir les hypothèses du théorème 21.1 ; la même démonstration permet de démontrer le

Théorème 21.2. (Théorème-limite central.) *Soit* $(\mathrm{X}_j)_{j\geqslant 1}$ *une suite de v.a. réelles indépendantes, mais pas nécessairement de même loi. On suppose qu'elles sont centrées :* $\mathrm{E}(\mathrm{X}_j) = 0$, *et on pose* $\sigma_j^2 = \sigma_{\mathrm{X}_j}^2$. *Supposons aussi*

$$\sup_j \mathrm{E}\left(|\mathrm{X}_j|^{2+\varepsilon}\right) < \infty, \qquad \sum_{j=1}^{\infty} \sigma_j^2 = \infty$$

pour un $\varepsilon > 0$. *Si alors* $\mathrm{S}_n = \mathrm{X}_1 + \cdots + \mathrm{X}_n$, *la suite* $\dfrac{\mathrm{S}_n}{\sqrt{\sum_{j=1}^{n} \sigma_j^2}}$ *converge en loi vers une v.a.* $\mathrm{N}(0,1)$.

Le théorème 21.1 est le théorème-limite central « classique », tandis que le théorème 21.2 montre qu'il est possible d'affaiblir les hypothèses tout en obtenant le même résultat. Il existe de nombreuses versions du théorème-limite central, toutes semblables par le fait qu'elles exhibent des conditions suffisantes pour qu'une somme de v.a. convenablement normalisée converge en loi vers une variable normale. Par exemple, la théorie des martingales permet d'affaiblir notablement les hypothèses du théorème 21.2 en introduisant une certaine dépendance entre les variables X_j : voir le théorème 27.7. L'extension à d'autres formes de dépendance, connues sous le nom de « conditions de mélange » est possible mais plus difficile.

Finalement, il est important pour les applications de souligner que le théorème 21.1 admet une version d-dimensionnelle, qui est démontrée exactement de la même manière :

Théorème 21.3. (Théorème-limite central.) *Soit $(X_j)_{j\geq 1}$ des variables aléatoires i.i.d. de vecteur moyenne $\mu = E(X_j)$ (i.e., μ_k est l'espérance de la k-ième composante X_j^k des X_j) et de matrice de covariance $Q = (q_{k,\ell})_{1\leq k,\ell\leq d}$ (donc $q_{k,\ell} = \mathrm{Cov}(X_j^k, X_j^\ell)$). Alors si $S_n = X_1 + \cdots + X_n$, les variables aléatoires $\frac{S_n - n\mu}{\sqrt{n}}$ convergent en loi vers une variable Z (à valeurs dans $\mathbb{R}^d$) de loi $N(0, Q)$.*

Dans ce théorème nous n'avons fait aucune hypothèse de non dégénérescence de la matrice de covariance Q : la limite Z est gaussienne, mais peut ne pas avoir de densité (rappelons que la loi $N(0, Q)$ admet une densité dans $\mathbb{R}^d$ si et seulement la matrice Q est inversible).

Exemples.

1. Soit $(X_j)_{j\geq 1}$ des v.a. indépendantes avec $P(X_j = 1) = p$ et $P(X_j = 0) = q = 1 - p$ (variables de Bernoulli). La v.a. $S_n = \sum_{j=1}^n X_j$ est binomiale de paramètres (n, p). On a $\mu = E(X_j) = p$ et $\sigma^2 = \sigma^2_{X_j} = pq = p(1-p)$. La loi des grands nombres dit que

$$\lim_{n\to\infty} \frac{S_n}{n} = p \qquad \text{p.s.,}$$

et le théorème-limite central dit que

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{L}} Z, \qquad \text{où Z est } N(0, 1).$$

2. Soit $(X_j)_{j\geq 1}$ une suite i.i.d. de variables aléatoires de carré intégrable et de fonction de répartition F. On suppose que cette fonction F est inconnue, et on souhaite « l'estimer ». Nous décrivons ci-dessous une technique standard pour résoudre ce problème. Posons

$$Y_j(x) = 1_{\{X_j \leq x\}}.$$

Les v.a. $Y_j(x)$ sont aussi i.i.d. et de carré intégrable : ce sont en fait des variables de Bernoulli de paramètre $p = F(x)$. Posons ensuite

$$F_n(x) = \frac{1}{n}\sum_{j=1}^n Y_j(x).$$

Les $F_n(x)$ sont des variables aléatoires pour tout x fixé ; pour ω fixé, la fonction $x \mapsto F_n(x) = F_n(x)(\omega)$ et clairement une fonction de répartition (celle de la probabilité $\frac{1}{n}\sum_{j=1}^n \varepsilon_{X_j(\omega)}$), appelée *fonction de répartition empirique.* La loi des grands nombres entraîne la convergence p.s. et dans L^2 de $F_n(x) = \frac{1}{n}\sum_{j=1}^n Y_j(x)$ vers $E(Y_1(x)) = F(x)$:

$$\lim_{n\to\infty} F_n(x) = F(x) \qquad \text{p.s. et dans } L^2.$$

Ainsi, les *fonctions de répartition empiriques* convergent vers la vraie fonction de répartition, pour chaque x fixé, au sens p.s. et L^2. Avec un petit travail supplémentaire on peut montrer la convergence uniforme :

$$\lim_{n\to\infty} \sup_x |F_n(x) - F(x)| = 0 \qquad \text{p.s. et dans } L^2.$$

Ce résultat est connu sous le nom de *théorème de Glivenko-Cantelli.* En utilisant le théorème-limite central on obtient de plus que la vitesse de convergence est $\sqrt{n}$: on a en effet

$$\begin{aligned}\sqrt{n}(F_n(x) - F(x)) &= \sqrt{n}\left(\frac{1}{n}\sum_{j=1}^n Y_j(x) - E(Y_1(x))\right)\\ &= \frac{\sum_{j=1}^n Y_j(x) - nE(Y_1(x))}{\sqrt{n}}.\end{aligned}$$

qui d'après le théorème 21.1 converge en loi vers une variable $N(0, \sigma^2(x))$, où $\sigma^2(x) = \operatorname{Var}(Y_1(x)) = F(x)(1 - F(x))$.

Ainsi qu'il a été dit plus haut, le théorème-limite central peut être interprété comme donnant la vitesse de convergence dans la loi des grands nombres. On peut aller un cran plus loin, en se posant le problème de la « vitesse de convergence » dans le théorème-limite central lui-même : à quelle vitesse la loi de $Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ approche-t-elle la loi $N(0,1)$? Le théorème suivant est l'une des réponses « classiques » à ce problème, lorsqu'on prend comme mesure de la « distance » entre deux lois sur $\mathbb{R}$ le supremum de la différence des fonctions de répartition.

Théorème 21.4. (Théorème de Berry-Esseen.) *Soit* $(X_j)_{j\geqslant 1}$ *des variables aléatoires réelles i.i.d., vérifiant* $E(|X_j|^3) < \infty$ *et* $\sigma^2 = \sigma^2_{X_j} > 0$. *Soit* $G_n(x) = P(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leqslant x)$, *où* $\mu = E(X_j)$ *et* $S_n = X_1 + \cdots + X_n$. *Soit enfin* Φ *la fonction de répartition de la loi* $N(0,1)$. *On a alors*

$$\sup_x |G_n(x) - \Phi(x)| \leqslant c\frac{E(|X_1|^3)}{\sigma^3\sqrt{n}}$$

pour une constante c.

La démonstration de ce théorème est un peu difficile pour ce livre, et nous l'omettons. Le lecteur peut consulter, par exemple [11, p. 108], où il est prouvé qu'on peut prendre $c = 3$. La meilleure approximation de c connue à ce jour est $c = 0.7975$.

Exercices

1. Soit $(X_j)_{j\geqslant 1}$ des variables indépendantes telles que $P(X_j = 1) = P(X_j = 0) = \frac{1}{2}$, et $S_n = \sum_{j=1}^{n} X_j$. La variable $Z_n = 2S_n - n$ représente le surplus de « faces » par rapport aux « piles » après n jets d'une pièce, si « Face » correspond à $X_j = 1$. Montrer que
$$\lim_{n\to\infty} P\left(\frac{Z_n}{\sqrt{n}} < x\right) = \Phi(x)$$
où
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du.$$

2. Soit $(X_j)_{j\geqslant 1}$ des v.a. indépendantes de loi exponentielle double de paramètre 1 (la densité est $\frac{1}{2}e^{-|x|}$). Montrer que $\sqrt{n}\left(\frac{\sum_{j=1}^{n} X_j}{\sum_{j=1}^{n} X_j^2}\right)$ converge en loi vers une variable $N(0, \frac{1}{2})$. (*Indication :* Utiliser le théorème de Slutsky (exercice 14 du chapitre 18).)

3. Construire une suite $(X_j)_{j\geqslant 1}$ de v.a. réelles indépendantes de carré intégrable, telle que $\lim_{j\to\infty} X_j = 1$ en probabilité et $E(X_j^2) \geqslant j$. Soit Y une v.a. de loi $N(0, 1)$, indépendantes de la suite (X_j). On pose $Z_j = YX_j$. Montrer que
   a) $E(Z_j) = 0$.
   b) $\lim_{j\to\infty} \sigma_{Z_j}^2 = \infty$
   c) Z_j converge en loi vers une v.a. $N(0, 1)$.

   (*Indication :* Pour construire X_j prendre $(\Omega, \mathcal{A}, P) = ([0, 1], \mathcal{B}[0, 1], m(ds))$ où m est la mesure de Lebesgue ; poser
$$X_j(\omega) = (j + 1)1_{[0,1/j]}(\omega) + 1_{(1/j,1]}(\omega).$$
   Pour (c) utiliser le théorème de Slutsky.) [Noter que les hypothèses du théorème-limite central présenté dans ce livre ne sont pas satisfaites ici : ces hypothèses sont suffisantes, mais pas nécessaires, pour avoir une limite normale.]

4. (Durrett, [11].) Soit $(X_j)_{j\geqslant 1}$ des v.a. réelles i.i.d. avec $E(X_1) = 1$ et $\sigma^2_{X_1} = \sigma^2 \in]0, \infty[$. Montrer que

$$\frac{2}{\sigma}(\sqrt{S_n} - \sqrt{n}) \xrightarrow{\mathcal{L}} Z,$$

où Z est N(0, 1). (*Indication :* $\frac{S_n - n}{\sqrt{n}} = \frac{(\sqrt{S_n}+\sqrt{n})}{\sqrt{n}}(\sqrt{S_n} - \sqrt{n})$.)

5. Soit (X_j) des variables indépendantes de loi de Poisson de paramètre $\lambda = 1$ et $S_n = \sum_{j=1}^{n} X_j$. Montrer que $\lim_{n\to\infty} \frac{S_n - n}{\sqrt{n}} = Z$, où Z est N(0, 1).

6. Soit Y^λ une v.a. de Poisson de paramètre $\lambda > 0$. Montrer que $\frac{Y^\lambda - \lambda}{\sqrt{\lambda}}$ converge en loi vers une v.a. N(0, 1) quand $\lambda \to \infty$. (*Indication :* Utiliser l'exercice 5 et comparer Y^λ avec $S_{[\lambda]}$ et $S_{[\lambda]+1}$, où $[\lambda]$ désigne la « partie entière » de λ(= le plus grand entier $n \leqslant \lambda$).)

7. Montrer que

$$\lim_{n\to\infty} e^{-n}\left(\sum_{k=0}^{n}\frac{n^k}{k!}\right) = \frac{1}{2}.$$

(*Indication :* utiliser l'exercice 5.)

8. Soit $(X_j)_{j\geqslant 1}$ des v.a. réelles i.i.d. avec $E(X_1) = 0$ et $\sigma^2_{X_1} = \sigma^2 \in]0, \infty[$, et $S_n = \sum_{i=1}^{n} X_i$. Montrer que $\frac{S_n}{\sqrt{n}}$ ne converge pas en probabilité.

9.* Soit $(X_j)_{j\geqslant 1}$ des v.a. réelles i.i.d. avec $E(X_1) = 0$ et $\sigma^2_{X_1} = \sigma^2 \in]0, \infty[$, et $S_n = \sum_{i=1}^{n} X_i$. Montrer que

$$\lim_{n\to\infty} E\left(\frac{|S_n|}{\sqrt{n}}\right) = \sqrt{\frac{2}{\pi}}\sigma.$$

(*Indication :* Calculer $E(|Z|)$ pour une v.a. Z de loi $N(0, \sigma^2)$.)

10. (Gut, [14].) Soit $(X_j)_{j\geqslant 1}$ des v.a. indépendantes uniformes sur $[-1, 1]$, et

$$Y_n = \frac{\sum_{j=1}^{n} X_j}{\sum_{j=1}^{n} X_j^2 + \sum_{j=1}^{n} X_j^3}.$$

Montrer que $\sqrt{n}Y_n$ converge. (*Rép. :* $\sqrt{n}Y_n$ converge en loi vers une v.a. N(0, 3).)

11. Soit $(X_j)_{j\geqslant 1}$ des v.a. indépendantes, chaque X_j étant uniforme sur $[-j, j]$, et $S_n = \sum_{i=1}^{n} X_i$.

a) Montrer que $\frac{S_n}{n^{3/2}}$ converge en loi vers une v.a. $N(0, \frac{1}{9})$. (*Indication :* Montrer que la fonction caractéristique de X_j est $\varphi_{X_j}(u) = \frac{\sin uj}{uj}$; calculer $\varphi_{S_n}(u)$, puis $\varphi_{S_n/n^{3/2}}(u)$, et prouver que la limite est $e^{-u^2/18}$ en utilisant $\sum_{j=1}^n j^2 = \frac{n(n+1)(2n+1)}{6}$).

b) Montrer que $\frac{S_n}{\sqrt{\sum_{j=1}^n \sigma_{X_j}^2}}$ converge en loi vers une v.a. $N(0, 1)$. [*Commentaire :* Ceci n'est *pas* un cas particulier du théorème 21.2].

12.* Soit $X \in L^2$ et supposons que les v.a. X et $\frac{1}{\sqrt{2}}(Y+Z)$ ont même loi, dès que Y et Z sont indépendantes et de même loi que X. Montrer que X suit une loi normale. (*Indication :* Montrer par récurrence que X a même loi que $\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ si les (X_i) sont i.i.d. de même loi que X, pour tout $n = 2^m$.)

Chapitre 22

L^2 et les espaces de Hilbert

Soit un espace de probabilité $(\Omega, \mathcal{F}, P)$. Rappelons que L^2 désigne l'ensemble des (classes d'équivalence pour la relation « égalité presque sûre » de) v.a. réelles de carré intégrable. Dans la suite on identifie de manière systématique une variable aléatoire et sa classe d'équivalence : ainsi, au lieu d'écrire $X = 0$ p.s. on écrit simplement $X = 0$.

L'espace L^2 est un espace vectoriel (cf. théorème 9.3), et on peut le munir d'un « produit scalaire » de la manière suivante : pour X et Y dans L^2 on pose

$$\langle X, Y\rangle = E(XY). \tag{22.1}$$

On a $|E(XY)| \leqslant E(X^2)^{\frac{1}{2}}E(Y^2)^{\frac{1}{2}} < \infty$ par l'inégalité de Cauchy-Schwarz. et la formule (22.1) a un sens. On a aussi $\langle X, Y\rangle = \langle Y, X\rangle$ (symétrie), et comme l'espérance est linéaire, il vient

$$\langle \alpha X + \beta Y, Z\rangle = \alpha\langle X, Z\rangle + \beta\langle Y, Z\rangle.$$

Finalement, on voit que

$$\langle X, X\rangle \geqslant 0, \qquad \text{et} \quad \langle X, X\rangle = 0 \ \text{ si et seulement si } \ X = 0$$

compte tenu de notre convention d'identifier une variable p.s. nulle avec 0. Cela conduit à poser pour $X \in L^2$:

$$\|X\| = \langle X, X\rangle^{\frac{1}{2}} = \big(E(X^2)\big)^{\frac{1}{2}}. \tag{22.2}$$

Donc $\|X\| = 0$ si et seulement si $X = 0$; de plus la bilinéarité du produit scalaire, vue ci-dessus, et l'inégalité de Cauchy-Schwarz conduisent à

$$\begin{aligned}
\|X + Y\|^2 &= E(X^2) + 2E(XY) + E(Y^2) \\
&\leqslant \|X\|^2 + 2\|X\|\,\|Y\| + \|Y\|^2 \\
&= (\|X\| + \|Y\|)^2,
\end{aligned}$$

et on obtient l'inégalité de Minkowski :

$$\|X + Y\| \leqslant \|X\| + \|Y\|.$$

On a donc ainsi défini une norme, et en fait montré le :

Théorème 22.1. L^2 *est un espace vectoriel normé, la norme étant associée à un produit scalaire, ce qui signifie que* $\|\cdot\| = \langle \cdot,\cdot \rangle^{\frac{1}{2}}$.

La convergence d'une suite de v.a. pour cette norme est clairement la « convergence dans L^2 » définie au chapitre 17. On peut faire mieux que dans le précédent théorème, en montrant que l'espace L^2 est, de plus, *complet* : cela signifie que toute suite de Cauchy dans L^2 est une suite convergente (une suite de Cauchy (X_n) est une suite telle que $\|X_n - X_m\| \to 0$ quand m et n tendent tous deux vers l'infini ; rappelons aussi que toute suite convergente est une suite de Cauchy). Le théorème ci-dessous est parfois connu sous le nom de théorème de *Riesz-Fischer*.

Théorème 22.2. L^2 *est complet pour la norme* (22.2).

Preuve. Soit X_n une suite de Cauchy dans L^2. Pour tout $\varepsilon > 0$ il existe N tel que $n, m \geqslant N$ entraîne $\|X_n - X_m\| \leqslant \varepsilon$. Choisissons une suite de epsilons de la forme $\frac{1}{2^n}$: il existe alors une sous-suite $(X_{n_k})_{k\geqslant 1}$ telle que $\|X_{n_k} - X_{n_{k+1}}\| \leqslant \frac{1}{2^k}$.

Posons
$$Y_n = \sum_{p=1}^{n} |X_{n_p} - X_{n_{p+1}}|.$$
L'inégalité triangulaire entraîne
$$E(Y_n^2) \leqslant \left(\sum_{p=1}^{n} \|X_{n_p} - X_{n_{p+1}}\|\right)^2 \leqslant 1.$$
La limite $Y(\omega) = \lim_{n\to\infty} Y_n(\omega)$ existe pour tout ω, puisque la suite Y_n est croissante. Comme $E(Y_n^2) \leqslant 1$ pour chaque n, d'après le théorème de limite monotone 9.1(d)) il vient aussi $E(Y^2) \leqslant 1$. Donc $Y < \infty$ p.s., donc la série $X_{n_1} + \sum_{p=1}^{\infty}(X_{n_{p+1}} - X_{n_p})$ converge absolument pour presque tout ω. Comme c'est une série télescopique, on en déduit que $X_{n_p}(\omega)$ converge vers une limite $X(\omega)$ quand $p \to \infty$, et de plus $|X(\omega)| \leqslant |X_{n_1}(\omega)| + Y(\omega)$, tout ceci pour presque tout ω. Comme X_{n_1} et Y sont dans L^2, on a donc aussi $X \in L^2$.

Ensuite observons que, presque sûrement,
$$X - X_{n_p} = \lim_{m\to\infty} Z_m^p, \qquad \text{où } Z_m^p = \sum_{q=p}^{m}(X_{n_{q+1}} - X_{n_q}).$$
Comme $|Z_m^p| \leqslant Y$ pour tous p, m, le théorème de convergence dominée implique

$$\|X - X_{n_p}\| = \lim_{m\to\infty} \|Z^p_m\| \leqslant \lim_m \sum_{q=p}^{m} \|X_{n_{q+1}} - X_{n_q}\| \leqslant \frac{1}{2^{p-1}},$$

et par suite $\lim_{p\to\infty} \|X - X_{n_p}\| = 0$. Donc X_{n_p} converge vers X dans L^2.

Finalement

$$\|X_n - X\| \leqslant \|X_n - X_{n_p}\| + \|X_{n_p} - X\|.$$

Donc en faisant tendre n et p vers l'infini on en déduit que $X_n \to X$ dans L^2. ■

Définition 22.1. *Un* espace de Hilbert $\mathcal{H}$ *est un espace vectoriel normé complet, dont la norme provient d'un produit scalaire :* $\|x\| = \langle x, x\rangle^{\frac{1}{2}}$.

On a donc établi le

Théorème 22.3. L^2 *est un espace de Hilbert.*

Ci-dessous nous décrivons une série de propriétés des espaces de Hilbert ; : chacune s'applique évidemment à l'espace L^2. Ci-dessous, $\mathcal{H}$ *désigne un espace de Hilbert* de norme $\|\cdot\|$ et de produit scalaire $\langle\cdot,\cdot\rangle$, tandis que α et β désigne toujours des nombres réels. La convergence des vecteurs de $\mathcal{H}$ est toujours au sens de la norme.

Définition 22.2. *Deux vecteurs x et y de $\mathcal{H}$ sont* orthogonaux *si* $\langle x, y\rangle = 0$. *Un vecteur x est* orthogonal à une partie Γ *de $\mathcal{H}$ si* $\langle x, y\rangle = 0$ *pour tout* $y \in \Gamma$.

Si $\langle x, y\rangle = 0$ on a $\|x + y\|^2 = \|x\|^2 + \|y\|^2$: c'est la version « Hilbert » du théorème de Pythagore.

Théorème 22.4. (Continuité du produit scalaire.) *Si $x_n \to x$ et $y_n \to y$ dans $\mathcal{H}$, alors $\langle x_n, y_n\rangle \to \langle x, y\rangle$ dans* $\mathbb{R}$ (et donc aussi $\|x_n\| \to \|x\|$).

Preuve. L'inégalité de Cauchy-Schwarz, qui est valide dans tout espace de Hilbert (même démonstration que pour l'espace L^2) entraîne $\langle x, y\rangle \leqslant \|x\|\,\|y\|$, donc

$$\begin{aligned}|\langle x, y\rangle - \langle x_n, y_n\rangle| &= |\langle x - x_n, y_n\rangle + \langle x - x_n, y - y_n\rangle + \langle x_n, y - y_n\rangle|\\ &\leqslant \|x - x_n\|\,\|y_n\| + \|x - x_n\|\,\|y - y_n\| + \|x_n\|\,\|y - y_n\|.\end{aligned}$$

On a $\sup_n \|y_n\| < \infty$ et $\sup_n \|x_n\| < \infty$ puisque x_n et y_n convergent (par exemple, $\|x_n\| \leqslant \|x_n - x\| + \|x\|$ et $\|x\| < \infty$ et $\|x_n - x\| \to 0$). Donc le second membre de l'inégalité ci-dessus tend vers 0 quand $n \to \infty$. ■

Définition 22.3. *Une partie $\mathcal{L}$ de $\mathcal{H}$ est appelée un* sous-espace, *ou « sous-espace vectoriel », si $x, y \in \mathcal{L}$ implique $\alpha x + \beta y \in \mathcal{L}$. On dit en outre que le sous-espace $\mathcal{L}$ est* fermé *si pour toute suite (x_n) de vecteurs de $\mathcal{L}$ convergeant vers une limite x au sens de $\mathcal{H}$, on a $x \in \mathcal{L}$.*

Théorème 22.5. *Soit Γ une partie de $\mathcal{H}$. L'ensemble $\Gamma^\perp$ de tous les vecteurs orthogonaux à Γ est un sous-espace fermé de $\mathcal{H}$.*

Preuve. D'abord, $\Gamma^\perp$ est un sous-espace, même si Γ n'en n'est pas un : en effet si $x, y \in \Gamma^\perp$, alors $\langle x, z\rangle = 0$ et $\langle y, z\rangle = 0$ pour tout $z \in \Gamma$; donc

$$\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle = 0,$$

et $\alpha x + \beta y \in \Gamma^\perp$. Le fait que $\Gamma^\perp$ soit fermé découle du théorème 22.4. ■

Définition 22.4. *Si $\Gamma \subset \mathcal{H}$ on note $d(x, \Gamma) = \inf\{\|x - y\|; y \in \Gamma\}$ la distance de x à Γ.*

Si $x \in \Gamma$, on a bien sûr $d(x, \Gamma) = 0$. Inversement si $d(x, \Gamma) = 0$ et si Γ est fermé, alors $x \in \Gamma$.

