

Kit de survie des probabilités*

Fabien Campillo**

Octobre 2013

Ce kit propose un survol des notions de base de la théorie des probabilités : des fondements, aux chaînes de Markov à espace d'états au plus dénombrable, en passant par les martingales. Je me suis efforcé d'être concis, de proposer des définitions et des énoncés de théorèmes rigoureux sans en donner de démonstration, pour cela le lecteur consultera les excellentes références [1] et [3], toutes à taille humaine, et [2] pour des exercices avec rappels de cours. Pour les processus stochastiques on consultera également [4].

Quelques documents de qualité sont accessibles sur la toile, comme les supports de cours de l'UMPC¹. Les "w" désignent des liens vers Wikipedia si possible français mais parfois anglais lorsque l'article français n'existe pas ou est moins bon.

TABLE DES MATIÈRES

1	Espaces de probabilité	1
2	Variables aléatoires	2
3	Lois de probabilité	3
4	Espérance	4
5	Espaces $L^p(\Omega)$	4
6	Fonctions caractéristiques	5
7	Convergence de suites de variables aléatoires	6
8	Convergence en loi	6
9	Probabilités conditionnelles d'événements	7
10	Espérance conditionnelle	7
11	Quelques lois de probabilité usuelles	10
12	Martingales à temps discret	13
13	Chaînes de Markov à espace d'états dénombrable	14
	Références	17
	Index	18

1 Espaces de probabilité

On convient de représenter une expérience aléatoire \mathcal{E} , c'est-à-dire une expérience dont l'issue est soumise à l'aléa, par l'ensemble Ω des résultats possibles de cette expérience. Un élément ω de Ω est appelé *réalisation* et Ω est l'ensemble des réalisations.

Événements et tribus^w

Un événement aléatoire A correspondant à l'expérience aléatoire \mathcal{E} est le sous-ensemble des $\omega \in \Omega$ t.q. A est réalisé. La classe \mathcal{F} de tous les événements A correspondant à l'expérience aléatoire \mathcal{E} est donc un sous-ensemble de $\mathcal{P}(\Omega)$, l'ensemble des parties de Ω . Cette

classe $\mathcal{F} \subset \mathcal{P}(\Omega)$ doit être une tribu sur Ω , également appelée σ -algèbre, c'est-à-dire vérifier :

- (i) $\Omega \in \mathcal{F}$,
- (ii) \mathcal{F} est stable par passage au complémentaire, i.e. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$,
- (iii) \mathcal{F} est stable par réunion dénombrable, i.e. $A_n \in \mathcal{F} \Rightarrow \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{F}$.

Les éléments de \mathcal{F} sont appelés *événements*.

La tribu engendrée par une classe \mathcal{C} de sous-ensembles de Ω est la plus petite tribu, notée $\sigma(\mathcal{C})$, contenant \mathcal{C} . C'est aussi l'intersection de toutes les tribus sur Ω contenant \mathcal{C} .

La tribu borélienne $\mathcal{B}(\mathbb{R}^d)$ sur \mathbb{R}^d est la tribu engendrée par la classe des ouverts de \mathbb{R}^d . C'est aussi la tribu engendrée par la classe des pavés $]a_1, b_1[\times \cdots \times]a_d, b_d[$, pour tout $-\infty \leq a_i < b_i \leq +\infty$. Les éléments de $\mathcal{B}(\mathbb{R}^d)$ sont appelés les (ensembles) *boréliens*. Plus généralement, la tribu borélienne associée à un espace topologique (E, \mathcal{T}) est la tribu $\mathcal{B}(E)$ engendrée par la classe \mathcal{T} des ouverts.

Espace de probabilité^w

Le couple (Ω, \mathcal{F}) est appelé *espace probabilisable*. Une *mesure de probabilité*, simplement appelée probabilité, sur l'espace (Ω, \mathcal{F}) est une application \mathbb{P} de \mathcal{F} dans \mathbb{R}_+ t.q. :

- (i) $\mathbb{P}(\Omega) = 1$;
- (ii) pour toute suite $(A_n)_{n \in \mathbb{N}}$ d'événements de \mathcal{F} deux à deux disjoints :

$$\mathbb{P}\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} \mathbb{P}(A_n). \quad (\sigma\text{-additivité})$$

Une telle application ne vérifiant que (ii) mais pas (i) est appelée *mesure finie positive*. \mathbb{P} est donc une mesure positive de masse totale égale à 1.

Le triplet $(\Omega, \mathcal{F}, \mathbb{P})$ est appelé *espace de probabilité*, $\mathbb{P}(A)$ est la probabilité de l'événement A . \mathbb{P} possède les propriétés élémentaires suivantes :

- $\mathbb{P}(\emptyset) = 0$;
 - $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$;
 - $A \subset B \Rightarrow \mathbb{P}(A \setminus B) = \mathbb{P}(B) - \mathbb{P}(A)$ et $\mathbb{P}(B) \geq \mathbb{P}(A)$;
 - $\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B)$;
 - $\mathbb{P}(\sum_{n \in \mathbb{N}} A_n) \leq \sum_{n \in \mathbb{N}} \mathbb{P}(A_n)$.
- pour tout $A, B, A_n \in \mathcal{F}$.

Sur un espace $(\Omega, \mathcal{F}, \mathbb{P})$, un sous-ensemble A de Ω est dit *négligeable* s'il existe $B \in \mathcal{F}$ t.q. $\mathbb{P}(B) = 0$ et $A \subset B$. Un négligeable n'est donc pas nécessairement un élément de \mathcal{F} . Une tribu est dite *complète* lorsqu'elle contient tous les sous-ensembles négligeables. Une propriété \mathbb{P}

*© Fabien Campillo

**INRIA, Fabien.Campillo@inria.fr

1. <http://www.proba.jussieu.fr/supports.php>

est dite vraie \mathbb{P} -presque sûrement, noté \mathbb{P} -p.s., lorsque l'ensemble des éléments pour lesquels elle est fautive est négligeable.

L'espace de probabilité produit $(\Omega, \mathcal{F}, \mathbb{P}) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \times \mathbb{P}_2)$ est défini à l'aide de la tribu produit $\mathcal{F}_1 \otimes \mathcal{F}_2$ qui est la plus petite tribu sur Ω contenant les événements produits $A_1 \times A_2$ avec $A_i \in \mathcal{F}_i$.

Nous passerons sous silence la question pourtant importante de la construction des mesures de probabilités, notamment le théorème des classes monotones, qui est aussi fort utile en pratique [3, Ch. 6].

Événements et tribus indépendants

Deux événements A et B sont dits *indépendants*, noté $A \perp B$, lorsque :

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

Une famille finie A_1, A_2, \dots, A_n d'événements est dite *indépendante* si pour tout $1 \leq k \leq n$ et tout $\{i_1, \dots, i_k\} \subset \{1, \dots, n\}$ deux à deux distincts :

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \dots \mathbb{P}(A_{i_k}).$$

Des événements peuvent être deux à deux indépendants sans être indépendants.

Une famille finie $\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_n$ de sous-tribus de \mathcal{F} est dite *indépendante* si :

$$\mathbb{P}(A_1 \cap \dots \cap A_n) = \mathbb{P}(A_1) \dots \mathbb{P}(A_n)$$

pour tout $A_i \in \mathcal{F}_i$. Une famille de sous-tribus $(\mathcal{F}_n)_{n \in \mathbb{N}}$ de \mathcal{F} est dite indépendante si $\mathcal{F}_1, \dots, \mathcal{F}_n$ est indépendante pour tout $n \geq 1$.

Probabilités sur un ensemble fini

Si Ω est fini, on choisit généralement $\mathcal{F} = \mathcal{P}(\Omega)$. Dans ce cas toute probabilité \mathbb{P} sur (Ω, \mathcal{F}) est entièrement caractérisée par la donnée de $p(\omega) = \mathbb{P}(\{\omega\})$ t.q. $p(\omega) \geq 0$ pour tout $\omega \in \Omega$ et $\sum_{\omega \in \Omega} p(\omega) = 1$, ainsi $\mathbb{P}(A) = \sum_{\omega \in A} p(\omega)$.