Théorème 22.6. *Soit $\mathcal{L}$ un sous-espace fermé de $\mathcal{H}$ et $x \in \mathcal{H}$. Il existe un unique vecteur $y \in \mathcal{L}$ tel que $\|x - y\| = d(x, \mathcal{L})$.*

Preuve. Si $x \in \mathcal{L}$ on a évidemment $y = x$. Supposons donc que $x \notin \mathcal{L}$. On peut trouver une suite $y_n \in \mathcal{L}$ telle que $\lim_{n\to\infty} \|x - y_n\| = d(x, \mathcal{L})$. On va montrer que cette suite est de Cauchy. Observons d'abord que

$$\|y_n - y_m\|^2 = \|x - y_m\|^2 + \|x - y_n\|^2 - 2\langle x - y_m, x - y_n\rangle. \qquad (22.3)$$

Si on utilise l'inégalité

$$\left\|x - \frac{y_m + y_n}{2}\right\| \leqslant \frac{\|x - y_m\|}{2} + \frac{\|x - y_n\|}{2}$$

on obtient que

$$\lim_{n,m\to\infty} \left\|x - \frac{y_m + y_n}{2}\right\| \leqslant d(x, \mathcal{L}).$$

Donc

$$\lim_{n,m\to\infty} \left\|x - \frac{y_m + y_n}{2}\right\| = d(x, \mathcal{L}),$$

puisque $\frac{y_m+y_n}{2} \in \mathcal{L}$ (car $\mathcal{L}$ est un sous-espace). On a donc

$$\begin{aligned} d(x,\mathcal{L})^2 &= \lim_{m,n\to\infty} \left\| x - \frac{y_m+y_n}{2} \right\|^2 \\ &= \lim_{m,n\to\infty} \frac{1}{4} \left\{ \|x-y_m\|^2 + \|x-y_n\|^2 + 2\langle x-y_m, x-y_n\rangle \right\}, \end{aligned}$$

et donc

$$\lim_{n,m\to\infty} \langle x-y_m, x-y_n\rangle = d(x,\mathcal{L})^2. \tag{22.4}$$

Il reste à combiner (22.3) et (22.4) pour obtenir que la suite $(y_n)_{n\geqslant 1}$ est de Cauchy. Cette suite converge donc vers une limite y, qui appartient à $\mathcal{L}$ puisque ce sous-espace est fermé. De plus $d(x,\mathcal{L}) = \|x-y\|$ à cause de la continuité (évidente) de la fonction distance.

Il reste à montrer l'unicité de y. Supposons que z soit un autre vecteur vérifiant les mêmes propriétés. La suite

$$w_{2n} = y, \qquad w_{2n+1} = z,$$

est encore une suite de Cauchy par le même argument que ci-dessus, donc elle converge vers une limite *unique*, donc $z = y$. ■

Nous introduisons maintenant un concept important, celui de *projection (orthogonale)*. L'espace de Hilbert $\mathcal{H}$ est toujours fixé, ainsi que le sous-espace fermé $\mathcal{L}$. La *projection d'un vecteur x sur $\mathcal{L}$* est le point y obtenu dans le théorème 22.6, c'est-à-dire le point de $\mathcal{L}$ qui est « le plus proche de x ». On note Π l'opérateur de projection, i.e. $\Pi x = y$ ci-dessus. Noter que Π dépend de manière essentielle du sous-espace $\mathcal{L}$, et le résultat ci-dessous explique son nom de *projecteur orthogonal sur $\mathcal{L}$*.

Théorème 22.7. *L'opérateur de projection (ou projection orthogonale) Π de $\mathcal{H}$ sur un sous-espace fermé $\mathcal{L}$ vérifie les trois propriétés suivantes :*

(i) Π *est idempotent : i.e.,* $\Pi^2 = \Pi$ *;*

(ii) $\Pi x = x$ *pour* $x \in \mathcal{L}$ *;* $\Pi x = 0$ *pour* $x \in \mathcal{L}^\perp$ *;*

(iii) *Pour tout $x \in \mathcal{H}$, le vecteur $x - \Pi x$ est orthogonal à $\mathcal{L}$.*

Preuve. (i) suit immédiatement du fait que $\Pi x = x$ si $x \in \mathcal{L}$.
(ii) La propriété $\Pi x = x$ si $x \in \mathcal{L}$ est évidente. Si $x \in \mathcal{L}^\perp$ on a $\|x-y\|^2 = \langle x-y, x-y\rangle = \|x\|^2 + \|y\|^2$ pour $y \in \mathcal{L}$, donc $y = 0$ minimise $y \mapsto \|x-y\|$ sur $\mathcal{L}$, et donc $\Pi x = 0$.
(iii) On remarque d'abord que si $y \in \mathcal{L}$:

$$\begin{aligned} \|x-\Pi x\|^2 &\leqslant \|x-(\Pi x + y)\|^2 \\ &= \|x-\Pi x\|^2 + \|y\|^2 - 2\langle x-\Pi x, y\rangle, \end{aligned}$$

et donc

$$2\langle x - \Pi x, y\rangle \leqslant \|y\|^2.$$

Comme $y \in \mathcal{L}$ est arbitraire et comme $\mathcal{L}$ est un sous-espace, on peut remplacer y par αy pour tout $\alpha > 0$, de façon à obtenir

$$2\langle x - \Pi x, \alpha y\rangle \leqslant \|\alpha y\|^2,$$

et en divisant par α on obtient

$$2\langle x - \Pi x, y\rangle \leqslant \alpha\|y\|^2;$$

On fait tendre α vers 0 et on conclut que $\langle x - \Pi x, y\rangle \leqslant 0$. On montre exactement de la même manière que $\langle x - \Pi x, y\rangle \geqslant 0$, en considérant des $\alpha < 0$. Donc $x - \Pi x$ est orthogonal à $\mathcal{L}$. ■

Corollaire 22.1. *Soit Π l'opérateur de projection sur un sous-espace fermé $\mathcal{L}$ de $\mathcal{H}$. Pour tout $x \in \mathcal{H}$, $x = \Pi x + (x - \Pi x)$ est l'unique représentation de x comme la somme d'un vecteur de $\mathcal{L}$ et d'un vecteur de $\mathcal{L}^\perp$. De plus $x - \Pi x$ est la projection de x sur $\mathcal{L}^\perp$. Enfin on a $(\mathcal{L}^\perp)^\perp = \mathcal{L}$.*

Preuve. On a vu que $\Pi x \in \mathcal{L}$ et que $x - \Pi x \in \mathcal{L}^\perp$ dans le théorème 22.7(iii). Quant à l'unicité, soit $x = y + z$ une autre représentation avec $y \in \mathcal{L}$ et $z \in \mathcal{L}^\perp$. Le vecteur $y - \Pi x = z - (x - \Pi x)$ est à la fois dans $\mathcal{L}$ et dans $\mathcal{L}^\perp$; donc, étant orthogonal à lui-même, il est nul : cela montre l'unicité.

Ensuite on remarque que $\mathcal{L} \subset (\mathcal{L}^\perp)^\perp$: en effet si $x \in \mathcal{L}$ et $y \in \mathcal{L}^\perp$ on a $\langle x, y\rangle = 0$, donc $x \in (\mathcal{L}^\perp)^\perp$.

Enfin, si $x \in (\mathcal{L}^\perp)^\perp$ on a $x = y + z$ avec $y \in \mathcal{L}$ et $z \in \mathcal{L}^\perp$. On a $\langle x, z\rangle = \langle y, z\rangle + \langle z, z\rangle$, et comme $\langle y, z\rangle = 0$ (puisque $y \in \mathcal{L}$) et $\langle x, z\rangle = 0$ (puisque $z \in \mathcal{L}^\perp$ et $x \in (\mathcal{L}^\perp)^\perp$) il s'ensuit que $\langle z, z\rangle = 0$: donc $z = 0$, et $x = y$. Comme $y \in \mathcal{L}$ on a aussi $x \in \mathcal{L}$, donc $(\mathcal{L}^\perp)^\perp \subset \mathcal{L}$. ■

Corollaire 22.2. *Soit Π l'opérateur de projection sur un sous-espace fermé $\mathcal{L}$ de $\mathcal{H}$. On a alors :*

(i) $\langle \Pi x, y\rangle = \langle x, \Pi y\rangle$,

(ii) Π *est un opérateur linéaire :* $\Pi(\alpha x + \beta y) = \alpha \Pi x + \beta \Pi y$.

Preuve. (i) Soit $x, y \in \mathcal{H}$. En vertu du corollaire 22.1 on peut écrire les vecteurs x et y de manière unique ainsi :

$$\begin{aligned} x &= x_1 + x_2, \qquad && x_1 \in \mathcal{L}, \quad x_2 \in \mathcal{L}^\perp, \\ y &= y_1 + y_2, \qquad && y_1 \in \mathcal{L}, \quad y_2 \in \mathcal{L}^\perp. \end{aligned} \tag{22.5}$$

Par suite
$$\langle \Pi x, y\rangle = \langle x_1, y\rangle = \langle x_1, y_1 + y_2\rangle = \langle x_1, y_1\rangle,$$
puisque $\langle x_1, y_2\rangle = 0$. Comme on a aussi $\langle x_2, y_1\rangle = 0$, il vient
$$= \langle x_1 + x_2, y_1\rangle = \langle x, y_1\rangle = \langle x, \Pi y\rangle.$$

(ii) Avec les notations (22.5) on obtient
$$\alpha x + \beta y = (\alpha x_1 + \beta y_1) + (\alpha x_2 + \beta y_2),$$
donc
$$\Pi(\alpha x + \beta y) = \alpha x_1 + \beta y_1 = \alpha \Pi x + \beta \Pi y.$$
■

Nous terminons cette exposition par une réciproque qui dit essentiellement que si un opérateur linéaire « se comporte comme une projection », au sens des propriétés précédentes, c'est en fait une projection.

Théorème 22.8. *Soit* T *une application de $\mathcal{H}$ dans un sous-espace fermé $\mathcal{L}$. Si pour tout x le vecteur $x - \mathrm{T}x$ est orthogonal à $\mathcal{L}$, alors* T *est la projection (orthogonale) sur le sous-espace $\mathcal{L}$.*

Preuve. On peut écrire $x = \mathrm{T}x + (x - \mathrm{T}x)$, avec $\mathrm{T}x \in \mathcal{L}$ et $(x - \mathrm{T}x) \in \mathcal{L}^{\perp}$. D'après le corollaire 22.1, $\mathrm{T}x$ doit être la projection orthogonale de x sur $\mathcal{L}$. ■

Exercices

1. En utilisant la propriété $(a-b)^2 \geqslant 0$, montrer que $(a+b)^2 \leqslant 2a^2 + 2b^2$.
2. Soit $x, y \in \mathcal{H}$ (un espace de Hilbert), avec $\langle x, y\rangle = 0$. Montrer le « théorème de Pythagore » : $\|x+y\|^2 = \|x\|^2 + \|y\|^2$.
3. Montrer que $\mathbb{R}^n$ est un espace de Hilbert avec le produit scalaire « usuel » : si $\vec{x} = (x_1, \ldots, x_n)$ et $\vec{y} = (y, \ldots, y_n)$, alors $\langle \vec{x}, \vec{y}\rangle = \sum_{i=1}^n x_i y_i$.
4. Soit $\mathcal{L}$ un sous-espace fermé de l'espace de Hilbert $\mathcal{H}$, et Π la projection sur $\mathcal{L}$. Montrer que Πy est l'unique vecteur de $\mathcal{L}$ tel que $\langle \Pi y, z\rangle = \langle y, z\rangle$ pour tout $z \in \mathcal{L}$.

Chapitre 23

Espérance conditionnelle

Soit X et Y deux v.a., avec Y réelle et X prenant ses valeurs dans un ensemble *fini ou dénombrable.* Il arrive souvent qu'on connaisse la valeur prise par X et qu'on veuille calculer l'espérance de Y, en utilisant cette information supplémentaire, ce qui *a priori* modifie l'espérance. De manière plus formelle, cela revient à poser :

Définition 23.1. *Soit* X *une v.a. à valeurs dans un ensemble* E *fini ou dénombrable, dont les points sont notés* $\{x_1, x_2, \ldots, x_n, \ldots\}$. *Soit* Y *une autre v.a. réelle, définie sur le même espace de probabilité* $(\Omega, \mathcal{A}, \mathrm{P})$. *Si* $\mathrm{P}(\mathrm{X} = x_j) > 0$ l'espérance conditionnelle de Y sachant $\{\mathrm{X} = x_j\}$ *est le nombre*

$$\mathrm{E}(\mathrm{Y} \,|\, \mathrm{X} = x_j) = \mathrm{E}_{\mathrm{Q}_j}(\mathrm{Y}), \quad \text{espérance de Y pour } \mathrm{Q}_j,$$

où Q_j *est la probabilité sur* $(\Omega, \mathcal{A})$ *donnée par* $\mathrm{Q}(\Lambda) = \mathrm{P}(\Lambda \,|\, \mathrm{X} = x_j)$, *pourvu que* $\mathrm{E}_{\mathrm{Q}_j}(|\mathrm{Y}|) < \infty$ *(i.e.,* Y *est* Q_j*-intégrable).*

Lorsque Y est elle-même une v.a. discrète, prenant ses valeurs dans $\{y_1, y_2, \ldots\}$ (une partie dénombrable de $\mathbb{R}$), on peut utiliser la formule (5.1) pour Y, et on obtient immédiatement le

Théorème 23.1. *Dans la situation précédente, pour tout* j *tel que* $\mathrm{P}(\mathrm{X} = x_j) > 0$ *on a*

$$\mathrm{E}(\mathrm{Y} \,|\, \mathrm{X} = x_j) = \sum_{k=1}^{\infty} y_k \mathrm{P}(\mathrm{Y} = y_k \,|\, \mathrm{X} = x_j),$$

pourvu que la série ci-dessus soit absolument convergente.

Au lieu de « fixer » la valeur x_j prise par X, on peut modifier légèrement notre point de vue, et chercher l'espérance conditionnelle de Y, *connaissant la variable* X (toujours supposée à valeurs dans l'ensemble fini ou dénombrable E), plutôt que « sachant que $\mathrm{X} = x_j$ ». Pour cela, on pose

$$f(x_j) = \begin{cases} \mathrm{E}(\mathrm{Y} \,|\, \mathrm{X} = x_j) & \text{si } \mathrm{P}(\mathrm{X} = x_j) > 0 \\ \text{une valeur arbitraire} & \text{si } \mathrm{P}(\mathrm{X} = x_j) = 0, \end{cases} \tag{23.1}$$

ce qui définit une fonction f sur E.

Définition 23.2. *Soit* X *une v.a. à valeurs dans un espace* E *fini ou dénombrable, et* Y *une v.a. réelle définie sur le même espace de probabilité. L'espérance conditionnelle de* Y *sachant* X *est*

$$\mathrm{E}(\mathrm{Y} \mid \mathrm{X}) = f(\mathrm{X}),$$

où f est donnée par (23.1), et pourvu que cette fonction f soit bien définie (i.e., pour tout j tel que $\mathrm{P}(\mathrm{X} = x_j) > 0$*, la v.a.* Y *est intégrable par rapport à la probabilité* $\mathrm{Q}_j(\Lambda) = \mathrm{P}(\Lambda \mid \mathrm{X} = x_j)$*).*

Remarque 23.1. Cette définition ne définit pas réellement $\mathrm{E}(\mathrm{Y} \mid \mathrm{X})$ partout, mais seulement presque partout, car sa valeur est arbitraire sur les ensembles $\{\mathrm{X} = x\}$ tels que $\mathrm{P}(\mathrm{X} = x) = 0$: ce sera un trait caractéristique des espérances conditionnelles dans le cas général étudié plus bas.

Exemple. Soit X une v.a. de Poisson de paramètre λ. Soit une autre v.a. S définie ainsi : quand $\mathrm{X} = n$, on tire n fois (indépendamment) une pièce ayant la probabilité p de tomber sur Face ; S est alors le nombre total de « Faces ». On va calculer $\mathrm{E}(\mathrm{S} \mid \mathrm{X})$ et $\mathrm{E}(\mathrm{X} \mid \mathrm{S})$.

Calculons d'abord $\mathrm{E}(\mathrm{S} \mid \mathrm{X} = n)$. Si $\mathrm{X} = n$, la v.a. S est binomiale de paramètres n et p, donc $\mathrm{E}(\mathrm{S} \mid \mathrm{X} = n) = pn$ et $\mathrm{E}(\mathrm{S} \mid \mathrm{X}) = p\mathrm{X}$.

Pour calculer $\mathrm{E}(\mathrm{X} \mid \mathrm{S})$ il nous faut évaluer $\mathrm{E}(\mathrm{X} \mid \mathrm{S} = k)$, et nous commençons par chercher $\mathrm{P}(\mathrm{X} = n \mid \mathrm{S} = k)$:

$$\begin{aligned}
\mathrm{P}(\mathrm{X} = n \mid \mathrm{S} = k) &= \frac{\mathrm{P}(\mathrm{S} = k \mid \mathrm{X} = n)\mathrm{P}(\mathrm{X} = n)}{\mathrm{P}(\mathrm{S} = k)} \\
&= \frac{\mathrm{C}_n^k p^k (1-p)^{n-k} \left(\frac{\lambda^n}{n!}\right) e^{-\lambda}}{\sum_{m \geqslant k} \mathrm{C}_m^k p^k (1-p)^{m-k} \left(\frac{\lambda^m}{m!}\right) e^{-\lambda}} \\
&= \frac{((1-p)\lambda)^{n-k}}{(n-k)!} e^{-(1-p)\lambda}
\end{aligned}$$

pour $n \geqslant k$, et bien sûr $\mathrm{P}(\mathrm{X} = n \mid \mathrm{S} = k) = 0$ si $n < k$. Donc

$$\mathrm{E}(\mathrm{X} \mid \mathrm{S} = k) = \sum_{n \geqslant k} n \frac{((1-p)\lambda)^{n-k}}{(n-k)!} e^{-(1-p)\lambda} = k + (1-p)\lambda.$$

et il vient

$$\mathrm{E}(\mathrm{X} \mid \mathrm{S}) = \mathrm{S} + (1-p)\lambda.$$

On peut vérifier directement que $\mathrm{E}(\mathrm{S}) = \mathrm{E}(\mathrm{E}(\mathrm{S} \mid \mathrm{X}))$, propriété qui découlera aussi du théorème 23.3 ci-après. Donc on a

$$\mathrm{E}(\mathrm{S}) = p\mathrm{E}(\mathrm{X}) = p\lambda.$$

Nous allons maintenant passer au *cas général* : nous voulons calculer $E(Y \mid X)$ quand X est une v.a. à valeurs dans un espace quelconque. L'approche précédente ne marche plus, car typiquement on a $P(X = x) = 0$ pour tout x. Cependant on a vu aussi dans le cas dénombrable que $E(Y \mid X) = f(X)$ pour une certaine fonction f, et c'est cette idée qui va se généraliser, grâce au théorème suivant.

Rappelons d'abord qu'on a introduit au chapitre 10 la notion de tribu engendrée par une v.a. : si par exemple X est une application de Ω dans $\mathbb{R}^n$, la *tribu engendrée par* X est

$$\sigma(X) = X^{-1}(\mathcal{B}^n) = \left\{A \subset \Omega : X^{-1}(B) = A, \text{ pour un } B \in \mathcal{B}^n\right\}.$$

Théorème 23.2. *Soit X une v.a. à valeurs dans $\mathbb{R}^n$ et Y une v.a. réelle, toutes deux définies sur le même espace $(\Omega, \mathcal{A})$. Alors Y est mesurable par rapport à $\sigma(X)$ si et seulement s'il existe une fonction borélienne f de $\mathbb{R}^n$ dans $\mathbb{R}$ telle que $Y = f(X)$.*

Preuve. Si la fonction f de l'énoncé existe on a $Y^{-1}(B) = X^{-1}(f^{-1}(B))$. Si $B \in \mathcal{B}$ alors $A = f^{-1}(B) \in \mathcal{B}^n$, donc $X^{-1}(A) \in \sigma(X)$ (on peut aussi utiliser le théorème 8.2).

Supposons inversement que $Y^{-1}(B) \in \sigma(X)$ pour tout $B \in \mathcal{B}$. Supposons d'abord que Y soit « simple », i.e. $Y = \sum_{i=1}^k a_i 1_{A_i}$ pour un $k < \infty$, des a_i réels tous différents, et des A_i deux à deux disjoints. On a $A_i = Y^{-1}(\{a_i\})$, donc $A_i \in \sigma(X)$ et il existe $B_i \in \mathcal{B}^n$ avec $A_i = X^{-1}(B_i)$. Il reste à poser $f(x) = \sum_{i=1}^k a_i 1_{B_i}(x)$ pour obtenir $Y = f(X)$ avec f borélienne. Ensuite, si on suppose Y positive, on sait (cf. chapitre 9) qu'il existe une suite croissante (Y_n) de v.a. simples positives croissant vers Y et $\sigma(X)$-mesurables. On vient de voir que $Y_n = f_n(X)$ pour des fonctions boréliennes f_n. On pose $f(x) = \limsup_{n\to\infty} f_n(x)$, qui est borélienne également, et on a

$$Y = \lim_{n\to\infty} Y_n = \lim_n f_n(X).$$

Mais

$$(\limsup_{n\to\infty} f_n)(X) = \limsup_n (f_n(X)),$$

et par suite $Y = f(X)$.

Pour Y de signe quelconque on écrit $Y = Y^+ - Y^-$, et on se ramène ainsi au cas précédent. ■

Dans ce qui suit l'espace de probabilité $(\Omega, \mathcal{A}, P)$ est fixé, et X est une v.a. sur cet espace, à valeurs dans $\mathbb{R}^n$. L'espace $L^2(\Omega, \mathcal{A}, P)$ est l'espace des (classes d'équivalence pour la relation d'équivalence « égalité p.s. » de) v.a. Y vérifiant $E(Y^2) < \infty$. Comme on l'a vu au chapitre 22 c'est un espace de Hilbert avec le produit scalaire

$$\langle Y, Z\rangle = E(YZ).$$

Comme $\sigma(X) \subset \mathcal{A}$, l'espace $L^2(\Omega, \sigma(X), P)$ est aussi un espace de Hilbert, et c'est un sous-espace fermé de $L^2(\Omega, \mathcal{A}, P)$. (Noter que le produit scalaire est le même sur ces deux espaces de Hilbert.)

Définition 23.3. *Soit* $Y \in L^2(\Omega, \mathcal{A}, P)$. L'espérance conditionnelle de Y sachant X *est l'unique élément* $\widehat{Y}$ *de* $L^2(\Omega, \sigma(X), P)$ *qui vérifie*

$$E(\widehat{Y}Z) = E(YZ) \qquad \textit{pour tout } \ Z \in L^2(\Omega, \sigma(X), P). \tag{23.2}$$

On écrit cette espérance conditionnelle (i.e., la variable $\widehat{Y}$*) ainsi :*

$$E(Y \mid X)$$

L'équation (23.2) signifie simplement que $\widehat{Y}$ est la projection orthogonale de Y sur le sous-espace fermé $L^2(\Omega, \sigma(X), P)$ de $L^2(\Omega, \mathcal{A}, P)$: l'existence et l'unicité de l'espérance conditionnelle suit donc du corollaire 22.1.

Comme $E(Y \mid X)$ est $\sigma(X)$-mesurable, d'après le théorème 23.2 il existe une fonction borélienne f telle que $E(Y \mid X) = f(X)$. Donc (23.2) équivaut au fait que

$$E(f(X)g(X)) = E(Yg(X)) \tag{23.3}$$

pour toute fonction borélienne g telle que $g(X) \in \mathcal{L}^2$.

Dans la définition précédente la v.a. X ne joue pas de rôle réel, c'est la tribu $\sigma(X)$ qui est importante. Il est donc plutôt naturel de la remplacer par une sous-tribu arbitraire $\mathcal{G}$ de $\mathcal{A}$. Là encore $L^2(\Omega, \mathcal{G}, P)$ est un sous-espace de Hilbert de $L^2(\Omega, \mathcal{A}, P)$, et on peut poser :

Définition 23.4. *Soit* $Y \in L^2(\Omega, \mathcal{A}, P)$ *et* $\mathcal{G}$ *une sous-tribu de* $\mathcal{A}$. L'espérance conditionnelle de Y sachant $\mathcal{G}$ *est l'unique élément* $E(Y \mid \mathcal{G})$ *de* $L^2(\Omega, \mathcal{G}, P)$ *qui vérifie*

$$E(YZ) = E(E(Y \mid \mathcal{G})Z) \tag{23.4}$$

pour tout $Z \in L^2(\Omega, \mathcal{G}, P)$.