La *probabilité uniforme* \mathbb{P} sur Ω est définie par :

$$\mathbb{P}(A) \stackrel{\text{déf}}{=} \frac{|A|}{|\Omega|} \quad (\text{probabilité uniforme})$$

où $|A|$ désigne le cardinal de A . Dans ce cadre, on fait appel aux formules de dénombrement classiques :

- le nombre d'applications d'un ensemble de cardinal p dans un ensemble de cardinal n est n^p ;
- le nombre de bijections d'un ensemble de cardinal n dans lui-même, i.e. le nombre de *permutations*, est $n!$;
- le nombre d'injections d'un ensemble de cardinal p dans un ensemble de cardinal n , i.e. le nombre d'*arrangements*, avec $n \geq p$ est :

$$A_n^p \stackrel{\text{déf}}{=} \frac{n!}{(n-p)!} \quad (\text{nombre d'arrangements})$$

- le nombre de sous-ensembles de cardinal p dans un ensemble de cardinal n ($p \leq n$), i.e. le nombre de *combinaisons*, est :

$$\binom{n}{p} \stackrel{\text{déf}}{=} \frac{A_n^p}{p!} = \frac{n!}{p!(n-p)!}$$

(nombre de combinaisons)

La notation anglo-saxonne $\binom{n}{p}$ est maintenant plus répandue que la notation classique C_n^p .

2 Variables aléatoires

Une *variable aléatoire* (v.a.) sur (Ω, \mathcal{F}) à valeurs dans un ensemble E muni d'une tribu \mathcal{E} est une application mesurable X de (Ω, \mathcal{F}) dans (E, \mathcal{E}) , c'est-à-dire vérifiant : $X^{-1}(B) \in \mathcal{F}$ pour tout $B \in \mathcal{E}$. On note :

$$\{X \in B\} \stackrel{\text{déf}}{=} X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}.$$

Lorsque $E = \mathbb{R}$ on parle de *variables aléatoires réelles* (v.a.r.), lorsque $E = \mathbb{R}^d$ on parle de *vecteurs aléatoires réels*.

Une v.a. est dite *discrète* lorsqu'elle ne prend qu'un nombre au plus dénombrable de valeurs ; elle est dite *étagée* ou *simple* lorsqu'elle ne prend qu'un nombre fini de valeurs. Les v.a.r. étagées se mettent sous la forme : $X(\omega) = \sum_{i=1}^n a_i \mathbf{1}_{A_i}(\omega)$ où $a_i \in \mathbb{R}$, $A_i \in \mathcal{F}$ et $\mathbf{1}_A(\omega) \stackrel{\text{déf}}{=} 1$ si $\omega \in A$, 0 sinon.

Si X est une v.a. à valeurs dans \mathbb{R}^d et ϕ une application mesurable $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, alors la composée $Y = \phi(X) = \phi \circ X$ est une v.a.r.

Soit X et Y deux v.a. définies sur un même espace de probabilité et à valeurs dans le même ensemble, on dira que X et Y sont p.s. égaux, noté $X \stackrel{\text{p.s.}}{=} Y$ ou $X = Y$ p.s., lorsque $\{X \neq Y\}$ est un ensemble négligeable. De la même façon on définit $X \geq Y$ p.s. etc.

Tribus engendrées

La *tribu engendrée* par une v.a. X , également appelée *tribu naturelle*, définie sur (Ω, \mathcal{F}) et à valeurs dans (E, \mathcal{E}) , notée $\sigma(X)$, est la tribu sur (E, \mathcal{E}) définie par :

$$\sigma(X) \stackrel{\text{déf}}{=} \{\{X \in B\} : B \in \mathcal{E}\}$$

C'est la plus petite tribu sur Ω qui rende X mesurable. C'est la *tribu image* de \mathcal{F} par X sur E . Soit \mathcal{G} une sous-tribu de \mathcal{F} , X est dite \mathcal{G} -mesurable lorsque $\sigma(X) \subset \mathcal{G}$.

Une v.a. Y est dite $\sigma(X)$ -mesurable lorsque $\sigma(Y) \subset \sigma(X)$. Soit X une v.a. à valeurs dans (E, \mathcal{E}) et Y une v.a. à valeurs dans \mathbb{R}^d : Y est $\sigma(X)$ -mesurable si et seulement si il existe une fonction mesurable $f : (E, \mathcal{E}) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ t.q. $Y = f(X)$.

La tribu engendrée par la famille finie $\{X_0, \dots, X_n\}$ est définie par :

$$\begin{aligned} \sigma(X_0, \dots, X_n) &\stackrel{\text{déf}}{=} \sigma\left(\bigcup_{i=0}^n \sigma(X_i)\right) \\ &= \sigma\left(\left\{\bigcap_{i=0}^n \{X_i \in B_i\} : B_i \in \mathcal{F}\right\}\right). \end{aligned}$$

La tribu engendrée par une suite $(X_n)_{n \in \mathbb{N}}$ de v.a. est définie par :

$$\sigma(X_n, n \geq 0) \stackrel{\text{d\u00e9f}}{=} \sigma\left(\bigcup_{n \geq 0} \sigma(X_n)\right) = \sigma\left(\left\{\bigcap_{i=0}^n \{X_i \in B_i\} : n \geq 0, B_i \in \mathcal{F}\right\}\right)$$

Attention $\bigcup_{n \in \mathbb{N}} \sigma(X_0, \dots, X_n)$ n'est pas nécessairement une tribu, c'est seulement une algèbre.