Avertissement important : L'espérance conditionnelle est un élément de L^2, c'est-à-dire une classe d'équivalence de v.a. : on peut supposer que c'est une vraie variable aléatoire, mais dans ce cas elle n'est définie qu'à un ensemble de mesure nulle près. Par suite les propriétés comme $E(Y \mid \mathcal{G}) \geqslant 0$ ou $E(Y \mid \mathcal{G}) = Z$, etc., doivent *toujours* être comprises au sens « presque sûr », bien qu'on omette en général de le mentionner. On dit aussi parfois : il existe une « version » de $E(Y \mid \mathcal{G})$ qui est positive, ou égale à Z, etc.

Théorème 23.3. *Soit* $Y \in L^2(\Omega, \mathcal{A}, P)$ *et* $\mathcal{G}$ *une sous-tribu de* $\mathcal{A}$.

(a) *Si* $Y \geqslant 0$ *alors* $E(Y \mid \mathcal{G}) \geqslant 0$.

(b) *Si* $\mathcal{G} = \sigma(X)$ *pour une v.a.* X *à valeurs dans* $\mathbb{R}^n$, *il existe une fonction borélienne* f *sur* $\mathbb{R}$ *telle que* $E(Y \mid \mathcal{G}) = f(X)$; .

(c) $E(E(Y \mid \mathcal{G})) = E(Y)$.

(d) *L'application* $Y \to E(Y \mid \mathcal{G})$ *est linéaire.*

Preuve. On a déjà vu (b). Pour (c) il suffit d'appliquer (23.4) avec $Z = 1$, et (d) découle aussi de (23.4) : en effet si U et V sont dans L^2, on a

$$\begin{aligned} E((U + \alpha V)Z) &= E(UZ) + \alpha E(VZ) \\ &= E(E(U \mid \mathcal{G})Z) + \alpha E(E(V \mid \mathcal{G})Z) \\ &= E((E(U \mid \mathcal{G}) + \alpha E(V \mid \mathcal{G}))Z). \end{aligned}$$

et donc $E(U + \alpha V \mid \mathcal{G}) = E(U \mid \mathcal{G}) + \alpha E(V \mid \mathcal{G})$ à cause de l'unicité de l'espérance conditionnelle (on peut aussi dire que $E(Y \mid \mathcal{G})$ est la projection de Y sur le sous-espace $L^2(\Omega, \mathcal{G}, P)$, et l'opérateur projection est linéaire, cf. le corollaire 22.2).

Finalement pour (a) on utilise encore (23.4) en prenant pour Z la fonction indicatrice $1_{\{E(Y \mid \mathcal{G})<0\}}$; si $Y \geqslant 0$ p.s. on a $E(YZ) \geqslant 0$ puisque Y et Z sont toutes deux positives, tandis que

$$E(E(Y \mid \mathcal{G})Z) = E(E(Y \mid \mathcal{G})1_{\{E(Y \mid \mathcal{G})<0\}}) < 0 \quad \text{si} \quad P(\{E(Y \mid \mathcal{G}) < 0\}) > 0.$$

Cela contredit (23.3), sauf si $P(\{E(Y \mid \mathcal{G}) < 0\}) = 0$. ■

Remarque 23.2. Comme on l'a vu dans la démonstration du théorème 23.3, la propriété cruciale de l'espérance conditionnelle est (23.4) ; prouver l'existence et l'unicité de l'espérance conditionnelle est la seule raison pour laquelle nous avons consacré un chapitre entier aux espaces de Hilbert.

Il nous reste à étendre la notion d'espérance conditionnelle aux v.a. qui ne sont pas nécessairement de carré intégrable. Les techniques d'espace de Hilbert font alors bien sûr défaut.

Rappelons une fois encore que $L^1(\Omega, \mathcal{A}, P)$ est l'espace des (classes d'équivalences de) v.a. intégrables, et de manière analogue nous allons noter $L^+(\Omega, \mathcal{A}, P)$ l'espace des (classes d'équivalences de) v.a. *positives*, toujours en identifiant deux v.a. qui sont p.s. égales. Les v.a. de L^+ peuvent prendre la valeur $+\infty$.

Lemme 23.1. *Soit $Y \in L^+(\Omega, \mathcal{A}, P)$ et soit $\mathcal{G}$ une sous-tribu de $\mathcal{A}$. Il existe un unique élément $E(Y \mid \mathcal{G})$ de $L^+(\Omega, \mathcal{G}, P)$, appelé l'espérance conditionnelle de Y sachant $\mathcal{G}$, tel que*

$$E(YX) = E(E(Y \mid \mathcal{G})X) \tag{23.5}$$

pour tout $X \in L^+(\Omega, \mathcal{G}, P)$, et cette espérance conditionnelle coïncide avec celle de la définition 23.4 si on a aussi $Y \in L^2(\Omega, \mathcal{A}, P)$.

Si de plus $0 \leqslant Y \leqslant Y'$, alors

$$E(Y \mid \mathcal{G}) \leqslant E(Y' \mid \mathcal{G}). \tag{23.6}$$

Preuve. Si Y est à la fois dans $L^2(\Omega, \mathcal{A}, P)$ et positive, on définit $E\{Y|\mathcal{G}\}$ par la définition 23.4. Si alors $X \in L^+(\Omega, \mathcal{G}, P)$, et comme les v.a. $X_n = X \wedge n$ sont de carré intégrable, il vient en appliquant deux fois le théorème de convergence monotone et (23.5) :

$$\begin{aligned} E(YX) &= \lim_n E(YX_n) \\ &= \lim_n E(E(Y \mid \mathcal{G})X_n) \\ &= E(E(Y \mid \mathcal{G})X) \end{aligned} \tag{23.7}$$

et (23.5) est satisfaite.

Soit maintenant $Y \in L^+(\Omega, \mathcal{A}, P)$. Chaque $Y_m = Y \wedge m$ est de carré intégrable, tandis que d'après le théorème 23.3 l'espérance conditionnelle est un opérateur positif, de sorte que la suite $E(Y_m \mid \mathcal{G})$ est croissante. Par suite on peut poser

$$E(Y \mid \mathcal{G}) = \lim_{m \to \infty} E(Y_m \mid \mathcal{G}). \tag{23.8}$$

Si $X \in L^+(\Omega, \mathcal{G}, P)$, on obtient alors en utilisant plusieurs fois le théorème de convergence monotone, ainsi que (23.8) :

$$\begin{aligned} E(YX) &= \lim_m E(Y_m X) \\ &= E\left(\lim_m E(Y_m \mid \mathcal{G})X\right) \\ &= E(E(Y \mid \mathcal{G})X), \end{aligned}$$

Si de plus on a $Y \leqslant Y'$, alors $Y \wedge m \leqslant Y' \wedge m$ pour chaque m, donc $E(Y \wedge m \mid \mathcal{G}) \leqslant E(Y' \wedge m \mid \mathcal{G})$ par le théorème 23.3(a), et on obtient (23.6).

Il reste à montrer l'unicité. Soit U et V deux versions de $E(Y \mid \mathcal{G})$, qui donc vérifient (23.8) pour toute v.a. positive $\mathcal{G}$-mesurable X. Soit $\Lambda_n = \{U < V \leqslant n\}$ et supposons que $P(\Lambda_n) > 0$. Comme on a $\Lambda_n \in \mathcal{G}$ il vient

$$E(Y1_{\Lambda_n}) = E(U1_{\Lambda_n}) = E(V1_{\Lambda_n}).$$

tandis que $0 \leqslant U1_{\Lambda_n} \leqslant V1_{\Lambda_n} \leqslant n$, et $P(\Lambda_n) > 0$ implique que les v.a. $V1_{\Lambda_n}$ et $U1_{\Lambda_n}$ ne sont pas p.s. égales : on en déduit que $E(U1_\Lambda) < E(V1_\Lambda)$, d'où une contradiction. Il faut donc que $P(\Lambda_n) = 0$ pour tout n, et comme $\{U > V\} = \cup_{n \geqslant 1} \Lambda_n$ il vient $P(U < V) = 0$; on vérifie de même que $P(V > U) = 0$, et l'unicité en découle. ■

Théorème 23.4. *Soit* $Y \in L^1(\Omega, \mathcal{A}, P)$ *et soit* $\mathcal{G}$ *une sous-tribu de* $\mathcal{A}$. *Il existe un unique élément* $E(Y \mid \mathcal{G})$ *de* $L^1(\Omega, \mathcal{G}, P)$, *appelé l'espérance conditionnelle de* Y *sachant* $\mathcal{G}$, *tel que*

$$E(YX) = E(E(Y \mid \mathcal{G})X) \tag{23.9}$$

pour tout v.a. X *borné* $\mathcal{G}$*-mesurable, et cette espérance conditionnelle coïncide avec celle de la définition 23.4 (resp. du lemme 23.1) si on a aussi* $Y \in L^2(\Omega, \mathcal{A}, P)$ *(resp.* $Y \in L^+(\Omega, \mathcal{A}, P)$*). Enfin, on a aussi*

(a) *Si* $Y \geqslant 0$ *alors* $E(Y \mid \mathcal{G}) \geqslant 0$.

(b) *L'application* $Y \mapsto E(Y \mid \mathcal{G})$ *est linéaire.*

Preuve. Si $Y \in L^1$, on a aussi $Y^+ \in L^1$ et $Y^- \in L^1$, où Y^+ et Y^- sont les parties positive et négative de Y. On a $Y = Y^+ - Y^-$, et on pose

$$E(Y \mid \mathcal{G}) = E(Y^+ \mid \mathcal{G}) - E(Y^- \mid \mathcal{G}).$$

Cette formule a bien en sens : en effet les v.a. Y^+ et Y^-, donc aussi les v.a. $E(Y^+ \mid \mathcal{G})$ et $E(Y^- \mid \mathcal{G})$ d'après le théorème 23.3(c), sont intégrables et donc p.s. finies. Il est clair que $E(Y \mid \mathcal{G})$ vérifie (23.9). Pour l'unicité, considérons deux versions U et V de $E(Y \mid \mathcal{G})$, et soit $\Lambda = \{U < V\}$. On a $\Lambda \in \mathcal{G}$, donc 1_Λ est bornée et $\mathcal{G}$-mesurable, donc $E(Y1_\Lambda) = E(E(Y \mid \mathcal{G})1_\Lambda) = E(U1_\Lambda) = E(V1_\Lambda)$. Mais si $P(\Lambda) > 0$ il vient $E(U1_\Lambda) < E(V1_\Lambda)$, ce qui fournit une contradiction. Par suite $P(U < V) = 0$ et on vérifie de même que $P(V < U) = 0$.

Les dernières assertions sont des conséquences triviales de la définition ci-dessus de $E(Y \mid \mathcal{G})$, du lemme 23.1 et du théorème 23.3. ■

Exemple. Soit (X, Z) une v.a. bidimensionnelle admettant une densité f. Soit g une fonction borélienne bornée et $Y = g(Z)$. On cherche à calculer $E(Y \mid X) = E(g(Z) \mid X)$. On rappelle que X admet la densité f_X suivante :

$$f_X(x) = \int f(x, z) dz,$$

et qu'on a défini au chapitre 12 (cf. théorème 12.2) la *densité conditionnelle de* Z *sachant* $X = x$ par :

$$f_{X=x}(z) = \frac{f(x, z)}{f_X(x)},$$

partout où $f_X(x) \neq 0$. Posons alors

$$h(x) = \int g(z) f_{X=x}(z) dz.$$

Pour toute fonction borélienne bornée k il vient

$$\begin{aligned} E(h(X)k(X)) &= \int h(x)k(x)f_X(x)dx \\ &= \iint g(z)f_{X=x}(z)dz\, k(x)f_X(x)dx \\ &= \iint g(z)\frac{f(x,z)}{f_X(x)}k(x)f_X(x)dz\, dx \\ &= \iint g(z)k(x)f(x,z)dz\, dx \\ &= E(g(Z)k(X)) = E(Yk(X)). \end{aligned}$$

Donc (23.9) implique que

$$E(Y \mid X) = h(X),$$

et on a ainsi décrit une méthode permettant de calculer les espérances conditionnelles lorsque les v.a. admettent des densités.

Théorème 23.5. *Soit* Y *une v.a. positive ou intégrable sur* $(\Omega, \mathcal{F}, P)$, *et* $\mathcal{G}$ *une sous-tribu. Alors* $E(Y \mid \mathcal{G}) = Y$ *si et seulement si* Y *est* $\mathcal{G}$*-mesurable.*

Preuve. Trivial à partir de la définition de l'espérance conditionnelle. ∎

Théorème 23.6. *Si* $\mathrm{Y} \in \mathrm{L}^1(\Omega, \mathcal{A}, \mathrm{P})$ *et si* X *est une v.a. indépendante de* Y*, on a*

$$\mathrm{E}(\mathrm{Y} \,|\, \mathrm{X}) = \mathrm{E}\{\mathrm{Y}\}.$$

Preuve. Soit g borélienne bornée. On a $\mathrm{E}(\mathrm{Y}g(\mathrm{X})) = \mathrm{E}(\mathrm{Y})\mathrm{E}(g(\mathrm{X}))$ par l'indépendance. Le résultat découle alors immédiatement de la définition de l'espérance conditionnelle et du théorème 23.2. ■

Théorème 23.7. *Soit* X, Y *des v.a. réelles sur* $(\Omega, \mathcal{A}, \mathrm{P})$, *et* $\mathcal{G}$ *une sous-tribu de* $\mathcal{A}$ *par rapport à laquelle* X *est mesurable. On a alors*

$$\mathrm{E}(\mathrm{XY} \,|\, \mathcal{G}) = \mathrm{X}\mathrm{E}(\mathrm{Y} \,|\, \mathcal{G}).$$

dans chacun des deux cas suivants :

(a) *les v.a.* X, Y *et* XY *sont intégrables,*

(b) *les v.a.* X *et* Y *sont positives.*

Preuve. Supposons d'abord (b). Pour toute v.a. $\mathcal{G}$-mesurable et positive Z on a

$$\mathrm{E}(\mathrm{XYZ}) = \mathrm{E}(\mathrm{XZ}\ \mathrm{E}(\mathrm{Y} \,|\, \mathcal{G}))$$

par (23.5). Comme $\mathrm{XE}(\mathrm{Y} \,|\, \mathcal{G})$ est aussi $\mathcal{G}$-mesurable, on en déduit le résultat par une nouvelle application de la caractérisation (23.5).

Dans le cas (a) on observe que les v.a. $\mathrm{X}^+\mathrm{Y}^+$, $\mathrm{X}^-\mathrm{Y}^+$, $\mathrm{X}^+\mathrm{Y}^-$ et $\mathrm{X}^-\mathrm{Y}^-$ sont toutes intégrables et positives, donc $\mathrm{E}(\mathrm{X}^+\mathrm{Y}^+ \,|\, \mathcal{G}) = \mathrm{X}^+\ \mathrm{E}(\mathrm{Y}^+ \,|\, \mathcal{G})$ d'après ce qui précède, et de même pour les trois autres produits, et toutes ces v.a. sont à valeurs finies. Il reste à appliquer la linéarité de l'espérance conditionnelle et la propriété $\mathrm{XY} = \mathrm{X}^+\mathrm{Y}^+ + \mathrm{X}^-\mathrm{Y}^- - \mathrm{X}^+\mathrm{Y}^- - \mathrm{X}^-\mathrm{Y}^+$. ■

Ensuite, nous insistons sur le fait – important – que les principaux résultats de convergence pour les espérances (ou intégrales) sont aussi valides pour les espérances conditionnelles, à condition de rajouter le qualificatif « p.s. » (qui est implicite dès qu'on parle d'espérance conditionnelle, mais que nous écrivons explicitement dans le théorème suivant).

Théorème 23.8. *Soit* $(\mathrm{Y}_n)_{n \geqslant 1}$ *une suite de v.a. réelles sur* $(\Omega, \mathcal{A}, \mathrm{P})$, *et* $\mathcal{G}$ *une sous-tribu de* $\mathcal{A}$.

(a) *(Convergence monotone.) Si les* Y_n *sont positives et croissent p.s. vers une limite* Y*, alors*

$$\lim_{n \to \infty} \mathrm{E}(\mathrm{Y}_n \,|\, \mathcal{G}) = \mathrm{E}(\mathrm{Y} \,|\, \mathcal{G}) \quad p.s.;$$

(b) *(Lemme de Fatou.) Si les* Y_n *sont positives, on a*

$$E(\liminf_{n\to\infty} Y_n \,|\, \mathcal{G}) \leqslant \liminf_{n\to\infty} E(Y_n \,|\, \mathcal{G}) \quad \textit{p.s.};$$

(c) *(Théorème de convergence dominée de Lebesgue.) Si les* Y_n *convergent p.s. vers une limite* Y, *et si on a* $|Y_n| \leqslant Z$ *pour tout* n *et pour une certaine v.a.* $Z \in L^1(\Omega, \mathcal{A}, P)$, *alors*

$$\lim_{n\to\infty} E(Y_n \,|\, \mathcal{G}) = E(Y \,|\, \mathcal{G}) \quad \textit{p.s.}.$$

Preuve. (a) Par (23.6) on a $E(Y_{n+1} \,|\, \mathcal{G}) \geqslant E(Y_n \,|\, \mathcal{G})$ p.s. pour chaque n. Donc $U = \lim_{n\to\infty} E(Y_n \,|\, \mathcal{G})$ existe p.s. et pour toute v.a. positive et $\mathcal{G}$-mesurable X on a :

$$\begin{aligned} E(UX) &= \lim_{n\to\infty} E(E(Y_n \,|\, \mathcal{G})X) \\ &= \lim_{n\to\infty} E(Y_n X) \\ &= \lim_{n\to\infty} E(YX) \end{aligned}$$

par le théorème de convergence monotone (usuel) et (23.5). Donc $U = E(Y \,|\, \mathcal{G})$, en appliquant de nouveau (23.5).

Pour les démonstrations de (b) et (c) il suffit de reproduire les preuves des résultats similaires pour les espérances (non conditionnelles) et d'utiliser (23.5) dans le même esprit que pour (a) ci-dessus. ■

Terminons ce chapitre avec trois inégalités utiles. Les deux dernières, bien qu'ayant une version pour les espérances conditionnelles, sont données pour les espérances « ordinaires » seulement. Elles se trouvent placées ici parce qu'elles découlent naturellement de la première qui, quant à elle, est très utile dans sa version « espérances conditionnelles ».

Théorème 23.9. (Inégalité de Jensen.) *Soit* $\varphi : \mathbb{R} \to \mathbb{R}$ *une fonction convexe, et soit* X *une v.a. réelle, intégrable et telle que* $\varphi(X)$ *soit aussi intégrable. Pour toute sous-tribu* $\mathcal{G}$ *on a alors*

$$\varphi \circ E(X \,|\, \mathcal{G}) \leqslant E(\varphi(X) \,|\, \mathcal{G}).$$

Preuve. Un résultat d'analyse réelle dit que, si φ est convexe, il existe deux suites (a_n) et (b_n) de réels telles que $\varphi(x) = \sup_n (a_n x + b_n)$ pour tout x. On a alors

$$E(a_n X + b_n \,|\, \mathcal{G}) = a_n E(X \,|\, \mathcal{G}) + b_n.$$

Mais $E(a_n X + b_n \mid \mathcal{G}) \leqslant E(\varphi(X) \mid \mathcal{G})$, donc $a_n E(X \mid \mathcal{G}) + b_n \leqslant E(\varphi(X) \mid \mathcal{G})$ pour chaque n. En prenant le supremum en n, on obtient immédiatement le résultat. ∎

Noter que $\varphi(x) = x^2$ est une fonction convexe. Par suite on déduit de l'inégalité de Jensen que

$$(E(X \mid \mathcal{G}))^2 \leqslant E(X^2 \mid \mathcal{G}).$$

Une conséquence importante de l'inégalité de Jensen est l'inégalité de Hölder pour les variables aléatoires.

Théorème 23.10. (Inégalité de Hölder.) *Soit* X *et* Y *des v.a. réelles, et* $p, q \in]1, \infty[$ *avec* $\frac{1}{p} + \frac{1}{q} = 1$. *On a alors* (avec les conventions $a \times +\infty = +\infty$ si $a > 0$ et $0 \times +\infty = 0$).

$$E(|XY|) \leqslant E(|X|^p)^{\frac{1}{p}} E(|Y|^q)^{\frac{1}{q}}. \tag{23.10}$$

(En particulier, si $X \in L^p$ *et* $Y \in L^q$ *avec* p, q *comme ci-dessus, le produit* XY *appartient à* L^1*).*

Preuve. Si le membre de droite de (23.10) est infini il n'y a rien à démontrer. S'il est nul l'un des deux termes du produit est nul, disons le premier par exemple; donc $X = 0$ p.s. et le membre de gauche de (23.10) est aussi nul. On peut donc supposer que $0 < E(X^p) < \infty$ et que $0 < E(Y^q) < \infty$, et également, sans perte de généralité, que X et Y sont positives. Posons $C = E(X^p) < \infty$, et définissons une nouvelle probabilité Q en posant pour $A \in \mathcal{A}$:

$$Q(A) = \frac{1}{C}\, E(1_A X^p).$$

Posons ensuite $Z = \frac{Y}{X^{p-1}} 1_{\{X>0\}}$. Comme $\varphi(x) = |x|^p$ est convexe, l'inégalité de Jensen donne

$$(E_Q(Z))^q \leqslant E(Z^q).$$

Donc

$$\begin{aligned}
\frac{1}{C^q}\, E(XY)^q &= \frac{1}{C^q}\, E\left(\frac{Y}{X^{p-1}} X^p\right)^q \\
&= \left(E_Q\left(\frac{Y}{X^{p-1}}\right)\right)^q \\
&\leqslant E_Q\left(\left(\frac{Y}{X^{p-1}}\right)^q\right)
\end{aligned}$$

$$= \frac{1}{\mathrm{C}}\,\mathrm{E}\left(\left(\frac{\mathrm{Y}}{\mathrm{X}^{p-1}}\right)^q \mathrm{X}^p\right)$$
$$= \frac{1}{\mathrm{C}}\,\mathrm{E}\left(\mathrm{Y}^q \frac{1}{\mathrm{X}^{(p-1)q}}\mathrm{X}^p\right),$$

et comme $q = \frac{p}{p-1}$ et $(p-1)q = p$,

$$\frac{1}{\mathrm{C}^q}\,\mathrm{E}(\mathrm{XY})^q = \frac{1}{\mathrm{C}}\,\mathrm{E}\left(\mathrm{Y}^q \frac{1}{\mathrm{X}^p}\mathrm{X}^p\right)$$
$$= \frac{1}{\mathrm{C}}\,\mathrm{E}(\mathrm{Y}^q).$$

Par suite

$$\mathrm{E}(\mathrm{XY})^q \leqslant \mathrm{C}^{q-1}\mathrm{E}(\mathrm{Y}^q),$$

et en prenant la racine q-ième on obtient

$$\mathrm{E}(\mathrm{XY}) \leqslant \mathrm{C}^{\frac{q-1}{q}}\,\mathrm{E}(\mathrm{Y}^q)^{\frac{1}{q}}.$$

Comme $\frac{q-1}{q} = \frac{1}{p}$ et $\mathrm{C} = \mathrm{E}(\mathrm{X}^p)$, on a le résultat. ∎

Corollaire 23.1. (Inégalité de Minkowski.) *Soit* X *et* Y *deux v.a. réelles. On a alors*

$$\mathrm{E}(|\mathrm{X}+\mathrm{Y}|^p)^{\frac{1}{p}} \leqslant \mathrm{E}(\mathrm{X}^p)^{\frac{1}{p}} + \mathrm{E}(\mathrm{Y}^p)^{\frac{1}{p}}.$$

Preuve. Si $p = 1$ le résultat est évident. On suppose donc dans la suite que $1 < p < \infty$, et on prend q tel que $\frac{1}{p} + \frac{1}{q} = 1$. D'après l'inégalité de Hölder (théorème 23.10) on a

$$\begin{aligned}\mathrm{E}(|\mathrm{X}+\mathrm{Y}|^p) &= \mathrm{E}\left(|\mathrm{X}|\,|\mathrm{X}+\mathrm{Y}|^{p-1}\right) + \mathrm{E}\left(|\mathrm{Y}|\,|\mathrm{X}+\mathrm{Y}|^{p-1}\right)\\ &\leqslant \mathrm{E}(|\mathrm{X}|^p)^{\frac{1}{p}}\mathrm{E}(|\mathrm{X}+\mathrm{Y}|^{(p-1)q})^{\frac{1}{q}} + \mathrm{E}(|\mathrm{Y}|^p)^{\frac{1}{p}}\mathrm{E}(|\mathrm{X}+\mathrm{Y}|^{(p-1)q})^{\frac{1}{q}}.\end{aligned}$$

Comme $(p-1)q = p$ et $\frac{1}{q} = 1 - \frac{1}{p}$,

$$= \left(\mathrm{E}(|\mathrm{X}|^p)^{\frac{1}{p}} + \mathrm{E}(|\mathrm{Y}|^p)^{\frac{1}{p}}\right)\mathrm{E}(|\mathrm{X}+\mathrm{Y}|^p)^{1-\frac{1}{p}}$$

et on a le résultat. ∎

L'inégalité de Minkowski permet de définir une norme (satisfaisant l'inégalité triangulaire) sur l'espace L^p des (classes d'équivalences de) v.a. vérifiant $\mathrm{E}(|\mathrm{X}|^p) < \infty$:

Définition 23.5. *Pour* $X \in L^p$ *on pose*

$$\|X\|_p = E(|X^p|)^{\frac{1}{p}}.$$

Avec cette norme l'espace vectoriel L^p est *complet* (on dit que c'est un « espace de Banach ») ; mais, pour $p \neq 2$, ce n'est pas un espace de Hilbert : la norme n'est pas associée à un produit scalaire.