Convergence simple de suite de v.a.

Une suite $(X_n)_{n \in \mathbb{N}}$ à valeurs dans \mathbb{R} converge simplement vers une v.a. X , noté $X_n \rightarrow X$ ou $\lim_{n \rightarrow \infty} X_n = X$, lorsque :

$$X_n(\omega) \xrightarrow[n \rightarrow \infty]{} X(\omega), \quad \forall \omega \in \Omega.$$

On notera $X_n \uparrow X$ ou $X_n \downarrow X$ lorsque cette convergence est croissante ou décroissante.

Nous allons considérer des v.a.r. pouvant prendre des valeurs infinies, i.e. des fonctions $X : \Omega \mapsto \bar{\mathbb{R}}$ telles que $X^{-1}(\{-\infty\})$, $X^{-1}(\{+\infty\})$ et $X^{-1}(B) \in \mathcal{F}$ pour tout $B \in \mathcal{B}(\mathbb{R})$. On utilisera la convention " $0 \times \infty = 0$ ".

On peut montrer que :

- Pour toute v.a.r. X positive il existe une suite croissante $(X_n)_{n \in \mathbb{N}}$ de v.a.r. étagées positives t.q. $X_n \uparrow X$.
- Pour toute v.a.r. X il existe une suite $(X_n)_{n \in \mathbb{N}}$ de v.a.r. étagées t.q. $X_n \rightarrow X$ et $\sup_{n \in \mathbb{N}} |X_n| \leq |X|$.

Ces résultats sont utilisés pour la construction de l'espérance.

3 Lois de probabilité

On considère une v.a. X de (Ω, \mathcal{F}) à valeurs dans (E, \mathcal{E}) . On munit (Ω, \mathcal{F}) d'une mesure de probabilité \mathbb{P} , on appelle *loi de probabilité* de X la mesure de probabilité définie sur (E, \mathcal{E}) par :

$$\mathbb{P}_X(B) \stackrel{\text{d\u00e9f}}{=} \mathbb{P}(X \in B), \quad \forall B \in \mathcal{E}.$$

C'est la mesure image de \mathbb{P} par X sur (E, \mathcal{E}) .

Fonctions de répartition

La *fonction de répartition* d'une v.a. X à valeurs dans \mathbb{R} est la fonction définie par :

$$F_X(x) \stackrel{\text{d\u00e9f}}{=} \mathbb{P}(X \leq x), \quad x \in \mathbb{R}.$$

La fonction de répartition d'une v.a. X à valeurs dans \mathbb{R}^d est la fonction définie par :

$$F_X(x) \stackrel{\text{d\u00e9f}}{=} \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d)$$

pour $x = (x_1, \dots, x_d) \in \mathbb{R}^d$. On peut naturellement définir la fonction de répartition d'une mesure de probabilité Q sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ par $F_X(x) \stackrel{\text{d\u00e9f}}{=} Q(\prod_{i=1}^d]-\infty, x_i])$ pour tout $x \in \mathbb{R}^d$.

La fonction de répartition d'un v.a. réelle X vérifie :

- (i) F_X est croissante,

(ii) F_X est continue à droite,

(iii) $F_X(x) \uparrow 1$ si $x \uparrow +\infty$ et $F_X(x) \downarrow 0$ si $x \uparrow -\infty$.

Dans le cas vectoriel on définit :

$$\Delta_{a_i, b_i} \Psi(x) \stackrel{\text{d\u00e9f}}{=} \Psi(x + \gamma_i b_i) - \Psi(x + \gamma_i a_i).$$

où $\gamma_i = (0, \dots, 0, 1, 0, \dots, 0)$ (1 en i ème position). La fonction de répartition vérifie :

- (i) $\Delta_{a_1, b_1} \cdots \Delta_{a_n, b_n} F_X(x) \geq 0$ pour tout $a, b \in \mathbb{R}^d$ t.q. $a_i \leq b_i$;
- (ii) si $x_i^k \uparrow x_i$ pour $i = 1, \dots, n$ alors $F_X(x^k) \uparrow F_X(x)$;
- (iii) $F_X(x) \rightarrow 0$ dès qu'une composante de x tend vers $-\infty$ et $F_X(x) \rightarrow 1$ lorsque toutes les composantes de x convergent vers $+\infty$.