Exercices

Pour les exercices 1 à 6 on considère une v.a. positive Y sur un espace $(\Omega, \mathcal{A}, P)$, et une sous-tribu $\mathcal{G}$ de $\mathcal{A}$.

1. Montrer que $|E(Y \mid \mathcal{G})| \leqslant E(|Y| \mid \mathcal{G})$.
2. Si $\mathcal{H}$ est une sous-tribu de $\mathcal{G}$, montrer que

$$E\left(E(Y \mid \mathcal{G}) \mid \mathcal{H}\right) = E(Y \mid \mathcal{H}).$$

3. Montrer que $E(Y \mid Y) = Y$ p.s.
4. Montrer que si $|Y| \leqslant c$ p.s., alors $|E(Y \mid \mathcal{G})| \leqslant c$ p.s. également.
5. Si $Y = \alpha$ p.s., où α est une constante, montrer que $E(Y \mid \mathcal{G}) = \alpha$ p.s.
6. Montrer que $\{E(Y \mid \mathcal{G}) = 0\} \subset \{Y = 0\}$ et $\{Y = +\infty\} \subset \{E(Y \mid \mathcal{G}) = +\infty\}$ p.s. (rappelons que $Y \geqslant 0$).
7.* Soit X et Y des v.a. réelles indépendantes, et f une fonction borélienne telle que $f(X, Y) \in L^1(\Omega, \mathcal{A}, P)$. On pose

$$g(x) = \begin{cases} E(f(x, Y)) & \text{si } |E(f(x, Y))| < \infty, \\ 0 & \text{sinon.} \end{cases}$$

Montrer que g est borélienne et que

$$E(f(X, Y) \mid X) = g(X).$$

8. Soit $Y \in L^2(\Omega, \mathcal{A}, P)$ et supposons que $E(Y^2 \mid X) = X^2$ et $E(Y \mid X) = X$. Montrer que $Y = X$ p.s.
9.* Soit Y une v.a. exponentielle de paramètre 1 (i.e., $P(Y > t) = e^{-t}$ pour $t > 0$). Calculer $E(Y \mid Y \wedge t)$.
10. (Inégalité de Bienaymé-Chebyshev.) Montrer que si $X \in L^2$ et $a > 0$, on a $P(|X| \geqslant a \mid \mathcal{G}) \leqslant \frac{E(X^2 \mid \mathcal{G})}{a^2}$.

11. (Cauchy-Schwarz.) Si $X, Y \in L^2$ montrer que

$$(E(XY \mid \mathcal{G}))^2 \leqslant E(X^2 \mid \mathcal{G})E(Y^2 \mid \mathcal{G}).$$

12. Si $X \in L^2$, montrer que

$$E\big((X - E(X \mid \mathcal{G}))^2\big) \leqslant E\big((X - E(X))^2\big).$$

13. Si $r \geqslant p \geqslant 1$, montrer que $L^p(\Omega, \mathcal{A}, P) \supset L^r(\Omega, \mathcal{A}, P)$.

14.* Soit Z une v.a. positive sur $(\Omega, \mathcal{F}, P)$, avec $E(Z) = 1$. On définit une nouvelle probabilité Q en posant $Q(\Lambda) = E(1_\Lambda Z)$. Soit $\mathcal{G}$ une sous-tribu de $\mathcal{F}$, et $U = E(Z \mid \mathcal{G})$. Montrer que $E_Q(X \mid \mathcal{G}) = \frac{E(XZ \mid \mathcal{G})}{U}$, pour toute v.a. bornée et $\mathcal{F}$-mesurable X. (Ici, $E_Q(X \mid \mathcal{G})$ désigne l'espérance conditionnelle relativement à la probabilité Q ; on utilise la convention $\frac{0}{0} = 0$.)

15. Montrer que l'espace vectoriel normé L^p est complet. (*Indication :* voir la preuve du théorème 22.2.)

16. Soit $X \in L^1(\Omega, \mathcal{F}, P)$ et soit $\mathcal{G}$ et $\mathcal{H}$ des sous-tribus de $\mathcal{F}$. Supposons de plus $\mathcal{H}$ indépendante de $\sigma(\sigma(X), \mathcal{G})$. Montrer que

$$E(X \mid \sigma(\mathcal{G}, \mathcal{H})) = E(X \mid \mathcal{G}).$$

17. Soit $(X_n)_{n \geqslant 1}$ des v.a. indépendantes intégrables, et posons $S_n = \sum_{i=1}^n X_i$, et $\mathcal{G}_n = \sigma(S_n, S_{n+1}, \dots)$. Montrer que $E(X_1 \mid \mathcal{G}_n) = E(X_1 \mid S_n)$ et que $E(X_j \mid \mathcal{G}_n) = E(X_j \mid S_n)$ pour $1 \leqslant j \leqslant n$. Montrer aussi que $E(X_j \mid \mathcal{G}_n) = E(X_1 \mid S_n)$ pour $1 \leqslant j \leqslant n$. (*Indication :* Utiliser l'exercice 16.)

Chapitre 24

Martingales

Rappelons la loi forte des grands nombres (théorème 20.2) : si les v.a. $(X_n)_{n\geqslant 1}$ sont i.i.d. avec $E(X_n) = \mu$ et $\sigma^2_{X_n} < \infty$ (cette dernière hypothèse pouvant d'ailleurs être supprimée), et si $S_n = \sum_{j=1}^n X_j$, alors $\lim_{n\to\infty} \frac{S_n}{n} = \mu$ p.s. Comme les X_n sont indépendantes, la limite doit effectivement être p.s. constante, en vertu de la loi zéro-un (théorème 10.6). Il est évidemment intéressant d'étudier aussi des suites de v.a. qui convergent p.s. vers une limite qui est une « vraie » variable aléatoire, non constante.

La propriété qui va permettre de généraliser la loi des grands nombres est la suivante : si on considère les tribus $\mathcal{F}_n = \sigma\{S_k; k \leqslant n\}$, alors

$$E(S_{n+1} - (n+1)\mu \,|\, \mathcal{F}_n) = S_n - n\mu, \tag{24.1}$$

(voir l'exemple 1 ci-dessous). C'est la propriété (24.1) qui est essentielle quand on veut étudier les convergences en affaiblissant l'hypothèse d'indépendance.

Dans la suite de ce chapitre on fixe l'espace de probabilité $(\Omega, \mathcal{F}, P)$, ainsi que la suite croissante $(\mathcal{F}_n)_{n\geqslant 0}$ de sous-tribus de $\mathcal{F}$: on a donc $\mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \mathcal{F}$ pour tout $n \geqslant 0$. Une telle suite de tribus s'appelle une *filtration*. On considérera aussi des suites $(X_n)_{n\geqslant 0}$ de v.a. réelles indicées par $\mathbb{N}$: dans le contexte de la théorie des martingales, une telle suite s'appelle aussi un *processus*, en référence à la description d'un phénomène qui évolue de manière aléatoire au cours du temps (les instants sont les entiers).

Définition 24.1. *Une suite $(X_n)_{n\geqslant 0}$ de v.a. réelles (autrement dit, un processus) est appelée une* martingale *si*

(i) $E(|X_n|) < \infty$ *pour chaque n ;*

(ii) X_n *est $\mathcal{F}_n$-mesurable pour chaque n ;*

(iii) $E(X_n \,|\, \mathcal{F}_m) = X_m$ *p.s. pour tous $m \leqslant n$.*

Noter que (ii) est « presque » entraîné par (iii), qui implique que X_m est p.s. égale à une variable $\mathcal{F}_m$-mesurable.

Exemple 1. Soit $(X_n)_{n\geqslant 1}$ des v.a. indépendantes intégrables centrées ; soit $\mathcal{F}_n = \sigma\{X_k ; k \leqslant n\}$ et $S_n = \sum_{k=1}^n X_k$ (avec $\mathcal{F}_0 = \{\emptyset, \Omega\}$ = « la tribu triviale » et $S_0 = 0$). La suite $(S_n)_{n\geqslant 0}$ est alors une martingale, ce qui suit de :

$$\begin{aligned}
E(S_n \mid \mathcal{F}_m) &= E(S_m + (S_n - S_m) \mid \mathcal{F}_m) \\
&= S_m + E(S_n - S_m \mid \mathcal{F}_m) \\
&= S_m + E\left(\sum_{k=m+1}^{n} X_k \mid \mathcal{F}_m\right) \\
&= S_m + \sum_{k=m+1}^{n} E(X_k) \\
&= S_m.
\end{aligned}$$

Si les X_n ne sont pas centrées, en utilisant les v.a. $X_n - \mu$ (où $\mu = E(X_1)$) on obtient de la même manière (24.1), i.e. la suite $(S_n - n\mu)$ est une martingale.

Exemple 2. Soit Y une v.a. intégrable et $(\mathcal{F}_n)_{n\geqslant 0}$ une famille croissante de sous-tribus. Si on pose

$$X_n = E(Y \mid \mathcal{F}_n)$$

on obtient une martingale. En effet on a $E(|X_n|) \leqslant E(|Y|) < \infty$ et pour $m \leqslant n$:

$$\begin{aligned}
E(X_n \mid \mathcal{F}_m) &= E(E(Y \mid \mathcal{F}_n) \mid \mathcal{F}_m) \\
&= E(Y \mid \mathcal{F}_m) \\
&= X_m
\end{aligned}$$

(voir les exercices 1 et 2 du chapitre précédent).

Définition 24.2. *Une martingale* $X = (X_n)_{n\geqslant 0}$ *est dite* fermée *par une variable aléatoire* Y *si cette variable est intégrable et si* $X_n = E(Y|\mathcal{F}_n)$ *pour tout* n.

L'exemple 2 montre que toute v.a. Y intégrable donne une martingale fermée, en posant précisément $X_n = E(Y|\mathcal{F}_n)$, $n \geqslant 0$.

Une des propriétés importantes des martingales est que leur espérance est constante :

Théorème 24.1. *Si* $(X_n)_{n\geqslant 0}$ *est une martingale, alors* $n \mapsto E(X_n)$ *est constant, i.e.* $E(X_n) = E(X_0)$ *pour tout* $n \geqslant 0$.

Preuve. $E(X_n) = E(E(X_n \mid \mathcal{F}_0)) = E(X_0)$. ∎

La réciproque de ce théorème est grossièrement fausse, mais il existe une réciproque partielle qui utilise les temps d'arrêt : voir le théorème 24.7.

Définition 24.3. *Une variable aléatoire* $T : \Omega \to \overline{\mathbb{N}} = \mathbb{N} \cup \{+\infty\}$ *est appelée un temps d'arrêt si* $\{T \leqslant n\} \in \mathcal{F}_n$ *pour chaque* n.

Toute constante entière, ou égale à $+\infty$, est évidemment un temps d'arrêt. Les temps d'arrêt sont parfois plus utiles que les temps « fixes ». Cette terminologie suggère clairement que $\mathbb{N}$ est un ensemble de « temps » tandis que $\mathcal{F}_n$ représente la famille des événements qu'on peut observer avant ou à l'instant n. On peut alors imaginer un temps d'arrêt comme le premier instant où une certaine suite de v.a. X_n, avec chaque X_n observable à l'instant n (i.e. $\mathcal{F}_n$-mesurable), vérifie une certaine propriété, avec la convention que le temps d'arrêt vaut $+\infty$ si cela n'arrive jamais. Par exemple (X_n) est une martingale, et T est le premier temps auquel elle atteint ou dépasse 12 ; on peut écrire

$$T = \begin{cases} \inf\limits_{n \geqslant 0}\{n : X_n \geqslant 12\} & \text{si } X_n \geqslant 12 \text{ pour un } n \in \mathbb{N} \\ +\infty & \text{sinon,} \end{cases}$$

ou encore,

$$T(\omega) = \inf_{n \geqslant 0}\{n : X_n(\omega) \geqslant 12\}$$

si $X_n(\omega) \geqslant 12$ pour au moins un entier, et $T(\omega) = +\infty$ sinon. Noter que

$$\{T \leqslant n\} = \bigcup_{k=0}^{n}\{X_k \geqslant 12\} \in \mathcal{F}_n$$

parce que $\{X_k \geqslant 12\} \in \mathcal{F}_k \subset \mathcal{F}_n$ si $k \leqslant n$, donc T est bien un temps d'arrêt. La terminologie « temps d'arrêt » vient de la modélisation des jeux de hasard : un joueur peut arrêter de jouer à un temps aléatoire (dépendant de ses pertes antérieures par exemple), mais uniquement sur la base des événements effectivement observables avant l'instant ou à l'instant où il s'arrête : le lecteur vérifiera que mathématiquement, cette notion équivaut à celle de la définition 24.3.

Le théorème 24.1 s'étend sans difficulté aux temps d'arrêts T qui sont bornés (i.e., il existe une constante c telle que $T \leqslant c$ identiquement). Si T est un temps d'arrêt *fini*, on notera X_T la variable $X_T(\omega) = X_{T(\omega)}(\omega) = \sum_{n=0}^{\infty} X_n(\omega) 1_{\{T(\omega)=n\}}$.

Théorème 24.2. *Soit* T *un temps d'arrêt borné par* c, *et* $(X_n)_{n\geqslant 0}$ *une martingale. On a alors* $E(X_T) = E(X_0)$.

Preuve. En supposant (sans perte de généralité) que c est un entier, on a

$$\begin{aligned} E(X_T) &= E\left(\sum_{n=0}^{\infty} X_n 1_{\{T=n\}}\right) \\ &= E\left(\sum_{n=0}^{c} X_n 1_{\{T=n\}}\right) \\ &= \sum_{n=0}^{c} E(X_n 1_{\{T=n\}}). \end{aligned}$$

Comme $\{T = n\} = \{T \leqslant n\} \setminus \{T \leqslant n-1\}$ on voit que $\{T = n\} \in \mathcal{F}_n$, et par suite

$$\begin{aligned} &= \sum_{n=0}^{c} E(E(X_c \,|\, \mathcal{F}_n)\ 1_{\{T=n\}}) \\ &= \sum_{n=0}^{c} E(X_c 1_{\{T=n\}}) \\ &= E\left(X_c \sum_{n=0}^{c} 1_{\{T=n\}}\right) \\ &= E(X_c) = E(X_0), \end{aligned}$$

où la dernière égalité vient du théorème 24.1. ■

La tribu $\mathcal{F}_n$ représente la classe des événements observables jusqu'à l'instant n ; on veut généraliser cette notion aux temps d'arrêt.

Définition 24.4. *Soit* T *un temps d'arrêt. La* tribu antérieure à T *est la tribu*

$$\mathcal{F}_T = \{A \in \mathcal{F} : A \cap \{T \leqslant n\} \in \mathcal{F}_n \quad \textit{pour tout } n\}.$$

Pour que cette définition ait un sens, nous avons besoin du résultat suivant :

Théorème 24.3. *Si* T *est un temps d'arrêt, la classe* $\mathcal{F}_T$ *définie ci-dessus est une tribu.*

Preuve. Que Ω soit dans $\mathcal{F}_T$ est évident. Si $A \in \mathcal{F}_T$ on a

$$A^c \cap \{T \leqslant n\} = \{T \leqslant n\} \setminus (A \cap \{T \leqslant n\}),$$

donc $A^c \in \mathcal{F}_T$. Enfin si les $(A_i)_{i\geqslant 1}$ sont dans $\mathcal{F}_T$, alors

$$\left(\bigcup_{i=1}^{\infty} A_i\right) \cap \{T \leqslant n\} = \bigcup_{i=1}^{\infty}(A_i \cap \{T \leqslant n\}) \in \mathcal{F}_n.$$

Donc $\mathcal{F}_T$, étant stable par complémentation et par réunions dénombrables, est une tribu. ■

Théorème 24.4. *Si* S *et* T *sont des temps d'arrêt vérifiant* $S \leqslant T$*, on a* $\mathcal{F}_S \subset \mathcal{F}_T$.

Preuve. Comme $S \leqslant T$ on a $\{T \leqslant n\} \subset \{S \leqslant n\}$. Donc si $A \in \mathcal{F}_S$ il vient

$$A \cap \{T \leqslant n\} = A \cap \{S \leqslant n\} \cap \{T \leqslant n\};$$

mais $A \cap \{S \leqslant n\} \in \mathcal{F}_n$ et $\{T \leqslant n\} \in \mathcal{F}_n$, donc $A \cap \{T \leqslant n\} \in \mathcal{F}_n$, donc $A \in \mathcal{F}_T$. ■

Dans le théorème suivant on suppose donnée une suite $(X_n)_{n\geqslant 0}$ de v.a. réelles, telle que chaque X_n soit $\mathcal{F}_n$-mesurable. Soit T un temps d'arrêt *fini*, de sorte que X_T soit bien défini.

Théorème 24.5. X_T *est* $\mathcal{F}_T$*-mesurable.*

Preuve. Soit $\Lambda \in \mathcal{B}$. On veut montrer que $\{X_T \in \Lambda\} \in \mathcal{F}_T$, ce qui signifie que pour tout n on a $\{X_T \in \Lambda\} \cap \{T \leqslant n\} \in \mathcal{F}_n$. Mais

$$\begin{aligned}\{X_T \in \Lambda\} \cap \{T \leqslant n\} &= \bigcup_{k=1}^{n} \{X_T \in \Lambda\} \cap \{T = k\} \\ &= \bigcup_{k=1}^{n} \{X_k \in \Lambda\} \cap \{T = k\},\end{aligned}$$

et $\{X_k \in \Lambda\} \cap \{T = k\} \in \mathcal{F}_k \subset \mathcal{F}_n$ pour $k \leqslant n$. ■

Les deux théorèmes suivants montrent que la « propriété de martingale » est valide pour les temps d'arrêt aussi bien que pour les temps fixes. Il s'agit de résultats simples, mais très puissants.

Théorème 24.6. (Théorème d'arrêt de Doob.) *Soit* $(X_n)_{n\geqslant 0}$ *une martingale. Soit* S *et* T *deux temps d'arrêt bornés par une constante* c*, avec* $S \leqslant T$ *p.s. On a alors*

$$E(X_T \mid \mathcal{F}_S) = X_S \qquad p.s.$$

Preuve. D'abord, $|X_T| \leqslant \sum_{n=0}^{c} |X_n|$ est intégrable (on a supposé, sans perte de généralité, que c est un entier). De plus X_S est $\mathcal{F}_S$-mesurable en vertu du théorème précédent. Il reste donc à prouver que $E(X_T Z) = E(X_S Z)$ pour toute v.a. Z bornée et $\mathcal{F}_S$-mesurable. Grâce à un argument standard il suffit même de montrer que pour tout $A \in \mathcal{F}_S$, alors

$$E(X_T 1_A) = E(X_S 1_A)$$

(ceci implique par linéarité qu'on a $E(X_T Z) = E(X_S Z)$ pour les v.a. Z qui sont « simples », puis par limite croissante que c'est vrai pour Z positive, et enfin par différence que c'est vrai pour Z bornée).

Soit donc $A \in \mathcal{F}_S$. Définissons un nouveau « temps aléatoire » R par

$$R(\omega) = S(\omega) 1_A(\omega) + T(\omega) 1_{A^c}(\omega).$$

R est aussi un temps d'arrêt : en effet, on a

$$\{R \leqslant n\} = (A \cap \{S \leqslant n\}) \cup (A^c \cap \{T \leqslant n\}),$$

et $A \cap \{S \leqslant n\} \in \mathcal{F}_n$ car $A \in \mathcal{F}_S$. Comme $A \in \mathcal{F}_S$ on a $A^c \in \mathcal{F}_S$, donc $A^c \in \mathcal{F}_T$ par le théorème 24.4. Donc $A^c \cap \{T \leqslant n\} \in \mathcal{F}_n$, d'où on déduit que $\{R \leqslant n\} \in \mathcal{F}_n$. Donc $E(X_R) = E(X_T) = E(X_0)$ par le théorème 24.2. Mais

$$\begin{aligned} E(X_R) &= E(X_S 1_A + X_T 1_{A^c}), \\ E(X_T) &= E(X_T 1_A + X_T 1_{A^c}) \end{aligned}$$

et en soustrayant on obtient

$$E(X_S 1_A) - E(X_T 1_A) = 0. \qquad \blacksquare$$

On peut maintenant établir une réciproque partielle au théorème 24.1.

Théorème 24.7. *Soit $(X_n)_{n \geqslant 0}$ une suite de v.a. intégrables, telle que chaque X_n soit $\mathcal{F}_n$-mesurable. Si $E(X_T) = E(X_0)$ pour tout temps d'arrêt T borné, alors (X_n) est une martingale.*

Preuve. Soit $0 \leqslant m < n < \infty$ et $\Lambda \in \mathcal{F}_m$. Posons

$$\mathbf{T}(\omega) = \begin{cases} m & \text{si } \omega \in \Lambda^c, \\ n & \text{si } \omega \in \Lambda \end{cases}$$

T est un temps d'arrêt (vérification immédiate), donc

$$E(X_0) = E(X_T) = E(X_m 1_{\Lambda^c} + X_n 1_\Lambda).$$

Cependant on a aussi $\mathrm{E}(\mathrm{X}_0) = \mathrm{E}(\mathrm{X}_m 1_{\Lambda^c} + \mathrm{X}_m 1_\Lambda)$ (appliquer l'hypothèse avec $\mathrm{T} \equiv m$). En soustrayant, on obtient $\mathrm{E}(\mathrm{X}_n 1_\Lambda) = \mathrm{E}(\mathrm{X}_m 1_\Lambda)$, et comme Λ est arbitraire dans $\mathcal{F}_m$ on obtient $\mathrm{E}(\mathrm{X}_n \mid \mathcal{F}_m) = \mathrm{X}_m$ p.s. ∎

Exercices

Dans les exercices 1 à 11 on considère deux temps d'arrêt S et T relatifs à une suite croissante donnée $(\mathcal{F}_n)_{n\geqslant 0}$ de sous-tribus.

1. Si $\mathrm{T} \equiv n$ montrer que $\mathcal{F}_\mathrm{T} = \mathcal{F}_n$.
2. Montrer que $\mathrm{S} \wedge \mathrm{T} = \min(\mathrm{S}, \mathrm{T})$ est un temps d'arrêt.
3. Montrer que $\mathrm{S} \vee \mathrm{T} = \max(\mathrm{S}, \mathrm{T})$ est un temps d'arrêt.
4. Montrer que $\mathrm{S} + \mathrm{T}$ est un temps d'arrêt.
5. Montrer que $\alpha\mathrm{T}$ est un temps d'arrêt pour a entier supérieur ou égal à 1.
6. Montrer que $\mathcal{F}_{\mathrm{S}\wedge\mathrm{T}} \subset \mathcal{F}_\mathrm{T} \subset \mathcal{F}_{\mathrm{S}\vee\mathrm{T}}$.
7. Montrer que T est un temps d'arrêt si et seulement si $\{\mathrm{T} = n\} \in \mathcal{F}_n$ pour tout $n \geqslant 0$.
8. Soit $\Lambda \in \mathcal{F}_\mathrm{T}$ et posons
$$\mathrm{T}_\Lambda(\omega) = \begin{cases} \mathrm{T}(\omega) & \text{si } \omega \in \Lambda, \\ \infty & \text{si } \omega \notin \Lambda. \end{cases}$$
Montrer que T_Λ est aussi un temps d'arrêt.
9. Montrer que T est $\mathcal{F}_\mathrm{T}$-mesurable.
10. Montrer que $\{\mathrm{S} < \mathrm{T}\}$, $\{\mathrm{S} \leqslant \mathrm{T}\}$, et $\{\mathrm{S} = \mathrm{T}\}$ sont tous dans $\mathcal{F}_\mathrm{S} \cap \mathcal{F}_\mathrm{T}$.
11.* Montrer que si Y est une v.a. intégrable,
$$\mathrm{E}(\mathrm{E}(\mathrm{Y} \mid \mathcal{F}_\mathrm{T}) \mid \mathcal{F}_\mathrm{S}) = \mathrm{E}(\mathrm{E}(\mathrm{Y}_m id \mathcal{F}_\mathrm{S}) \mid \mathcal{F}_\mathrm{T}) = \mathrm{E}(\mathrm{Y} \mid \mathcal{F}_{\mathrm{S}\wedge\mathrm{T}}).$$
12. Soit $(\mathrm{M}_n)_{n\geqslant 0}$ une martingale avec $\mathrm{M}_n \in \mathrm{L}^2$ pour tout n. Soit S et T des temps d'arrêt bornés avec $\mathrm{S} \leqslant \mathrm{T}$. Montrer que M_S et M_T sont dans L^2, que
$$\mathrm{E}((\mathrm{M}_\mathrm{T} - \mathrm{M}_\mathrm{S})^2 \mid \mathcal{F}_\mathrm{S}) = \mathrm{E}(\mathrm{M}_\mathrm{T}^2 - \mathrm{M}_\mathrm{S}^2 \mid \mathcal{F}_\mathrm{S}),$$
et que
$$\mathrm{E}((\mathrm{M}_\mathrm{T} - \mathrm{M}_\mathrm{S})^2) = \mathrm{E}(\mathrm{M}_\mathrm{T}^2) - \mathrm{E}(\mathrm{M}_\mathrm{S}^2).$$

13. Soit φ une fonction convexe et soit $(M_n)_{n\geqslant 0}$ une martingale. Montrer que $n \to E(\varphi(M_n))$ est une fonction croissante. (*Indication :* Utiliser l'inégalité de Jensen [théorème 23.9].)
14. Soit X_n une suite de v.a. réelles intégrables avec $E(X_n \mid \mathcal{F}_{n-1}) = 0$ et X_n $\mathcal{F}_n$-mesurable pour chaque n. Soit $S_n = \sum_{k=0}^{n} X_k$. Montrer que $(S_n)_{n\geqslant 0}$ est une martingale.

Chapitre 25

Surmartingales et sous-martingales

Au chapitre 24 nous avons défini les martingales via une égalité pour certaines espérances conditionnelles. Si on remplace cette égalité par une inégalité, on obtient les surmartingales et les sous-martingales. Nous supposons de nouveau donné un espace de probabilité $(\Omega, \mathcal{F}, \mathrm{P})$ muni d'une suite croissante $(\mathcal{F}_n)_{n>0}$ de sous-tribus de $\mathcal{F}$.

Définition 25.1. *Une suite de v.a. réelles* $(\mathrm{X}_n)_{n\geqslant 0}$ *(un processus) est appelée une* sous-martingale *(resp. une* surmartingale*) si*

(i) $\mathrm{E}(|\mathrm{X}_n|) < \infty$ *pour chaque* n *;*

(ii) X_n *est* $\mathcal{F}_n$*-mesurable pour chaque* n *;*

(iii) $\mathrm{E}(\mathrm{X}_n \,|\, \mathcal{F}_m) \geqslant \mathrm{X}_m$ *p.s. (resp.* $\leqslant \mathrm{X}_m$ *p.s.) pour tous* $m \leqslant n$*.*

La suite $(\mathrm{X}_n)_{n\geqslant 0}$ est une martingale si et seulement si c'est à la fois une sous-martingale *et* une surmartingale.

Théorème 25.1. *Si* $(\mathrm{M}_n)_{n\geqslant 0}$ *est une martingale et si* φ *est une fonction convexe telle que chaque* $\varphi(\mathrm{M}_n)$ *soit intégrable, alors* $(\varphi(\mathrm{M}_n))_{n\geqslant 0}$ *est une sous-martingale.*

Preuve. Soit $m \leqslant n$. On a $\mathrm{E}(\mathrm{M}_n \,|\, \mathcal{F}_m) = \mathrm{M}_m$ p.s., donc $\varphi(\mathrm{E}(\mathrm{M}_n \,|\, \mathcal{F}_m)) = \varphi(\mathrm{M}_m)$ p.s., et comme φ est convexe l'inégalité de Jensen (théorème 23.9) donne

$$\mathrm{E}(\varphi(\mathrm{M}_n) \,|\, \mathcal{F}_m) \geqslant \varphi(\mathrm{E}(\mathrm{M}_n \,|\, \mathcal{F}_m)) = \varphi(\mathrm{M}_m).$$

■

Corollaire 25.1. *Si* $(\mathrm{M}_n)_{n\geqslant 0}$ *est une martingale, alors* $(|\mathrm{M}_n|)_{n\geqslant 0}$ *est une sous-martingale.*

Preuve. Appliquer le théorème 25.1 avec $\varphi(x) = |x|$. ■

Théorème 25.2. *Soit* T *un temps d'arrêt borné par* $\mathrm{C} \in \mathbb{N}$*, et* $(\mathrm{X}_n)_{n\geqslant 0}$ *une sous-martingale. On a alors* $\mathrm{E}(\mathrm{X_T}) \leqslant \mathrm{E}(\mathrm{X_C})$*.*

Preuve. La démonstration, analogue à celle du théorème 24.2, est omise. ■

Le résultat suivant explique un autre type de rapports entre sous-martingales et martingales.

Théorème 25.3. (Décomposition de Doob.) *Soit* $X = (X_n)_{n\geqslant 0}$ *une sous-martingale. Il existe une martingale* $M = (M_n)_{n\geqslant 0}$ *avec* $M_0 = 0$ *et une suite de v.a.* $A = (A_n)_{n\geqslant 0}$ *avec* $A_{n+1} \geqslant A_n$ *p.s. et* $A_0 = 0$ *(on dit :* un processus croissant*), chaque variable* A_{n+1} *étant en outre* $\mathcal{F}_n$*-mesurable, tels que*

$$X_n = X_0 + M_n + A_n.$$

De plus, cette décomposition est p.s. unique.

Preuve. Posons $A_0 = 0$ et

$$A_n = \sum_{k=1}^{n} E(X_k - X_{k-1} \,|\, \mathcal{F}_{k-1}) \quad \text{pour } n \geqslant 1.$$

X étant une sous-martingale, on a $E\{X_k - X_{k-1} | \mathcal{F}_{k-1}\} \geqslant 0$ p.s., donc $A_{k+1} \geqslant A_k$ p.s., tandis que par construction A_{k+1} est $\mathcal{F}_k$-mesurable. Par ailleurs on a

$$E(X_n \,|\, \mathcal{F}_{n-1}) - X_{n-1} = E(X_n - X_{n-1} \,|\, \mathcal{F}_{n-1}) = A_n - A_{n-1},$$

et donc

$$E(X_n \,|\, \mathcal{F}_{n-1}) - A_n = X_{n-1} - A_{n-1};$$

mais $A_n \in \mathcal{F}_{n-1}$, donc

$$E(X_n - A_n \,|\, \mathcal{F}_{n-1}) = X_{n-1} - A_{n-1}. \tag{25.1}$$

Si on pose $M_n = X_n - A_n - X_0$ il suit de (25.1) que $M = (M_n)_{n\geqslant 0}$ est une martingale avec $M_0 = 0$, et on a la décomposition souhaitée.

Quant à l'unicité, supposons que

$$\begin{aligned} X_n &= X_0 + M_n + A_n, \quad n \geqslant 0, \\ X_n &= X_0 + L_n + C_n, \quad n \geqslant 0, \end{aligned}$$

soient deux décompositions. En soustrayant l'une de l'autre, on arrive à

$$L_n - M_n = A_n - C_n. \tag{25.2}$$

Comme A_n et C_n sont $\mathcal{F}_{n-1}$-mesurables, il en est de même de $L_n - M_n$, et donc

$$L_n - M_n = E(L_n - M_n \,|\, \mathcal{F}_{n-1}) = L_{n-1} - M_{n-1} = A_{n-1} - C_{n-1} \quad \text{p.s.}$$

Par une récurrence évidente on obtient alors que $L_n - M_n = L_0 - M_0 = 0$ p.s. (car $L_0 = M_0 = 0$). Par suite $L_n = M_n$ p.s., donc aussi $A_n = C_n$ p.s., et l'unicité est démontrée. ∎

Corollaire 25.2. *Soit* $X = (X_n)_{n\geqslant 0}$ *une surmartingale. Il existe une (p.s.) unique décomposition*

$$X_n = X_0 + M_n - A_n, \quad n \geqslant 0$$

avec $(M_n)_{n\geqslant 0}$ *et* $(A_n)_{n\geqslant 0}$ *comme dans le théorème 25.3.*

Preuve. Soit $Y_n = -X_n$. La suite $(Y_n)_{n\geqslant 0}$ est une sous-martingale, qui admet donc une décomposition de Doob p.s. unique

$$Y_n = Y_0 + L_n + C_n,$$

et il suffit de poser $M_n = -L_n$ et $A_n = C_n$. ■

Exercices

1. Montrer que $X = (X_n)_{n\geqslant 0}$ est une sous-martingale si et seulement si $-X = (-X_n)_{n\geqslant 0}$ est une surmartingale.
2. Montrer que si $X = (X_n)_{n\geqslant 0}$ est à la fois une sous-martingale et une surmartingale, c'est une martingale.
3. Soit $(X_n)_{n\geqslant 0}$ une sous-martingale de décomposition de Doob $X_n = X_0 + M_n + A_n$. Montrer que $E(A_n) < \infty$ pour tout n.
4. Soit $M = (M_n)_{n>0}$ une martingale avec $M_0 = 0$, telle que $E(M_n^2) < \infty$ pour tout n, et soit $X_n = M_n^2$. Montrer que $X = (X_n)_{n\geqslant 0}$ est une sous-martingale, et que sa décomposition de Doob $X_n = L_n + A_n$ vérifie $E(M_n^2) = E(A_n)$.
5. Soit M et A comme dans l'exercice 4. Montrer que $A_n - A_{n-1} = E((M_n - M_{n-1})^2 \,|\, \mathcal{F}_{n-1})$.
6. Soit $(X_n)_{n\geqslant 0}$ une sous-martingale. Montrer que si φ est une fonction convexe croissante sur $\mathbb{R}$ telle que chaque v.a. $\varphi(X_n)$ soit intégrable, alors $(\varphi(X_n))_{n\geqslant 0}$ et aussi une sous-martingale.
7. Soit $(X_n)_{n\geqslant 0}$ une suite croissante de v.a. intégrables, chaque X_n étant $\mathcal{F}_n$-mesurable. Montrer que cette suite est une sous-martingale.

Chapitre 26

Les inégalités de martingales

L'une des raisons qui font que les martingales jouent un rôle central en probabilité est que leur structure engendre plusieurs inégalités très puissantes. Nous suivons ci-dessous la présentation de Bass [1].

L'espace de probabilité $(\Omega, \mathcal{F}, \mathrm{P})$ et la suite croissante $(\mathcal{F}_n)_{n \geqslant 0}$ de sous-tribus de $\mathcal{F}$ sont encore fixés. Soit $\mathrm{M} = (\mathrm{M}_n)_{n \geqslant 0}$ une suite de v.a. intégrables, chaque M_n étant $\mathcal{F}_n$-mesurable. On pose $\mathrm{M}_n^* = \sup_{j \leqslant n} |\mathrm{M}_j|$. Remarquer que $(\mathrm{M}_n^*)_{n \geqslant 0}$ est un « processus croissant », et donc une sous-martingale (voir l'exercice 7 du chapitre 25), puisque

$$\mathrm{E}(\mathrm{M}_n^*) \leqslant \mathrm{E}\left(\sum_{j=1}^{n} |\mathrm{M}_j|\right) < \infty.$$

Grâce à l'inégalité de Markov (corollaire 5.1), on a

$$\mathrm{P}(\mathrm{M}_n^* \geqslant \alpha) = \mathrm{E}(1_{\{\mathrm{M}_n^* \geqslant \alpha\}}) \leqslant \frac{\mathrm{E}(\mathrm{M}_n^*)}{\alpha}.$$

Notre première inégalité indique que dans le cas d'une martingale on peut remplacer M_n^* par $|\mathrm{M}_n|$ dans le membre de droite : on a donc un contrôle du supremum pour les temps $0, 1, \ldots, n$ par la variable « terminale » $|\mathrm{M}_n|$.

Théorème 26.1. (Inégalité maximale de Doob.) *Soit* $\mathrm{M} = (\mathrm{M}_n)_{n \geqslant 0}$ *une martingale, ou une sous-martingale positive. On a alors*

$$\mathrm{P}(\mathrm{M}_n^* \geqslant \alpha) \leqslant \frac{\mathrm{E}(|\mathrm{M}_n|)}{\alpha}.$$

Preuve. Soit $\mathrm{T} = \min\{j : |\mathrm{M}_j| \geqslant \alpha\}$ (rappelons notre convention selon laquelle l'infimum d'un ensemble vide d'entiers vaut $+\infty$). Comme la fonction $\varphi(x) = |x|$ est convexe et croissante sur $\mathbb{R}_+$, le processus $(|\mathrm{M}_n|)$ est une sous-martingale (théorème 25.1 si M est une martingale, exercice 24.6 si M est une sous-martingale positive). Les ensembles $\{\mathrm{T} \leqslant n, |\mathrm{M}_\mathrm{T}| \geqslant \alpha\}$ et $\{\mathrm{M}_n^* \geqslant \alpha\}$ sont égaux, donc

$$\mathrm{P}(\mathrm{M}_n^* \geqslant \alpha) = \mathrm{P}(\mathrm{T} \leqslant n, |\mathrm{M}_\mathrm{T}| \geqslant \alpha) \leqslant \mathrm{E}\left(\frac{|\mathrm{M}_\mathrm{T}|}{\alpha} 1_{\{\mathrm{T} \leqslant n\}}\right).$$

et comme $M_T = M_{T\wedge n}$ sur $\{T \leqslant n\}$, on a

$$P(M_n^* \geqslant \alpha) \leqslant \frac{1}{\alpha}\, E(|M_{T\wedge n}|1_{\{T\leqslant n\}}) \leqslant \frac{E(|M_{T\wedge n}|)}{\alpha} \leqslant \frac{E(|M_n|)}{\alpha}$$

par le théorème 25.2. ■

Avant de prouver notre seconde inégalité, nous avons besoin d'un lemme intéressant en lui-même.

Lemme 26.1. *Soit* X *une v.a. positive et* $p > 0$. *On a alors*

$$E(X^p) = \int_0^\infty p\lambda^{p-1} P(X > \lambda)d\lambda.$$

(Les deux membres de l'égalité ci-dessus peuvent être égaux à $+\infty$.)

Preuve. On a

$$\int_0^\infty p\lambda^{p-1}P(X > \lambda)d\lambda = \int_0^\infty p\lambda^{p-1}E(1_{(X>\lambda)})d\lambda,$$

et le théorème de Fubini (voir l'exercice 15 du chapitre 10) donne

$$= E\left(\int_0^\infty p\lambda^{p-1}1_{(X>\lambda)}d\lambda\right) = E\left(\int_0^X p\lambda^{p-1}d\lambda\right) = E(X^p). \quad ■$$

Théorème 26.2. (Inégalité de Doob.) *Soit* $M = (M_n)_{n\geqslant 0}$ *une sous-martingale positive ou une martingale (de signe quelconque). Soit* $1 < p < \infty$. *Il existe une constante* c_p *dépendant seulement de* p *(mais pas de* M*) telle que*

$$E((M_n^*)^p) \leqslant c_p E(|M_n|^p).$$

Preuve. Le processus $(|M_n|)$ est une sous-martingale : c'est évident si M est une sous-martingale positive, et lorsque M est une martingale c'est le corollaire 25.1. Soit $X_n = M_n 1_{(|M_n|>\frac{\alpha}{2})}$. Pour n fixé, posons

$$Z_j = E(X_n \,|\, \mathcal{F}_j), \quad 0 \leqslant j \leqslant n.$$

Noter que $(Z_j)_{0\leqslant j\leqslant n}$ est une martingale, et aussi que $M_n^* \leqslant Z_n^* + \frac{\alpha}{2}$, puisque

$$\begin{aligned}
|M_j| &\leqslant E(|M_n| \,|\, \mathcal{F}_j)| \\
&= E(|M_n|1_{(|M_n|>\frac{\alpha}{2})} + |M_n|1_{(|M_n|\leqslant\frac{\alpha}{2})} \,|\, \mathcal{F}_j) \\
&= E(X_n + |M_n|1_{(|M_n|\leqslant\frac{\alpha}{2})} \,|\, \mathcal{F}_j) \\
&\leqslant E(X_n \,|\, \mathcal{F}_j) + \frac{\alpha}{2} \\
&= Z_j + \frac{\alpha}{2}.
\end{aligned}$$

Le théorème 26.1 implique alors que

$$\begin{aligned}
P(M_n^* > \alpha) &\leqslant P\left(Z_n^* > \frac{\alpha}{2}\right) \\
&\leqslant \frac{2}{\alpha}E(Z_n) = \frac{2}{\alpha}E(X_n) \\
&= \frac{2}{\alpha}E(|M_n|1_{\{|M_n|>\frac{\alpha}{2}\}}).
\end{aligned}$$

D'après le lemme 26.1 il vient

$$\begin{aligned}
E((M_n^*)^p) &= \int_0^\infty p\lambda^{p-1}P(M_n^* > \lambda)d\lambda \\
&\leqslant \int_0^\infty 2p\lambda^{p-2}E(|M_n|1_{\{|M_n|>\frac{\lambda}{2}\}})d\lambda
\end{aligned}$$

et le théorème de Fubini (voir l'exercice 15 du chapitre 10) entraîne

$$\begin{aligned}
E((M_n^*)^p) &= E\left(\int_0^{2|M_n|} 2p\lambda^{p-2}d\lambda\ |M_n|\right) \\
&= \frac{2^p p}{p-1}E(|M_n|^p).
\end{aligned}$$
■

Remarquer que cette démonstration donne la valeur $\frac{2^p p}{p-1}$ à la constante c_p. Avec un peu plus de travail, on peut montrer que $c_p = (\frac{p}{p-1})^p$ (c'est la valeur optimale de c_p), de sorte qu'on peut reformuler le théorème 26.2 ainsi :

Théorème 26.3. (Inégalité de Doob.) *Soit* $M = (M_n)_{n\geqslant 0}$ *une sous-martingale positive ou une martingale. Soit* $1 < p < \infty$. *On a alors*

$$E((M_n^*)^p)^{\frac{1}{p}} \leqslant \frac{p}{p-1}\ E(|M_n|^p)^{\frac{1}{p}},$$

ou en utilisant la notation de la norme de L^p :

$$\|M_n^*\|_p \leqslant \frac{p}{p-1}\ \|M_n\|_p.$$

La dernière inégalité de ce chapitre sera utilisée pour démontrer le théorème de convergence des martingales au chapitre 27. Introduisons d'abord la notion de « montées » : soit $X = (X_n)_{n\geqslant 0}$ un processus, et soit $a < b$. Le nombre de montées de X de a à b est le nombre de fois où le processus X « monte » d'un point en dessous de a jusqu'à un point au dessus de b à un instant ultérieur. On peut exprimer ceci d'une manière plus mathématique en utilisant les temps d'arrêt. Posons

$$\mathrm{T}_0 = 0,$$

et définissons par récurrence sur $j \geqslant 0$ les S_{j+1} et T_{j+1} :

$$\mathrm{S}_{j+1} = \min\{k \geqslant \mathrm{T}_j : \mathrm{X}_k \leqslant a\}, \qquad \mathrm{T}_{j+1} = \min\{k > \mathrm{S}_{j+1} : \mathrm{X}_k \geqslant b\}, \tag{26.1}$$

avec la convention usuelle que l'infimum de l'ensemble vide vaut $+\infty$; avec la convention symétrique que le supremum de l'ensemble vide vaut 0, on peut enfin poser

$$\mathrm{U}_n = \max\{j : \mathrm{T}_j \leqslant n\} \tag{26.2}$$

et U_n est exactement le nombre de montées de a à b pour X, entre les instants 0 et n.

Théorème 26.4. (Inégalité des montées de Doob.) *Soit* $\mathrm{X} = (\mathrm{X}_n)_{n\geqslant 0}$ *une sous-martingale. Soit* $a < b$ *et notons* U_n *le nombre de montées de* a *à* b *pour* X, *entre les temps* 0 *et* n *(cf.* (26.2)*). On a alors*

$$\mathrm{E}(\mathrm{U}_n) \leqslant \frac{1}{b-a}\ \mathrm{E}((\mathrm{X}_n - a)^+)$$

où $(\mathrm{X}_n - a)^+ = \max(\mathrm{X}_n - a, 0)$.

Preuve. Soit $\mathrm{Y}_n = (\mathrm{X}_n - a)^+$. La fonction $\varphi(x) = (x-a)^+$ étant convexe et croissante, d'après l'exercice 6 du chapitre 25, $\mathrm{Y} = (\mathrm{Y}_n)_{n\geqslant 0}$ est une sous-martingale positive. Comme à l'évidence $\mathrm{S}_{n+1} > n$, on a l'égalité suivante :

$$\mathrm{Y}_n = \mathrm{Y}_{\mathrm{S}_1\wedge n} + \sum_{i=1}^{n} (\mathrm{Y}_{\mathrm{T}_i\wedge n} - \mathrm{Y}_{\mathrm{S}_i\wedge n}) + \sum_{i=1}^{n} \left(\mathrm{Y}_{\mathrm{S}_{i+1}\wedge n} - \mathrm{Y}_{\mathrm{T}_i\wedge n}\right). \tag{26.3}$$

Par ailleurs chaque montée de X entre les instants 0 et n correspond à l'existence d'un entier i tel que $\mathrm{S}_i < \mathrm{T}_i \leqslant n$, avec $\mathrm{Y}_{\mathrm{S}_i} = 0$ et $\mathrm{Y}_{\mathrm{T}_i} = \mathrm{Y}_{\mathrm{T}_i\wedge n} \geqslant b-a$, tandis qu'on a par construction $\mathrm{Y}_{\mathrm{T}_i\wedge n} - \mathrm{Y}_{\mathrm{S}_i\wedge n} \geqslant 0$ pour tout i. Par suite

$$\sum_{i=1}^{n} (\mathrm{Y}_{\mathrm{T}_i\wedge n} - \mathrm{T}_{\mathrm{S}_i\wedge n}) \geqslant (b-a)\mathrm{U}_n.$$

Étant donné (26.3), et comme $Y_{S_1 \wedge n} \geqslant 0$, il vient alors

$$(b-a)U_n \leqslant Y_n - \sum_{i=1}^{n} \left(Y_{S_{i+1}\wedge n} - Y_{T_i \wedge n}\right).$$

On prend alors l'espérance des deux membres. Comme Y est une sous-martingale, comme les temps d'arrêt $T_i \wedge n$ et $S_{i+1} \wedge n$ sont bornés (par n) et comme $T_i \wedge n \leqslant S_{i+1} \wedge n$, le théorème d'arrêt 25.2 implique que l'espérance de chacun des termes de la somme à droite de (26.3) est positive, de sorte qu'on obtient

$$(b-a)E(U_n) \leqslant E(Y_n). \qquad \blacksquare$$

Exercices

1. Soit $Y_n \in L^2$ des variables aléatoires telles que $\lim_{n\to\infty} E(Y_n^2) = 0$. Soit $(\mathcal{F}_k)_{k\geqslant 0}$ une suite croissante de sous-tribus et posons $X_k^n = E(Y_n \mid \mathcal{F}_k)$. Montrer que $\lim_{n\to\infty} E(\sup_k (X_k^n)^2) = 0$.
2. Soit X et Y des v.a. positives vérifiant

$$\alpha P(X \geqslant \alpha) \leqslant E(Y 1_{\{X \geqslant \alpha\}}),$$

pour tout $\alpha > 0$. Montrer que si $p > 1$ et $\frac{1}{p} + \frac{1}{q} = 1$, on a $E(X^p) \leqslant E(qX^{p-1}Y)$.
3. Soit X et Y comme dans l'exercice 2. Supposons que $\|X\|_p < \infty$ et que $\|Y\|_p < \infty$. Montrer que $\|X\|_p \leqslant q\|Y\|_p$. (*Indication :* Utiliser l'exercice 2 et l'inégalité de Hölder.)
4. Établir le résultat de l'exercice 3 sans l'hypothèse $\|X\|_p < \infty$.
5.* Utiliser l'exercice 4 pour démontrer le théorème 26.3.

Chapitre 27

Les théorèmes de convergence de martingales

Dans le chapitre 17 nous avons étudié des théorèmes de convergence du type suivant : une certaine forme de convergence, plus éventuellement telle ou telle condition supplémentaire, entraîne une autre forme de convergence. Pour les martingales, on a un phénomène un peu surprenant : aucune forme de convergence n'est *a priori* supposée – seulement une certaine structure – et on obtient la convergence. Cela rend ce genre de théorème singulier en analyse, la seule situation un peu analogue ne se rencontrant qu'en théorie ergodique.

Théorème 27.1. (Théorème de convergence des sous-martingales.) *Soit* $(X_n)_{n\geqslant 0}$ *une sous-martingale telle que* $\sup_n E(X_n^+) < \infty$. *Alors la suite* X_n *converge p.s. vers une limite* X *à valeurs réelles. De plus* X *est dans* L^1. Attention : nous n'affirmons pas que X_n converge vers X dans L^1 ; ceci est *faux* en général.

Preuve. Pour $a < b$ notons $U_n(a,b)$ le nombre de montées de la suite (X_k) de a à b, entre les instants 0 et n, ainsi que nous l'avons défini en (26.2). La suite $U_n(a,b)$ est évidemment croissante en n, et elle converge donc (pour tout ω) vers une limite $U(a,b)$, à valeurs dans $\mathbb{N} \cup \{+\infty\}$. En utilisant le théorème de convergence monotone, le théorème 26.4 et le fait que $(x-a)^+ \leqslant x^+ + |a|$ pour tous réels a, x, on obtient

$$\begin{aligned} E(U(a,b)) &= \lim_{n\to\infty} E(U_n(a,b)) \\ &\leqslant \frac{1}{b-a} \sup_n E((X_n - a)^+) \\ &\leqslant \frac{1}{b-a}\left(\sup_n E(X_n^+) + |a|\right) \leqslant \frac{c}{b-a} < \infty \end{aligned}$$

pour une certaine constante c finie. Comme $E(U(a,b)) < \infty$, on a aussi $P(U(a,b) < \infty) = 1$. C'est-à-dire que la suite X_n « monte » de a à b presque sûrement un nombre fini de fois le long de *tous les entiers*. Il s'ensuit que si

$$\Lambda_{a,b} = \{\limsup_{n\to\infty} X_n \geqslant b\ ;\ \liminf_{n\to\infty} X_n \leqslant a\},$$

alors $P(\Lambda_{a,b}) = 0$. Si $\Lambda = \bigcup_{\substack{a<b \\ a,b\in\mathbb{Q}}} \Lambda_{a,b}$ (où $\mathbb{Q}$ désigne l'ensemble des rationnels), et comme la famille des couples de rationnels est dénombrable, on a aussi $P(\Lambda) = 0$. Mais

$$\Lambda = \{\limsup_n X_n > \liminf_n X_n\}.$$

et on en conclut que X_n converge p.s. vers une limite X.

Noter qu'il est pour le moment possible *a priori* que X prenne les valeurs $+\infty$ et $-\infty$, et nous devons montrer que ces deux éventualités sont en fait impossibles. Comme (X_n) est une sous-martingale on a $E(X_n) \geqslant E(X_0)$, et donc

$$\begin{aligned} E(|X_n|) &= E(X_n^+) + E(X_n^-) \\ &= 2E(X_n^+) - E(X_n) \\ &\leqslant 2E(X_n^+) - E(X_0), \end{aligned} \tag{27.1}$$

donc

$$E(|X|) = E(\lim_n |X_n|) \leqslant \liminf_{n\to\infty} E(|X_n|) \leqslant 2\sup_n E(X_n^+) - E(X_0) < \infty,$$

par le lemme de Fatou et (27.1) combiné à l'hypothèse que $\sup_n E(X_n^+) < \infty$. Donc X est intégrable (dans L^1), et donc *a fortiori* p.s. fini. ■

Corollaire 27.1. *Si (X_n) est une surmartingale positive, ou une martingale bornée inférieurement ou bornée supérieurement (par une constante), la suite (X_n) converge p.s. vers une limite X appartenant à L^1.*

Preuve. Si (X_n) est une surmartingale positive, alors $(-X_n)$ est une sous-martingale négative et on peut appliquer le théorème 27.1.

Si (X_n) est une martingale bornée inférieurement, disons par a, alors $(X_n - a)$ est une martingale positive et il suffit d'appliquer ce qui précède. Si enfin (X_n) est une martingale bornée supérieurement, il suffit de considérer $(-X_n)$. ■

Le théorème 27.1 donne la convergence presque sûre vers une v.a. X qui est dans L^1. Mais il ne donne pas la convergence dans L^1. Pour l'obtenir il nous faut une hypothèse légèrement plus forte, pour laquelle nous avons besoin du concept d'intégrabilité uniforme.

Définition 27.1. *Un sous-ensemble $\mathcal{H}$ de* L^1 *est dit* uniformément intégrable *si*

$$\lim_{c\to\infty} \sup_{\mathrm{X}\in\mathcal{H}} \mathrm{E}(1_{\{|\mathrm{X}|\geqslant c\}}|\mathrm{X}|) = 0.$$

Voici deux conditions assurant l'uniforme intégrabilité.

Théorème 27.2. *Soit $\mathcal{H}$ une famille de v.a. réelles.*

(a) *Si* $\sup_{\mathrm{X}\in\mathcal{H}} \mathrm{E}(|\mathrm{X}|^p) < \infty$ *pour un $p > 1$, la famille $\mathcal{H}$ est uniformément intégrable.*

(b) *S'il existe une v.a.* Y *intégrable telle que* $|\mathrm{X}| \leqslant \mathrm{Y}$ *p.s. pour tout* $\mathrm{X} \in \mathcal{H}$, *la famille $\mathcal{H}$ est uniformément intégrable.*

Preuve. (a) Soit $k = \sup_{\mathrm{X}\in\mathcal{H}_n} \mathrm{E}(|\mathrm{X}|^p)$. Si $x \geqslant c > 0$, on a $x^{1-p} \leqslant c^{1-p}$, et en multipliant par x^p on obtient $x \leqslant c^{1-p}x^p$. Par suite

$$\mathrm{E}\left(|\mathrm{X}|1_{\{|\mathrm{X}|>c\}}\right) \leqslant c^{1-p}\mathrm{E}\left(|\mathrm{X}|^p 1_{\{|\mathrm{X}|>c\}}\right) \leqslant \frac{k}{c^{p-1}},$$

donc $\lim_{c\to\infty} \sup_{\mathrm{X}\in\mathcal{H}} \mathrm{E}(|\mathrm{X}|1_{\{|\mathrm{X}|>c\}}) \leqslant \lim_{c\to\infty} \frac{k}{c^{p-1}} = 0$.

(b) Comme $|\mathrm{X}| \leqslant \mathrm{Y}$ a.s. pour tout $\mathrm{X} \in \mathcal{H}$, on a

$$|\mathrm{X}|1_{\{|\mathrm{X}|>c\}} \leqslant \mathrm{Y}1_{\{\mathrm{Y}>c\}}.$$

Mais $\lim_{c\to\infty} \mathrm{Y}1_{\{\mathrm{Y}>c\}} = 0$ p.s. ; donc le théorème de convergence dominée de Lebesgue entraîne que

$$\begin{aligned}\lim_{c\to\infty} \sup_{\mathrm{X}\in\mathcal{H}} \mathrm{E}(|\mathrm{X}|1_{\{|\mathrm{X}|>c\}}) &\leqslant \lim_{c\to\infty} \mathrm{E}(\mathrm{Y}1_{\{\mathrm{Y}>c\}}) \\ &= \mathrm{E}(\lim_{c\to\infty} \mathrm{Y}1_{\{\mathrm{Y}>c\}}) = 0.\end{aligned}$$

■

Pour des résultats complémentaires sur l'intégrabilité uniforme, nous recommandons [18, p. 16-21] ou [19].

Théorème 27.3. (Théorème de convergence des martingales.)

(a) *Soit* $(\mathrm{X}_n)_{n\geqslant 1}$ *une martingale. Si la suite de v.a.* $(\mathrm{X}_n)_{n\geqslant 1}$ *est uniformément intégrable, elle converge p.s. et dans* L^1 *vers une v.a.* X *intégrable, et de plus on a* $\mathrm{X}_n = \mathrm{E}(\mathrm{X} \mid \mathcal{F}_n)$ *pour tout n.*

(b) *Inversement, soit* $\mathrm{Y} \in \mathrm{L}^1$ *et considérons la martingale* $\mathrm{X}_n = \mathrm{E}(\mathrm{Y} \mid \mathcal{F}_n)$. *La suite* $(\mathrm{X}_n)_{n\geqslant 1}$ *est alors uniformément intégrable.*

En utilisant la terminologie de la définition 24.2, nous avons donc que (X_n) est une martingale *fermée* si et seulement si elle est uniformément intégrable.

Preuve. (a) Comme $(X_n)_{n\geqslant 1}$ est uniformément intégrable, pour tout $\varepsilon > 0$ il existe c tel que $\sup_n E(|X_n|1_{\{|X_n|>c\}}) \leqslant \varepsilon$. Donc

$$E(|X_n|) = E(|X_n|1_{\{|X_n|>c\}}) + E(|X_n|1_{\{|X_n|\leqslant c\}}) \leqslant \varepsilon + c.$$

Par suite $(X_n)_{n\geqslant 1}$ est « bornée dans L^1 ». En particulier $\sup_n E(X_n^+) < \infty$ et le théorème 27.1 implique

$$\lim_{n\to\infty} X_n = X \text{ existe p.s. et } X \text{ est dans } L^1.$$

Pour montrer la convergence $X_n \to X$ dans L^1, posons

$$f_c(x) = \begin{cases} c & \text{si } x > c, \\ x & \text{si } |x| \leqslant c, \\ -c & \text{si } x < -c. \end{cases}$$

La fonction f est lipschitzienne bornée. À cause de l'uniforme intégrabilité, pour tout $\varepsilon > 0$ il existe c tel que

$$E\left(|f_c(X_n) - X_n|\right) < \frac{\varepsilon}{3} \quad \text{pour tout } n\ ; \tag{27.2}$$

$$E\left(|f_c(X) - X|\right) < \frac{\varepsilon}{3}. \tag{27.3}$$

Comme $\lim X_n \to X$ p.s. on a $f_c(X_n) \to f_c(X)$ p.s., et le théorème de convergence dominée de Lebesgue implique que pour tout $n \geqslant N$, avec N assez grand,

$$E(|f_c(X_n) - f_c(X)|) < \frac{\varepsilon}{3}. \tag{27.4}$$

Donc en utilisant (27.2), (27.3) et (27.4), on obtient

$$E(|X_n - X|) < \varepsilon \quad \text{pour } n \geqslant N.$$

Donc $X_n \to X$ dans L^1. Il reste à montrer que $E(X \mid \mathcal{F}_n) = X_n$. Soit $\Lambda \in \mathcal{F}_n$ et $m \geqslant n$. Alors

$$E(X_n 1_\Lambda) = E(X_m 1_\Lambda)$$

par la propriété de martingale. Mais

$$\begin{aligned} |E(X_m 1_\Lambda) - E(X 1_\Lambda)| &\leqslant E(|X_m - X| 1_\Lambda) \\ &\leqslant E(|X_m - X|) \end{aligned}$$

qui tend vers 0 quand $m \to \infty$. Donc $E(X_n 1_\Lambda) = E(X 1_\Lambda)$ et on en conclut que $E(X \mid \mathcal{F}_n) = X_n$ p.s.

(b) On sait déjà que $(X_n)_{n\geqslant 0}$ est une martingale. Si $c > 0$ on a

$$X_n 1_{\{|X_n|>c\}} = E(Y 1_{\{|X_n|>c\}} \mid \mathcal{F}_n),$$

puisque $\{|X_n| > c\} \in \mathcal{F}_n$. Donc pour tout $d > 0$ on a

$$\begin{aligned} E(|X_n| 1_{\{|X_n|>c\}}) &\leqslant E(|Y| 1_{\{|X_n|>c\}}) \\ &\leqslant E(|Y| 1_{\{|Y|>d\}}) + d\, P(|X_n| > c) \\ &\leqslant E(|Y| 1_{\{|Y|>d\}}) + \frac{d}{c}\, E(|X_n|). \end{aligned} \tag{27.5}$$

Prenons $\varepsilon > 0$. Choisissons d tel que le premier terme dans (27.5) soit plus petit que $\varepsilon/2$, puis c tel que le second terme soit plus petit que $\varepsilon/2$: on a alors $E(|X_n| 1_{\{|X_n|>c\}}) \leqslant \varepsilon$ pour tout n, et nous avons terminé. ■

La propriété de martingale est $E(X_m \mid \mathcal{F}_n) = X_n$ p.s. pour $m \geqslant n$. Jusqu'à présent nous avons supposé que m et n étaient des entiers positifs, mais on peut aussi prendre des entiers négatifs, i.e. considérer que l'ensemble d'indices est $-\mathbb{N}$, l'ensemble des entiers négatifs. Dans ce cas si $|m| > |n|$, mais m et n sont des entiers négatifs, on a $m < n$. De manière à minimiser les risques de confusion, on supposera toujours que m et n sont des entiers *positifs* et on écrira X_{-n}. Ainsi, une *martingale inverse* sera un processus $X = (X_{-n})_{n\in\mathbb{N}}$ vérifiant $X_{-n} \in L^1$ et X_{-n} est $\mathcal{F}_{-n}$-mesurable pour tout n, et

$$E(X_{-n} \mid \mathcal{F}_{-m}) = X_{-m} \qquad \text{p.s.}, \tag{27.6}$$

si $0 \leqslant n < m$. Ci-dessus les sous-tribus vérifient $\mathcal{F}_{-m} \subset \mathcal{F}_{-n}$ si $0 \leqslant n < m$.

Théorème 27.4. (Théorème de convergence des martingales inverses.) *Soit une martingale inverse $(X_{-n}, \mathcal{F}_{-n})_{n\geqslant 0}$, et soit $\mathcal{F}_{-\infty} = \bigcap_{n=0}^{\infty} \mathcal{F}_{-n}$. Alors la suite (X_{-n}) converge p.s. et dans L^1 quand $n \to \infty$ vers une limite X qui est finie et intégrable.*

Preuve. Si $a < b$ notons $U_{-n}(a, b)$ le nombre de montées de a à b du processus $(X_{-n})_{n\geqslant 0}$ entre les instants $-n$ et 0. La suite $U_{-n}(a, b)$ croît vers une limite $U(a, b)$ à valeurs dans $\mathbb{N} \cup \{+\infty\}$ quand $n \to \infty$. Exactement comme dans la preuve du théorème 27.1 on peut écrire

$$\begin{aligned} E(U(a,b)) &= \lim_{n\to\infty} E(U_{-n}) \\ &\leqslant \frac{1}{b-a}\, E((X_0 - a)^+) < \infty, \end{aligned}$$

donc $P(U(a,b) < \infty) = 1$. À nouveau le même argument que dans le théorème 27.1 montre que X_{-n} converge p.s. vers une limite X, *a priori* à valeurs dans $[-\infty, +\infty]$.

La fonction $\varphi(x) = x^+ = x \vee 0$ est convexe, donc l'inégalité de Jensen (théorème 23.9) et (27.6) impliquent que $X_{-n}^+ \leqslant E(X_0^+ \,|\, \mathcal{F}_{-n})$, donc $E(X_{-n}^+) \leqslant E(X_0^+)$. En appliquant le lemme de Fatou et le fait que $X_{-n}^+ \geqslant 0$ et $X_{-n}^+ \to X^+$ p.s., on arrive à

$$E(X^+) \leqslant \liminf_n E(X_{-n}^+) \leqslant E(X_0^+) < \infty,$$

de sorte que $X^+ \in L^1$. Le même argument appliqué à la martingale $(-X_{-n})$ montre que $X^- \in L^1$, donc $X \in L^1$.

Il reste à montrer la convergence dans L^1. Pour cela, on remarque d'abord que dans la preuve du théorème 27.3 nous avons en particulier montré que si $X_{-n} \to X$ p.s., si $X \in L^1$, et si la suite (X_{-n}) est uniformément intégrable, alors on a aussi $X_{-n} \to X$ dans L^1. La partie (b) de ce théorème montre aussi que la famille des v.a. $E(X_0 \,|\, \mathcal{G})$, lorsque $\mathcal{G}$ décrit la classe de toutes les sous-tribus de $\mathcal{F}$, est uniformément intégrable. Comme $X_{-n} = E(X_0 \,|\, \mathcal{F}_{-n})$, on déduit de tous ces rappels la convergence dans L^1 souhaitée. ■

Comme application du théorème 27.4 nous prouvons la *loi des grands nombres de Kolmogorov.*

Théorème 27.5. (Loi des grands nombres.) *Soit $(X_n)_{n\geqslant 1}$ une suite i.i.d. de variables aléatoires réelles intégrables. On a alors*

$$\frac{X_1 + \cdots + X_n}{n} \to E(X_1) \qquad \textit{p.s. et dans}\, L^1.$$

Preuve. Soit $S_n = X_1 + \cdots + X_n$ et $\mathcal{F}_{-n} = \sigma(S_n, S_{n+1}, S_{n+2}, \ldots)$. On a $\mathcal{F}_{-n} \subset \mathcal{F}_{-m}$ si $n \geqslant m$, et le processus

$$M_{-n} = E(X_1 \,|\, \mathcal{F}_{-n})$$

est une martingale inverse. Noter que $E(M_{-n}) = E(X_1)$ pour chaque n. En utilisant le fait que les X_n sont i.i.d., un argument de symétrie permet de vérifier aisément que pour $1 \leqslant j \leqslant n$ on a

$$E(X_1 \,|\, \mathcal{F}_{-n}) = E(X_j \,|\, \mathcal{F}_{-n}) \qquad \text{p.s.} \tag{27.7}$$

(voir l'exercice 17 du chapitre 23). Donc

$$M_{-n} = E(X_1 \,|\, \mathcal{F}_{-n}) = E(X_2 \,|\, \mathcal{F}_{-n}) = \cdots = E(X_n \,|\, \mathcal{F}_{-n}),$$

et par suite

$$\mathrm{M}_{-n} = \frac{1}{n}\sum_{j=1}^{n}\mathrm{E}(\mathrm{X}_j \,|\, \mathcal{F}_{-n}) = \mathrm{E}\left(\frac{\mathrm{S}_n}{n} \,|\, \mathcal{F}_{-n}\right) = \frac{\mathrm{S}_n}{n} \quad \text{p.s.}$$

Le théorème 27.4 implique alors que $\frac{\mathrm{S}_n}{n}$ converge p.s. et dans L^1 vers une limite intégrable X, qui vérifie nécessairement $\mathrm{E}(\mathrm{X}) = \mathrm{E}(\mathrm{X}_1)$. De plus X est mesurable par rapport à la tribu asymptotique, donc la loi zéro-un de Kolmogorov (théorème 10.6) entraîne que X est p.s. égale à une constante, qui est donc nécessairement son espérance. ■

Le théorème 27.5 a été publié pour la première fois en 1933 [17], sans l'aide de la théorie des martingales ; cette théorie a été développée par J. L. Doob bien des années plus tard.

Voici une autre application du théorème de convergence des martingales, due aussi à Kolmogorov.

Théorème 27.6. (Kolmogorov.) *Soit* $(\mathrm{Y}_n)_{n\geqslant 1}$ *des v.a. réelles indépendantes, centrées et de carré intégrable. Supposons que* $\sum_{n=1}^{\infty}\mathrm{E}(\mathrm{Y}_n^2) < \infty$, *et soit* $\mathrm{S}_n = \sum_{j=1}^{n}\mathrm{Y}_j$. *La suite* S_n *converge alors p.s. vers une limite finie, notée* $\sum_{j=1}^{\infty}\mathrm{Y}_j$.

Preuve. Soit $\mathcal{F}_n = \sigma(\mathrm{Y}_1, \ldots, \mathrm{Y}_n)$. On a vu au chapitre 24 que $(\mathrm{S}_n)_{n\geqslant 1}$ est une $\mathcal{F}_n$-martingale. De plus, $\sup_n \mathrm{E}(\mathrm{S}_n^+) \leqslant \sup_n(\mathrm{E}(\mathrm{S}_n^2) + 1) \leqslant \sum_{n=1}^{\infty}\mathrm{E}(\mathrm{Y}_n^2) + 1 < \infty$. Le résultat découle alors du théorème 27.1. ■

Nous venons de voir ci-dessus un exemple de martingale, à savoir la suite des sommes partielles (S_n) de v.a. indépendantes et centrées. Si ces v.a. indépendantes sont de plus de même loi et de variance finie, disons σ^2, le théorème-limite central nous dit que la suite $\frac{\mathrm{S}_n}{\sqrt{n}}$ converge en loi vers une v.a. $\mathrm{N}(0, \sigma^2)$. La théorie des martingales nous permet d'affaiblir à la fois l'hypothèse d'indépendance, et l'hypothèse que ces v.a. ont même loi. En ce sens, elle permet de mieux cerner ce qui est réellement indispensable pour avoir un théorème-limite central. Comme dans tout ce chapitre, nous supposons donnée la suite croissante $(\mathcal{F}_n)_{n\geqslant 0}$ de sous-tribus de $\mathcal{F}$, sur l'espace $(\Omega, \mathcal{F}, \mathrm{P})$.

Théorème 27.7. (Théorème-limite central pour les martingales.) *Soit* $(\mathrm{X}_n)_{n\geqslant 1}$ *une suite de v.a. réelles telles que pour tout* $n \geqslant 1$ *:*

(i) $\mathrm{E}(\mathrm{X}_n \,|\, \mathcal{F}_{n-1}) = 0$

(ii) $\mathrm{E}(\mathrm{X}_n^2 \,|\, \mathcal{F}_{n-1}) = 1$

(iii) $\mathrm{E}(|\mathrm{X}_n|^3 \,|\, \mathcal{F}_{n-1}) \leqslant \mathrm{K} < \infty$.

Soit $S_n = \sum_{i=1}^n X_i$ *pour* $n \geqslant 1$. *La suite* $\frac{1}{\sqrt{n}} S_n$ *converge alors en loi vers une v.a.* $N(0, 1)$.

Preuve. Nous allons utiliser les fonctions caractéristiques pour prouver ce théorème. Pour $u \in \mathbb{R}$, rappelons que la fonction caractéristique d'une v.a. réelle X est $\varphi_X(u) = E(e^{iuX})$. On définit la « fonction caractéristique conditionnelle » de la v.a. $X_j/\sqrt{n}$ comme étant la v.a.

$$\varphi_{n,j}(u) = E\left(e^{iu\frac{1}{\sqrt{n}}X_j} \mid \mathcal{F}_{j-1}\right).$$

On peut faire un développement de Taylor au voisinage de 0 :

$$e^{iu\frac{1}{\sqrt{n}}X_j} = 1 + iu\frac{1}{\sqrt{n}}X_j - \frac{u^2}{2n}X_j^2 - \frac{iu^3}{6n^{\frac{3}{2}}}\overline{X}_j^3 \tag{27.8}$$

où $\overline{X}_j$ est un nombre (aléatoire) situé entre 0 et X_j. En prenant l'espérance conditionnelle des deux membres de (27.8), on obtient

$$\varphi_{n,j}(u) = 1 + iu\frac{1}{\sqrt{n}}E(X_j \mid \mathcal{F}_{j-1}) - \frac{u^2}{2n}E(X_j^2 \mid \mathcal{F}_{j-1}) - \frac{iu^3}{6n^{\frac{3}{2}}}E(\overline{X}_j^3 \mid \mathcal{F}_{j-1})$$

et les hypothèses (i) et (ii) donnent

$$\varphi_{n,j}(u) - 1 - \frac{u^2}{2n} = \frac{u^3}{6n^{\frac{3}{2}}}E(\overline{X}_j^3 \mid \mathcal{F}_{j-1}). \tag{27.9}$$

Rappelons que $S_p = \sum_{j=1}^p X_j$, donc pour $1 \leqslant p \leqslant n$ on a

$$\begin{aligned} E\left(e^{iu\frac{1}{\sqrt{n}}S_p}\right) &= E\left(e^{iu\frac{1}{\sqrt{n}}S_{p-1}}e^{iu\frac{1}{\sqrt{n}}X_p}\right) \\ &= E\left(e^{iu\frac{1}{\sqrt{n}}S_{p-1}}E\left(e^{iu\frac{1}{\sqrt{n}}X_p} \mid \mathcal{F}_{p-1}\right)\right) \\ &= E\left(e^{iu\frac{1}{\sqrt{n}}S_{p-1}}\varphi_{n,p}(u)\right). \end{aligned} \tag{27.10}$$

En utilisant (27.10) et (27.9) on arrive à

$$E\left(e^{i\frac{u}{\sqrt{n}}S_p}\right) = E\left(e^{i\frac{u}{\sqrt{n}}S_{p-1}}\left(1 - \frac{u^2}{2n} - \frac{iu^3}{6n^{\frac{3}{2}}}\overline{X}_j^3\right)\right)$$

et donc

$$E\left(e^{i\frac{u}{\sqrt{n}}S_p} - \left(1 - \frac{u^2}{2n}\right)e^{i\frac{u}{\sqrt{n}}S_{p-1}}\right) = E\left(e^{i\frac{u}{\sqrt{n}}S_{p-1}}\frac{iu^3}{6n^{\frac{3}{2}}}\overline{X}_j^3\right). \tag{27.11}$$

En prenant le module dans chaque membre de (27.11) et en utilisant le fait que $|\overline{\mathrm{X}}_j| \leqslant |\mathrm{X}_j|$ et l'hypothèse (iii), il vient

$$\begin{aligned}\left|\mathrm{E}\left(e^{iu\frac{1}{\sqrt{n}}\mathrm{S}_p} - \left(1-\frac{u^2}{2n}\right)e^{iu\frac{1}{\sqrt{n}}\mathrm{S}_{p-1}}\right)\right| &\leqslant \mathrm{E}\left(\left|e^{iu\frac{1}{\sqrt{n}}\mathrm{S}_{p-1}}\right|\frac{|u|^3}{6n^{\frac{3}{2}}}\mathrm{E}\big(|\mathrm{X}_j|^3 \mid \mathcal{F}_{j-1}\big)\right)\\ &\leqslant \mathrm{K}\frac{|u|^3}{6n^{\frac{3}{2}}}. \qquad (27.12)\end{aligned}$$

Fixons $u \in \mathbb{R}$. Lorsque $n \to \infty$, pour n assez grand on a $n \geqslant \frac{u^2}{2}$ et donc $0 \leqslant 1 - \frac{u^2}{2n} \leqslant 1$. En multipliant (27.12) par $(1-\frac{u^2}{2n})^{n-p}$ on arrive, pour n assez grand, à

$$\left|\left(1-\frac{u^2}{2n}\right)^{n-p}\mathrm{E}\left(e^{iu\frac{1}{\sqrt{n}}\mathrm{S}_p}\right) - \left(1-\frac{u^2}{2n}\right)^{n-p+1}\mathrm{E}\left(e^{iu\frac{1}{\sqrt{n}}\mathrm{S}_{p-1}}\right)\right| \leqslant \mathrm{K}\frac{|u|^3}{6n^{\frac{3}{2}}}. \qquad (27.13)$$

Noter que (27.13) est vraie également pour chaque n fixé, pour un certain K_n : en prenant K assez grand on peut donc supposer (27.13) vraie pour tout n.

Finalement en additionnant et en utilisant le caractère télescopique des différents termes, on voit que

$$\begin{aligned}\mathrm{E}\left(e^{iu\frac{1}{\sqrt{n}}\mathrm{S}_n}\right) - \left(1-\frac{u^2}{2n}\right)^n = \sum_{p=1}^{n}\Bigg\{&\left(1-\frac{u^2}{2n}\right)^{n-p}\mathrm{E}\left(e^{iu\frac{1}{\sqrt{n}}\mathrm{S}_p}\right)\\ &- \left(1-\frac{u^2}{2n}\right)^{n-(p-1)}\mathrm{E}\left(e^{iu\frac{1}{\sqrt{n}}\mathrm{S}_{p-1}}\right)\Bigg\}\end{aligned}$$

de sorte que par l'inégalité triangulaire et (27.13) on voit que

$$\left|\mathrm{E}\left(e^{iu\frac{1}{\sqrt{n}}\mathrm{S}_n}\right) - \left(1-\frac{u^2}{2n}\right)^n\right| \leqslant n\frac{\mathrm{K}|u|^3}{6n^{\frac{3}{2}}} = \mathrm{K}\frac{|u|^3}{6\sqrt{n}}. \qquad (27.14)$$

Comme le membre de droite de (27.14) tend vers 0 et que

$$\lim_{n\to\infty}\left(1-\frac{u^2}{2n}\right)^n = e^{-\frac{u^2}{2}},$$

on obtient

$$\lim_{n\to\infty}\mathrm{E}(e^{iu\frac{\mathrm{S}_n}{\sqrt{n}}}) = e^{-\frac{u^2}{2}}.$$

Le théorème de Lévy 19.1 donne alors le résultat. ∎

Remarque. Si S_n est la martingale du théorème 22.4, on sait que le théorème de convergence p.s. ne peut pas être vérifié : en effet, si on avait $\lim_{n\to\infty} S_n = S$ p.s. avec $S \in L^1$, on aurait aussi $\lim_{n\to\infty} \frac{S_n}{\sqrt{n}} = 0$ p.s., et la convergence en loi de $\frac{S_n}{\sqrt{n}}$ vers une v.a. normale non dégénérée serait impossible. Ce qui rend la convergence p.s. de S_n impossible est le comportement des variances conditionnelles des accroissements X_n, i.e. l'hypothèse (ii) du théorème 22.4.

Nous terminons notre exposé sur les martingales par un exemple tiré de l'analyse. Cet exemple montre la souplesse d'utilisation et la variété des applications de la théorie des martingales : nous utilisons le résultat « probabiliste » de convergence des martingales pour montrer de manière simple une résultat « analytique » de convergence de fonctions.

Exemple ([13]). Soit f une fonction dans $L^p[0,1]$ pour la mesure de Lebesgue sur $[0,1]$. Définissons les *fonctions de Rademacher* sur $[0,1]$ comme suit. Posons $R_0(x) = 1$ pour $0 \leqslant x \leqslant 1$. Si $n \geqslant 1$, nous posons pour $0 \leqslant x \leqslant 1$:

$$R_n(x) = \begin{cases} 1 & \text{si } \frac{2j-1}{2^n} \leqslant x < \frac{2j}{2^n}, \quad \text{pour un } j \text{ dans } \{1,\ldots,2^n\} \\ -1 & \text{sinon.} \end{cases}$$

Notons P la probabilité sur $[0,1]$ égale à la mesure de Lebesgue. $\mathcal{F}$ désigne la tribu borélienne. On a alors

$$E(R_n) = \int_0^1 R_n(x)dx = 0$$

et

$$\mathrm{Var}\,(R_n) = E(R_n^2) = \int_0^1 R_n(x)^2 dx = 1.$$

Finalement on observe que R_n et R_m (qui peuvent être ici considérées comme des variables aléatoires) sont indépendantes si $n \neq m$. (voir l'exercice 8.)

Ensuite, nous définissons les *fonctions de Haar* comme suit :

$$H_0(x) = R_0(x),$$
$$H_1(x) = R_1(x);$$

Pour $n \geqslant 2$, écrivons $n = 1 + 2 + \cdots + 2^{r-2} + \lambda = 2^{r-1} - 1 + \lambda$, avec $r \geqslant 2$ et $1 \leqslant \lambda \leqslant 2^{r-1}$; posons alors

$$\mathrm{H}_n(x) = \begin{cases} \sqrt{2^{r-1}}\mathrm{R}_n(x) & \text{si } \frac{2\lambda-2}{2^r} \leqslant x < \frac{2\lambda}{2^r}, \\ 0 & \text{sinon.} \end{cases}$$

Ensuite, on note $\mathcal{F}_n = \sigma(\mathrm{H}_0, \mathrm{H}_1, \ldots, \mathrm{H}_n)$ la plus petite tribu rendant les fonctions $\mathrm{H}_0, \ldots, \mathrm{H}_n$ toutes mesurables. On vérifie aisément que

$$\int_\Lambda \mathrm{H}_{n+1}(x)dx = 0 \qquad \text{si } \Lambda \in \mathcal{F}_n; \tag{27.15}$$

(voir l'exercice 9). On a de plus

$$\int_0^1 \mathrm{H}_n(x)dx = 0,$$
$$\int_0^1 \mathrm{H}_n(x)^2dx = 1.$$

On a alors le théorème suivant :

Théorème 27.8. *Soit* (H_n) *les fonctions de Haar sur* $[0,1]$, *et soit* $f \in \mathrm{L}^p[0,1]$ *pour un* $p \geqslant 1$. *Posons*

$$\alpha_r = \int_0^1 \mathrm{H}_r(x)f(x)dx,$$
$$\mathrm{S}_n(x) = \sum_{r=0}^n \alpha_r \mathrm{H}_r(x). \tag{27.16}$$

Alors S_n *converge vers* f *presque partout (pour la mesure de Lebesgue). De plus si* $\mathrm{S}^*(x) = \sup_n |\mathrm{S}_n(x)|$, *on a*

$$\int_0^1 \mathrm{S}^*(x)^p dx \leqslant \left(\frac{p}{p-1}\right)^p \int_0^1 |f(x)|^p dx.$$

Preuve. Montrons d'abord que la suite (S_n) est une martingale. On a

$$\begin{aligned} \mathrm{E}(\mathrm{S}_{n+1} \,|\, \mathcal{F}_n) &= \mathrm{S}_n + \mathrm{E}(\alpha_{n+1}\mathrm{H}_{n+1} \,|\, \mathcal{F}_n) \\ &= \mathrm{S}_n + \alpha_{n+1}\mathrm{E}(\mathrm{H}_{n+1} \,|\, \mathcal{F}_n) \\ &= \mathrm{S}_n, \end{aligned}$$

où on a utilisé (27.15). On a même :

$$S_n = E(f \mid \mathcal{F}_n), \tag{27.17}$$

qui est le résultat clé. C'est l'endroit où nous avons besoin de la forme (27.16) pour les coefficients α_r (voir l'exercice 10.)

Nous montrons ensuite que S_n vérifie $\sup_n E(S_n^+) < \infty$, pour $p > 1$ (c'est l'hypothèse pour pouvoir appliquer le théorème de convergence des martingales 27.1). Nous montrons en fait un peu plus, grâce à l'inégalité de Jensen : comme $\varphi(u) = |u|^p$ est convexe pour $p > 1$, il vient

$$\begin{aligned}\int_0^1 |S_n(x)|^p dx &= E(|E(f \mid \mathcal{F}_n)|^p)\\ &\leqslant E(E(|f|^p \mid \mathcal{F}_n))\\ &= E(|f|^p)\\ &= \int_0^1 |f(x)|^p dx < \infty.\end{aligned}$$

Comme $S_n^+ \leqslant |S_n| \leqslant |S_n|^p + 1$, on a alors

$$\begin{aligned}\sup_n E(S_n^+) &\leqslant \sup_n (E(|S_n|^p) + 1)\\ &\leqslant E(|f|^p) + 1 < \infty.\end{aligned}$$

Le théorème 27.1 montre alors que

$$\lim_{n\to\infty} S_n = f \qquad \text{p.s.},$$

tandis que l'inégalité de Doob (théorème 26.2) donne

$$\begin{aligned}E(S^{*p}) &\leqslant \left(\frac{p}{p-1}\right)^p E(|S_n|^p)\\ &\leqslant \left(\frac{p}{p-1}\right)^p E(|f|^p),\end{aligned}$$

ou de manière équivalente :

$$\int_0^1 (S^*(x))^p dx \leqslant \left(\frac{p}{p-1}\right)^p \int_0^1 |f(x)|^p dx. \qquad \blacksquare$$

Remarquons enfin que des résultats analogues au théorème 27.8 sont vrais pour les séries de Fourier classiques, mais c'est un peu plus difficile à montrer.

Exercices

1. *(Une preuve de la loi zéro-un de Kolmogorov à l'aide des martingales.)* Soit (X_n) une suite de v.a. indépendantes, et $\mathcal{C}_\infty$ la tribu asymptotique correspondante (voir le théorème 10.6). Soit $C \in \mathcal{C}_\infty$. Montrer que $E(1_C \mid \mathcal{F}_n) = P(C)$ pour tout n, où $\mathcal{F}_n = \sigma(X_j; 0 \leqslant j \leqslant n)$. Montrer ensuite que $\lim_{n\to\infty} E(1_C \mid \mathcal{F}_n) = 1_C$ p.s., et en déduire que $P(C) = 0$ ou $P(C) = 1$.
2. Une martingale (X_n) est dite *bornée dans* L^2 si $\sup_n E(X_n^2) < \infty$. Soit (X_n) une martingale telle que chaque X_n soit dans L^2. Montrer que (X_n) est bornée dans L^2 si et seulement si
$$\sum_{n=1}^{\infty} E\big((X_n - X_{n-1})^2\big) < \infty.$$
(*Indication :* cf. l'exercice 12 du chapitre 24.)
3. Soit (X_n) une martingale bornée dans L^2. Montrer que
$$\sup_n E(|X_n|) < \infty$$
et en conclure que X_n converge p.s. et dans L^1 vers une limite X.
4.* Sous les hypothèses de l'exercice 3, montrer que X_n converge vers X dans L^2 également.
5. *(Signes aléatoires.)* Soit $(X_n)_{n\geqslant 1}$ une suite de v.a. indépendantes avec $P(X_n = 1) = P(X_n = -1) = \frac{1}{2}$. Soit $(\alpha_n)_{n\geqslant 1}$ une suite de réels. Montrer que la série $\sum_{n=1}^{\infty} \alpha_n X_n$ est p.s. convergente dès que $\sum_{n=1}^{\infty} \alpha_n^2 < \infty$.
6. Soit $(X_n)_{n\geqslant 1}$ une suite i.i.d. de variables positives vérifiant $E\{X_1\} = 1$. Soit $R_n = \prod_{i=1}^{n} X_i$. Montrer que (R_n) est une martingale.
7. Montrer que si $n \neq m$ les fonctions de Rademacher R_n et R_m sont indépendantes pour $P = \lambda$ (mesure de Lebesgue sur $[0, 1]$).
8. Soit (H_n) les fonctions de Haar, et soit $\Lambda \in \mathcal{F}_n = \sigma(H_0, H_1, \ldots, H_n)$. Montrer que si $n \geqslant 1$,
$$\int_\Lambda H_n(x)dx = 0.$$
9. Soit $f \in L^p[0, 1]$, et S_n défini en (27.16). Montrer que $E(f \mid \mathcal{F}_n) = S_n$. (*Indication :* Montrer que
$$\int_\Lambda f(x)dx = \int_\Lambda S_n(x)dx \qquad \text{si } \Lambda \in \mathcal{F}_n$$

en utilisant que les fonctions de Haar forment un système orthonormal, i.e.

$$\int_0^1 \mathrm{H}_n(x)\mathrm{H}_m(x)dx = 0 \qquad \text{si } n \neq m \text{ et } \int_0^1 \mathrm{H}_n(x)^2 dx = 1.)$$

10. Utiliser le théorème de convergence des martingales pour montrer la « loi zéro-un » suivante : soit $(\mathcal{F}_n)$ une suite croissante de sous-tribus, et $(\mathcal{G}_n)$ une autre suite, décroissante, de sous-tribus, avec $\mathcal{G}_1 \subset \sigma(\cup_{n=1}^{\infty} \mathcal{F}_n)$. Supposons que $\mathcal{F}_n$ et $\mathcal{G}_n$ soient indépendantes pour chaque n. Montrer que si $\Lambda \in \cap_{n=1}^{\infty} \mathcal{G}_n$, alors $\mathrm{P}(\Lambda) = 0$ ou $\mathrm{P}(\Lambda) = 1$.
11. Soit $\mathcal{H}$ un sous-ensemble de L^1. Soit G une fonction positive croissante sur $[0, \infty[$, telle que

$$\lim_{t \to \infty} \frac{\mathrm{G}(t)}{t} = \infty.$$

Supposons en outre que $\sup_{\mathrm{X} \in \mathcal{H}} \mathrm{E}(\mathrm{G}(\mathrm{X})) < \infty$. Montrer que $\mathcal{H}$ est uniformément intégrable. (Cela étend le théorème 27.2(a).)

Chapitre 28

Le théorème de Radon-Nikodym

Soit $(\Omega, \mathcal{F}, \mathrm{P})$ un espace de probabilité. Considérons une v.a. positive X d'espérance égale à 1. Pour tout $\Lambda \in \mathcal{F}$ on pose

$$\mathrm{Q}(\Lambda) = \mathrm{E}(1_\Lambda \mathrm{X}). \tag{28.1}$$

On obtient ainsi une nouvelle probabilité Q : en effet on a d'abord

$$\mathrm{Q}(\Omega) = \mathrm{E}(1_\Omega \mathrm{X}) = \mathrm{E}(\mathrm{X}) = 1,$$

et ensuite si $\mathrm{A}_1, \mathrm{A}_2, \mathrm{A}_3, \dots$ sont dans $\mathcal{F}$ et deux à deux disjoints il vient

$$\begin{aligned}
\mathrm{Q}\left(\bigcup_{i=1}^{\infty} \mathrm{A}_i\right) &= \mathrm{E}(1_{\{\cup_{i=1}^{\infty} \mathrm{A}_i\}} \mathrm{X}) \\
&= \mathrm{E}\left(\sum_{i=1}^{\infty} 1_{\mathrm{A}_i} \mathrm{X}\right) \\
&= \sum_{i=1}^{\infty} \mathrm{E}(1_{\mathrm{A}_i} \mathrm{X}) \\
&= \sum_{i=1}^{\infty} \mathrm{Q}(\mathrm{A}_i)
\end{aligned}$$

et on a la σ-additivité ; ci-dessus, on a pu échanger la somme et l'espérance grâce au théorème de convergence monotone (théorème 9.1(d)).

La probabilité Q jouit de deux propriétés remarquables :

(i) Si $\mathrm{P}(\Lambda) = 0$ alors $\mathrm{Q}(\Lambda) = 0$ (car $\mathrm{Q}(\Lambda) = \mathrm{E}(1_\Lambda \mathrm{X})$ et $1_\Lambda = 0$ p.s., donc aussi $1_\Lambda \mathrm{X} = 0$ p.s.)

(ii) Pour tout $\varepsilon > 0$ il existe $\delta > 0$ tel que si $\Lambda \in \mathcal{F}$ et $\mathrm{P}(\Lambda) < \delta$, alors $\mathrm{Q}(\Lambda) < \varepsilon$.

En fait, (ii) ci-dessus découle de manière générale de (i), comme nous allons l'énoncer formellement, et le montrer.

Théorème 28.1. *Soit* P, Q *deux probabilités telles que* $\mathrm{P}(\Lambda) = 0$ *implique* $\mathrm{Q}(\Lambda) = 0$. *Alors, pour tout* $\varepsilon > 0$ *il existe* $\delta > 0$ *tel que si* $\Lambda \in \mathcal{F}$ *et* $\mathrm{P}(\Lambda) < \delta$, *on ait* $\mathrm{Q}(\Lambda) < \varepsilon$.

Preuve. Supposons le résultat faux. Il existerait alors une suite $\Lambda_n \in \mathcal{F}$ avec $\mathrm{P}(\Lambda_n) < \frac{1}{2^n}$ et $\mathrm{Q}(\Lambda_n) \geqslant \varepsilon$ pour tout n et pour un certain $\varepsilon > 0$. Posons $\Lambda = \limsup_{n\to\infty} \Lambda_n$. Le lemme de Borel-Cantelli (théorème 10.5) entraîne que $\mathrm{P}(\Lambda) = 0$. Puis, en remarquant que le lemme de Fatou admet une version symétrique pour les lim sup (établie en passant dans la preuve du théorème 9.1(f)), on obtient

$$\mathrm{Q}(\Lambda) \geqslant \limsup_{n\to\infty} \mathrm{Q}(\Lambda_n) \geqslant \varepsilon,$$

et on arrive à une contradiction. ■

Cela vaut la peine de remarquer que les conditions (i) et (ii) sont en fait *équivalentes*. En effet, on vient de montrer que (i) implique (ii), et la réciproque est presque évidente : si $\mathrm{P}(\Lambda) = 0$, pour tout $\varepsilon > 0$ on a $\mathrm{P}(\Lambda) < \delta$ pour le δ associé au ε dans (ii), donc $\mathrm{Q}(\Lambda) < \varepsilon$; comme ε est arbitrairement petit on a donc $\mathrm{Q}(\Lambda) = 0$.

Définition 28.1. *Soit* P *une probabilité et* Q *une mesure finie. On dit que* Q est absolument continue par rapport à P *si, chaque fois qu'on a* $\mathrm{P}(\Lambda) = 0$ *pour un* $\Lambda \in \mathcal{F}$, *on a aussi* $\mathrm{Q}(\Lambda) = 0$. *On note cette propriété ainsi :* $\mathrm{Q} \ll \mathrm{P}$.

Exemple. On a vu que la formule (28.1) donne une probabilité Q vérifiant $\mathrm{Q} \ll \mathrm{P}$.

Un tel exemple se présentant naturellement est celui de la probabilité conditionnelle $\mathrm{Q}(\Lambda) = \mathrm{P}(\Lambda \mid \mathrm{A})$, lorsque $\mathrm{P}(\mathrm{A}) > 0$. L'absolue continuité est évidente à vérifier, mais c'est aussi un cas particulier de l'exemple précédent avec la v.a. $\mathrm{X} = \frac{1}{\mathrm{P}(\mathrm{A})} 1_{\mathrm{A}}$.

Le théorème de Radon-Nikodym caractérise en fait toutes les probabilités absolument continues par rapport à P : ce sont exactement celles qui se mettent sous la forme (28.1) : notre exemple initial couvre donc toutes les situations possibles. Nous montrons une version un peu simplifiée, pour les tribus dites « séparables », et notre démonstration suit celle de P. A. Meyer [18].

Définition 28.2. *Une tribu* $\mathcal{F}$ *est dite* séparable *si* $\mathcal{F} = \sigma(\mathrm{A}_1, ..., \mathrm{A}_n, ...)$ *: en d'autres termes,* $\mathcal{F}$ *est engendrée par une suite* dénombrable *d'événements.*

Théorème 28.2. (Radon-Nikodym.) *Soit* $(\Omega, \mathcal{F}, \mathrm{P})$ *un espace de probabilité avec une tribu* $\mathcal{F}$ *qui est séparable. Si* Q *est une mesure finie sur* $\mathcal{F}$ *telle que* $\mathrm{Q} \ll \mathrm{P}$, *il existe alors une v.a. positive intégrable* X *telle que pour tout* $\Lambda \in \mathcal{F}$ *on ait*

$$\mathrm{Q}(\Lambda) = \mathrm{E}(1_\Lambda \mathrm{X}).$$

De plus X *est p.s. unique (i.e. si* X′ *vérifie les mêmes propriétés, alors* $X' = X$ P*-p.s.), et on l'écrit souvent* $X = \frac{dQ}{dP}$.

Preuve. Si $Q(\Omega) = 0$, le résultat est évident avec $X \equiv 0$. On suppose donc que $Q(\Omega) > 0$, et on peut alors « normaliser » Q en posant $\tilde{Q} = \frac{1}{Q(\Omega)}Q$, ce qui revient à supposer que Q est une probabilité. Soit $(A_n)_{n \geqslant 0}$ une suite engendrant la tribu $\mathcal{F}$, et définissons la famille croissante de sous-tribus $(\mathcal{F}_n)_{n \geqslant 1}$ ainsi :

$$\mathcal{F}_n = \sigma(A_0, \ldots, A_n).$$

Il existe alors pour chaque n une partition finie $A_{n,1}, A_{n,2}, \ldots, A_{n,k_n}$ de Ω, telle que les éléments de la tribu $\mathcal{F}_n$ soient exactement les réunions (nécessairement finies!) d'éléments de cette partition. Les $A(n,i)$ sont appelés les *atomes* de $\mathcal{F}_n$. Posons alors

$$X_n(\omega) = \sum_{i=1}^{k_n} \frac{Q(A_{n,i})}{P(A_{n,i})} 1_{A_{n,i}}(\omega), \tag{28.2}$$

avec la convention $\frac{0}{0} = 0$ (comme $Q \ll P$ le numérateur est nul chaque fois que le dénominateur est nul). On va montrer que la suite $(X_n)_{n \geqslant 0}$ est en fait une martingale.

D'abord, le fait que X_n soit $\mathcal{F}_n$-mesurable est évident. Ensuite, soit $m \leqslant n$. Exactement comme dans la preuve du théorème 24.6, pour montrer que $E(X_n \,|\, \mathcal{F}_m) = X_m$ il suffit de montrer que pour tout $\Lambda \in \mathcal{F}_m$ on a

$$\int_\Lambda X_n dP = \int_\Lambda X_m dP. \tag{28.3}$$

On peut écrire

$$\begin{aligned}
\int_\Lambda X_n dP &= \int_\Lambda \sum_{i=1}^{k_n} \frac{Q(A_{n,i})}{P(A_{n,i})} 1_{A_{n,i}} dP \\
&= \int \sum_{i=1}^{k_n} \frac{Q(A_{n,i})}{P(A_{n,i})} 1_{A_{n,i} \cap \Lambda} dP \\
&= \sum_{i=1}^{k_n} \frac{Q(A_{n,i})}{P(A_{n,i})} P(A_{n,i} \cap \Lambda).
\end{aligned}$$

Mais comme $\Lambda \in \mathcal{F}_n$, l'ensemble Λ s'écrit comme la réunion (disjointe) des $A(m,i)$ pour i dans une partie I de $\{1, \ldots, k_n\}$. Donc $\Lambda \cap A_{n,i} = A_{n,i}$ si $i \in I$ et $\Lambda \cap A_{n,i} = \emptyset$ sinon, de sorte que

$$\begin{aligned}\int_\Lambda X_n dP &= \sum_{i\in I} \frac{Q(A_{n,i})}{P(A_{n,i})} P(A_{n,i} \cap \Lambda) \\ &= \sum_{i\in I} Q(A_{n,i}) = Q(\Lambda)\end{aligned}$$

où on a encore une fois utilisé le fait que $Q(A_{n,i}) = 0$ si $P(A_{n,i}) = 0$. Comme $\Lambda \in \mathcal{F}_m$ on a de même $\int_\Lambda X_m dP = Q(\Lambda)$. Donc (28.3) est vérifié, et si on prend de plus $\Lambda = \Omega$ on obtient $\int X_n dP = Q(\Omega) = 1 < \infty$, donc X_n est P-intégrable. Donc $(X_n)_{n\geqslant 1}$ est bien une martingale.

En fait, cette martingale (X_n) est uniformément intégrable. En effet on a

$$\int_{\{X_n > c\}} X_n dP = Q(X_n > c);$$

et par l'inégalité de Markov il vient

$$P(X_n > c) \leqslant \frac{E(X_n)}{c} = \frac{1}{c}.$$

Soit alors $\varepsilon > 0$ et le $\delta > 0$ qui lui est associé comme dans le théorème 28.1. Si $c > 1/\delta$ on a $P(X_n > c) < \delta$, donc $Q(X_n > c) \leqslant \varepsilon$, donc $\int_{\{X_n > c\}} X_n dP \leqslant \varepsilon$: en conséquence, la suite (X_n) est uniformément intégrable. On peut alors appliquer le second théorème de convergence des martingales 27.3, de façon à obtenir que X_n converge p.s. et dans L^1 vers une limite X, et de plus

$$E(X \mid \mathcal{F}_n) = X_n.$$

Si alors on pose $R(\Lambda) = E(1_\Lambda X)$ pour tout $\Lambda \in \mathcal{F}$, on définit une nouvelle probabilité R, qui coïncide avec Q sur chaque tribu $\mathcal{F}_n$ (puisque si $\Lambda \in \mathcal{F}_n$ on a $R(\Lambda) = E(1_\Lambda X) = E(1_\Lambda X_n) = Q(\Lambda)$). Le théorème des classes monotones 6.3 entraîne alors que $R = Q$, puisque $\mathcal{F} = \sigma(\mathcal{F}_n; n \geqslant 1)$. ■

Remarque. On peut utiliser le théorème précédent pour montrer une version plus générale, sans l'hypothèse de séparabilité. Pour une démonstration du théorème 28.3 ci-dessous, nous référons à [27, p. 147-149] par exemple.

Théorème 28.3. (Radon-Nikodym.) *Soit* $(\Omega, \mathcal{F}, P)$ *un espace de probabilité. Si* Q *est une mesure finie sur* $\mathcal{F}$ *telle que* $Q \ll P$, *il existe alors une v.a. positive intégrable* X *telle que pour tout* $\Lambda \in \mathcal{F}$ *on ait*

$$Q(\Lambda) = E(1_\Lambda X).$$

De plus X *est p.s. unique (i.e. si* X′ *vérifie les mêmes propriétés, alors* X′ = X P-*p.s.).*

Le théorème de Radon-Nikodym est directement en rapport avec les espérances conditionnelles : supposons donnés l'espace de probabilité $(\Omega, \mathcal{F}, \mathrm{P})$ et la sous-tribu $\mathcal{G}$ de $\mathcal{F}$. Pour toute v.a. positive intégrable X, la formule $\mathrm{Q}(\Lambda) = \mathrm{E}(\mathrm{X}1_\Lambda)$ pour $\Lambda \in \mathcal{G}$ définit une mesure finie Q sur $(\Omega, \mathcal{G})$, et $\mathrm{P}(\Lambda) = 0$ implique $\mathrm{Q}(\Lambda) = 0$. Donc $\mathrm{Y} = \frac{d\mathrm{Q}}{d\mathrm{P}}$ existe *sur l'espace* $(\Omega, \mathcal{G})$, ce qui implique en particulier que Y est $\mathcal{G}$-mesurable. Par ailleurs on a pour tout $\Lambda \in \mathcal{G}$:

$$\mathrm{E}(\mathrm{Y}1_\Lambda) = \mathrm{Q}(\Lambda) = \mathrm{E}(\mathrm{X}1_\Lambda).$$

Donc Y est une version de $\mathrm{E}(\mathrm{X} \mid \mathcal{G})$. En fait il est possible de démontrer le théorème de Radon-Nikodym à l'aide seulement de la théorie de la mesure (sans martingales et donc sans espérances conditionnelles) ; puis on peut *construire* l'espérance conditionnelle de la manière suggérée ci-dessus : cela évite de recourir à la théorie des espaces de Hilbert (mais la preuve « directe » du théorème de Radon-Nikodym n'est pas particulièrement aisée).

Signalons enfin que si P est une probabilité sur $\mathbb{R}$ admettant la densité f, alors comme $\mathrm{P}(\mathrm{A}) = \int_\mathrm{A} f(x)dx$ on a que la mesure P est absolument continue par rapport à la mesure de Lebesgue m (ici, m est une mesure σ-finie, mais le théorème de Radon-Nikodym « marche » encore dans ce cas), et on peut écrire $f = \frac{d\mathrm{P}}{dm}$.

Exercices

1. Supposons que $\mathrm{Q} \ll \mathrm{P}$ et que $\mathrm{P} \ll \mathrm{Q}$: on dit alors que P et Q sont *équivalentes* et on écrit $\mathrm{Q} \sim \mathrm{P}$. Montrer que $\mathrm{X} = \frac{d\mathrm{Q}}{d\mathrm{P}}$ vérifie $\mathrm{X} > 0$ P-p.s.
2. Supposons que $\mathrm{Q} \sim \mathrm{P}$, et soit $\mathrm{X} = \frac{d\mathrm{Q}}{d\mathrm{P}}$. Montrer que $\frac{1}{\mathrm{X}} = \frac{d\mathrm{P}}{d\mathrm{Q}}$.
3. Soit μ une mesure telle que $\mu = \sum_{n=1}^\infty \alpha_n \mathrm{P}_n$, pour une suite (P_n) de probabilités et une suite (α_n) de nombres strictement positifs vérifiant $\sum_n \alpha_n = 1$. Montrer que μ est une probabilité. Soit une autre suite (Q_n) de probabilités et une autre suite (β_n) de réels positifs avec $\sum_n \beta_n < \infty$. Si on a $\mathrm{Q}_n \ll \mathrm{P}_n$ pour chaque n, montrer que la mesure finie $\nu = \sum_{n=1}^\infty \beta_n \mathrm{Q}_n$ vérifie $\nu \ll \mu$.
4. Soit P et Q deux probabilités et $\mathrm{R} = \frac{\mathrm{P}+\mathrm{Q}}{2}$. Montrer que $\mathrm{P} \ll \mathrm{R}$.

Bibliographie

[1] R. Bass (1995), *Probabilistic Techniques in Analysis*; Springer-Verlag; New York.

[2] H. Bauer (1996), *Probability Theory*; Walter de Gruyter; Berlin.

[3] J. Bernoulli (1713), *Ars Conjectandi*; Thurnisiorum; Bâle.

[4] M. Briane, G. Pagès (1998), *Théorie de l'intégration*; Vuibert; Paris.

[5] G. Cardano (1961), *Liber de ludo aleae*; *The Book on Games of Chance*; Sidney Gould (Traducteur); Holt, Rinehart and Winston.

[6] H. Cartan (1967); *Calcul différentiel*; Hermann; Paris.

[7] G. Casella and R. L. Berger (1990), *Statistical Inference*; Wadsworth; Belmont, USA.

[8] G. Choquet (1969); *Topologie*; Masson; Paris.

[9] A. De Moivre (1718), *The Doctrine of Chances; or, A Method of Calculating the Probability of Events in Play*; W. Pearson; Londres. Et, en 1756, *The Doctrine of Chances (Third Edition)*, réimpression en 1967 : Chelsea, New York.

[10] J. Doob (1994), *Measure Theory*; Springer-Verlag; New York.

[11] R. Durrett (1991), *Probability : Theory and Examples*; Wadsworth and Brooks-Cole; Belmont, USA.

[12] W. Feller (1971), *An Introduction to Probability Theory and Its Applications* (Volume II); John Wiley; New York.

[13] A. Garsia (1970), *Topics in Almost Everywhere Convergence*; Markham; Chicago.

[14] A. Gut (1995), *An Intermediate Course in Probability*; Springer-Verlag; New York.

[15] N. B. Haaser and J. A. Sullivan (1991), *Real Analysis*; Dover; New York.

[16] C. Huygens (1657); Voir *Œuvres Complètes de Christiaan Huygens* (traduction française en 1920), Den Haage : Nijhoff.

[17] A. N. Kolmogorov (1933), *Grundbegriffe der Wahrscheinlichkeitrechnung*. Traduction anglaise : *Foundations of the Theory of Probability*, (1950), Nathan Morrison (traducteur); Chelsea; New York.

[18] P. A. Meyer (1966), *Probabilités et Potentiel*; Hermann; Paris.

[19] J. Neveu (1964), *Bases mathématiques du calcul des probabilités*. Paris.

[20] D. Pollard (1984), *Convergence of Stochastic Processes*; Springer-Verlag; New York.

[21] M. H. Protter and P. Protter (1988), *Calculus with Analytic Geometry* (Fourth Edition); Jones and Bartlett; Boston.

[22] M. Sharpe (1988), *General Theory of Markov Processes*; Academic Press; New York.

[23] G. F. Simmons (1963), *Introduction to Topology and Modern Analysis*; McGraw-Hill; New York.

[24] S. M. Stigler (1986), *The History of Statistics : The measurement of uncertainty before 1900*; Harvard University Press; Cambridge, USA.

[25] D. W. Stroock (1990), *A Concise Introduction to the Theory of Integration*; World Scientific; Singapour.

[26] S. J. Taylor (1973), *Introduction to Measure and Integration*; Cambridge University Press; Cambridge, Grande-Bretagne.

[27] D. Williams (1991), *Probability with Martingales*; Cambridge; Grande-Bretagne.

Index des notations

1_A fonction indicatrice, 11, 53
2^Ω classe de toutes les parties de Ω, 4, 7
A^* transposée de A, 98
$A_n \to A$ convergence des ensembles A_n vers A, 11
$E(X)$ espérance de X, 30, 55, 56
$E(X \mid \mathcal{G})$ espérance conditionnelle de X si $\mathcal{G}$, 206
$E_Q(X \mid \mathcal{G})$ espérance conditionnelle de X si $\mathcal{G}$ sous Q, 216
$f \star g$ produit de convolution des fonctions f et g, 129
H_n fonction de Haar, 244
J_g matrice jacobienne, 98
$L^1 := \mathcal{L}^1$ modulo l'égalité p.s., 58
$L^p := \mathcal{L}^p$ modulo l'égalité p.s., 58
m mesure de Lebesgue sur $\mathbb{R}$, 83
m_n mesure de Lebesgue sur $\mathbb{R}^n$, 93
$N(\mu, Q)$, 133
$N(\mu, \sigma^2)$ loi normale de moyenne μ et de variance σ^2, 46, 131
$P(A \mid B)$, 15, 16
$P \otimes Q$, 74
P_X loi de la variable X, 5, 29, 54
$P_X \otimes P_Y$ produit des lois P_X et P_Y, 75
$P_{(X,Y)}$ loi du couple (X, Y), 75
$Q \ll P$ Q absolument continu par rapport à P, 250
R_n fonction de Rademacher, 244
$X \vee Y$ maximum de X et Y, 53
$X \wedge Y$ minimum de X et Y, 53
X^+ maximum de X et 0, 56
X^- maximum de $-X$ et 0, 56
$X^{-1}(\mathcal{F})$ image inverse par X de la tribu $\mathcal{F}$, 51
$X_n \xrightarrow{L^p} X$ convergence dans L^p de X_n vers X, 149
$X_n \xrightarrow{P} X$ convergence en probabilité de X_n vers X, 149
$X_n \xrightarrow{\mathcal{L}} X$ convergence en loi de X_n vers X, 158
χ^2 loi du chi-deux, 89
χ^2_n loi du chi-deux à n degrés de liberté, 90
$\emptyset$ ensemble vide, 4, 7
$\frac{dQ}{dP}$ dérivée de Radon-Nykodim de Q par rapport à P, 251
$\Gamma(\alpha)$ fonction gamma, 46
$\Gamma^\perp$ ensemble des vecteurs orthogonaux à Γ, 198
$\hat{\mu}$ transformée de Fourier de μ, 110
$\langle\, , \rangle$ produit scalaire, 110, 195
$\mathrm{Cov}(X, Y)$ covariance de X et Y, 80, 97
$\mu_X \star \mu_Y$ produit de convolution, 123
$\otimes_{n=1}^{\infty} \mathcal{A}_n$, 76
Π opérateur de projection, 199
$\liminf_n A_n$ (limite inférieure des ensembles A_n), 11
$\limsup_n A_n$ (limite supérieure des ensembles A_n), 11, 78
$\sigma(\mathcal{C})$ tribu engendrée par $\mathcal{C}$, 7
σ^2, σ^2_X variance, 31, 63
$\{\omega\}$ singleton, 23
$\mathrm{Var}\,(X)$ variance de X, 63, *voir aussi* σ^2, σ^2_X
$\|X\|$ norme de X, 195
$\|X\|_p$ norme dans L^p, 215
$\mathbb{N}$ ensemble des entiers naturels, 5
$\mathbb{Q}$ ensemble des rationnels, 8
$\mathbb{R}$ ensemble des réels, 2
$\mathbb{Z}$ ensemble des entiers relatifs, 84
$\mathcal{B}$ tribu borélienne de $\mathbb{R}$, 8, 41
$\mathcal{B}^n$ tribu borélienne de $\mathbb{R}^n$, 93
C^∞ ensemble des fonctions indéfiniment différentiables, 167
$\mathcal{C}_\infty$ tribu asymptotique, 79
$\mathcal{E} \otimes \mathcal{F}$, 73
$\mathcal{F}_T$ tribu antérieure au temps d'arrêt T, 220
$\mathcal{L}^1$ ensemble des variables intégrables, 30, 56
$\mathcal{L}^1(\Omega, \mathcal{A}, P)$ espace $\mathcal{L}^1$ sur $(\Omega, \mathcal{A}, P)$, 56
$\mathcal{L}^p$ ensemble des variables de puissance $p$ième intégrable, 58
$\mathcal{N}$ classe des ensembles négligeables, 39

Index

A

additivité, 9
  σ-additivité, 9, 37
algèbre, 7, 37

B

Bayes, 18
Bernoulli
  fonction caractéristique, 113
  loi de, variable de, 32, 125, 182, 190
Bernoulli, Jacob, 1, 131
Berry-Esseen, 191
bêta (fonction, loi), 67
Bienaymé-Chebyshev (inégalité de), 32, 63, 215
binomiale
  fonction caractéristique, 113
  loi, 25, 27, 33, 126, 170, 190, 204
binomiale négative (loi), 33
Bolzano-Weierstrass (théorème de), 164
Bonferroni (inégalités de), 14
Borel (tribu de), 8
  sur $\mathbb{R}$, 8, 41
  sur $\mathbb{R}^n$, 93
Borel-Cantelli (lemme de), 77
borélien (ensemble), 8
Box-Muller (simulation de), 107

C

Cauchy (loi de), 47, 65, 90, 105, 155
  et loi normale, 105
Cauchy (suite de), 196
Cauchy-Schwarz (inégalité de), 62, 216
changement de variable (formule du), 99
chi-deux (loi du), 89, 90, 103, 127
Cobb-Douglas (loi de), 46
coefficient de corrélation, 97
complètement convergent, 81
convergence (de variables aléatoires)
  dans L^p, 148
  en loi, 158
  en moyenne d'ordre p, 149
  en probabilité, 149
  presque sûre, 148
  simple, 147
convergence étroite (de mesures), 157
convolution de fonctions (produit de), 129
convolution de mesures de probabilité (produit de), 123
covariance, 80, 97
  et indépendance, 97
  matrice de, 98

D

densité
  sur $\mathbb{R}$, 45, 84
  sur $\mathbb{R}^n$, 94
densité conditionnelle, 95, 210
deux à deux indépendants, 16
Dirac (masse de, mesure de), 45, 162
dominée (théorème de convergence), 57, 212
Doob
  décomposition de, 226
  inégalité de, 230, 231
  inégalité des montées de, 232
  inégalité maximale de, 229
  théorème d'arrêt de, 221

E

ergodique (loi forte des grands nombres), 182
espérance, 30, 55, 56
espérance conditionnelle
  définie dans L^1, 209
  définie dans L^2, 206
estimateur, 123, 138
événement, 4, 8
exponentielle
  fonction caractéristique, 115
  loi, 45, 65, 102
exponentielle double (loi), 47

F

Fatou (lemme de), 57, 212
fonction de répartition, 41, 54
  empirique, 191
Fourier (transformée de), 110
Fubini (théorème de), 74, 82

G

Galton-McAlister (loi de), 46
gamma (loi), 46, 68, 90, 127
  et χ^2, 90, 103, 127
  fonction caractéristique, 116
Gauss (loi de), *voir* loi normale
Gauss C.F., 131
géométrique (loi), 26, 33
Glivenko-Cantelli (théorème de), 191
Gosset, W., 103

H

Haar (fonctions de, systèmes de), 244
Helly (principe de), 164
Hölder (inégalité de), 213
hypergéométrique (loi), 24, 28

I

i.i.d. (indépendantes de même loi), 131
image de P par X, 54
indépendant(e)s
  deux à deux, 16
  événements, 15
  tribus, 71
  variables aléatoires, 71
indicatrice (fonction), 53
intensité (d'une variable aléatoire), 69

J

jacobien(ne) : matrice, déterminant, 98
Jensen (inégalité de), 47, 212

K

Kolmogorov, 181, 240, 241, 247
Kolmogorov (loi forte des grands nombres de), 181, 240

L

Laplace, 131
Laplace (loi de), *voir* exponentielle double
Lebesgue (mesure de)
  sur $\mathbb{R}$, 83
  sur $\mathbb{R}^n$, 93
Lebesgue (théorème de convergence dominée), 57, 212
Lévy (théorème de), 173
Lévy, P., 132
linéaire (estimateur), 138
linéaire (régression), 137
logistique (loi), 69
lognormale (loi), 46, 47, 68, 91, 121
loi d'une variable aléatoire, 5, 29, 54
loi forte des grands nombres, 179, 181, 240
loi zéro-un (0-1), 79

M

marginale (densité), 95
Markov (inégalité de), 31
martingale, 217
  fermée, 218, 237
  inverse, 239
  théorème de convergence, 235, 237
  théorème-limite central, 241
Mellin (transformée de), 121
mesurable (fonction), 51
Minkowski (inégalité de), 214
Moivre, A. de, 131
monotone (théorème de convergence), 57, 211
monotones (théorème des classes), 38, 39
Monte Carlo, 183
montées (nombre de), 232
multivariée (loi normale ou gaussienne), 132

N

négligeable (ensemble), 39
non corrélées (variables aléatoires), 137
normale (loi), 46, 65, 89, 100, 101, 103, 126, 127, 131
  fonction caractéristique, 114, 115, 126, 132
  multivariée, 132
  simulation de la, 107
normé (espace vectoriel), 195
normé complet (espace vectoriel), 196

O

orthogonale (matrice), 135
orthogonaux (vecteurs), 197

P

p.p., presque partout, 95
p.s., presque sûrement, 39, 57
Pareto (loi de), 33
Pascal (loi de), 33
Poincaré (identité de), 13
Poisson (loi de), 26, 27, 32, 126, 170, 204
  fonction caractéristique, 114
probabilité, 4, 9
probabilité conditionnelle, 16
projection (opérateur de), 199
Pythagore (théorème de), 197

R

Rademacher (fonction de), 244
Radon-Nikodym (théorème de, dérivée de), 250, 252
Rayleigh (loi de), 105, 106
régression, 137
régression (résidus de), 146
Riemann (fonction zêta de), 34
Riesz-Fischer (théorème de), 196

S

semi-définie positive (matrice), 98
signes aléatoires, 247
simple (variable aléatoire), 55
singleton, 23
Slutsky (théorème de), 167
sous-additivité, 13
sous-espace (de Hilbert), 198
sous-martingale, 225
stable
  par différences, 38
  par intersections finies, 38
  par limites croissantes, 38
statistique d'ordre, 107
Stone-Weierstrass (théorème de), 118
Student (loi de), 103
surmartingale, 225
symétrique
  densité, 91
  variable aléatoire, 91

T

temps d'arrêt, 219
  borné, 219
tension, tendue, 164
théorème-limite central, 187, 189
Tonelli-Fubini (théorème de), 74
topologique (espace), 52
transformation préservant la mesure, 182
triangulaire (loi), 52
tribu, 7
  engendrée par $\mathcal{C}$, 7
  triviale, 8
tribu asymptotique, 79

U

uniforme (loi)
  (fonction caractéristique), 114
  continue, 45
  discrète, 24, 34
uniforme intégrabilité, 237
unimodale (loi), 48

V

variable aléatoire, 5, 29, 51
variable aléatoire simple, 42, 55
variance, 31, 63

W

Weibull (loi de), 46

Z

zéro-un (loi), 79
Zêta (loi), 33

En 28 courts chapitres, cet ouvrage expose, dans un style simple et dépouillé, les notions fondamentales de la théorie des probabilités. Il conduit le lecteur des premiers rudiments aux principaux théorèmes-limites et à la notion d'espérance conditionnelle, aboutissement traditionnel des cours de licence ou de première année de master.

Les derniers chapitres sont consacrés à un aperçu de la théorie des martingales. Ils constituent une initiation aux processus stochastiques, en même temps que l'exposé d'une théorie qui est à la base de la plupart des applications actuelles des probabilités.

La lecture de ce livre ne présuppose aucune connaissance en probabilités. Les connaissances nécessaires en analyse sont celles des deux premières années d'université. Les notions plus avancées, comme l'intégrale de Lebesgue, les espaces de Hilbert, sont introduites et étudiées quand elles deviennent nécessaires à la progression de l'exposé, ce qui rend l'ouvrage accessible aux lecteurs non universitaires intéressés notamment par les applications.

Le cours est complété par 331 énoncés d'exercices.

Jean Jacod est professeur à l'Université Paris VI-Pierre et Marie Curie. Philip Protter est professeur à Cornell University (Ithaca, N. Y., États-Unis).

Collection enseignement des mathématiques

ISBN 2-84225-050-8

Graphisme : Massin