Soit X une v.a. à valeurs dans \mathbb{R}^d , un point $x \in \mathbb{R}^d$ est un appelé *atome* de la loi de probabilité \mathbb{P}_X si $\mathbb{P}_X(\{x\}) = \mathbb{P}(X = x) > 0$. La loi \mathbb{P}_X sera dite *diffuse* lorsqu'elle ne comporte pas d'atome. Pour une loi diffuse les ensembles finis sont négligeables. Les atomes correspondent aux points de discontinuité de la fonction de répartition.

Densités

On considère une loi Q sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, si Q est absolument continue par rapport à la mesure de Lebesgue $\lambda(dx)$ sur \mathbb{R}^d , alors le théorème de Radon-Nikodym assure qu'il existe une fonction $p : \mathbb{R}^d \mapsto \mathbb{R}_+$ t.q. :

$$Q(B) = \int_B p(x) dx, \quad \forall B \in \mathcal{B}(\mathbb{R}^d).$$

En particulier $\int_{\mathbb{R}^d} p(x) dx = 1$. La fonction p est unique à un Lebesgue-négligeable près, elle est appelée *densité* de Q . Si F est la fonction de répartition de Q , alors p est une densité de Q si est seulement si :

$$F(x) = \int_{\prod_{i=1}^d]-\infty, x_i]} p(x') dx', \quad \forall x \in \mathbb{R}^d$$

L'équivalence entre ces deux définitions est due au fait que $\{\prod_{i=1}^d]-\infty, x_i]; x \in \mathbb{R}^d\}$ est une classe génératrice de $\mathcal{B}(\mathbb{R}^d)$.

Une v.a.r. X de loi \mathbb{P}_X est dite à *densité* lorsque sa loi \mathbb{P}_X admet une densité, notée $x \mapsto p_X(x)$, par rapport à la mesure de Lebesgue.

Lois et densités marginales

On considère $X = (X_1, \dots, X_d)$ un vecteur aléatoire à valeurs dans \mathbb{R}^d de loi \mathbb{P}_X . La *loi de probabilité marginale* de la variable X_i est définie par :

$$\mathbb{P}_{X_i}(A) = \mathbb{P}_X(\mathbb{R} \times \cdots \times \mathbb{R} \times A \times \mathbb{R} \times \cdots \times \mathbb{R})$$

pour tout $A \in \mathcal{B}(\mathbb{R})$. Lorsque X admet une densité p_X , la densité p_{X_i} de la composante X_i est donnée par la *densité marginale* :

$$p_{X_i}(\xi) \stackrel{\text{d\u00e9f}}{=} \int_{\mathbb{R}^{d-1}} p_X(x_1, \dots, x_{i-1}, \xi, x_{i+1}, \dots, x_d) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_d$$

pour $\xi \in \mathbb{R}$. Il s'agit de la densité de la loi marginale \mathbb{P}_{X_i} , i.e. la loi de la composante X_i . Attention une loi n'est pas déterminée par la donnée de ses lois marginales.

Indépendance de v.a.

Une suite finie de v.a. (X_1, \dots, X_n) est dite *indépendante* si la loi de probabilité de $X = (X_1, \dots, X_n)$ est le produit des lois de probabilité des X_i , c'est-à-dire $\mathbb{P}_X = \mathbb{P}_{X_1} \times \dots \times \mathbb{P}_{X_n}$ i.e. :

$$\mathbb{P}(X_i \in B_i, i = 1 \dots n) = \prod_{i=1}^n \mathbb{P}(X_i \in B_i)$$

pour tout $B_i \in \mathcal{E}$, ce qui est aussi équivalent à :

$$F_X(x_1, \dots, x_n) \equiv F_{X_1}(x_1) \times \dots \times F_{X_n}(x_n).$$

Dans le cas à densité, cette propriété s'écrit $p_X(x) \equiv p_{X_1}(x_1) \times \dots \times p_{X_n}(x_n)$. Une suite $(X_n)_{n \in \mathbb{N}}$ est indépendante si et seulement si (X_0, \dots, X_n) est indépendante quelque soit $n \in \mathbb{N}$.

L'indépendance des v.a. est équivalente à l'indépendance des tribus qu'elles engendrent : une suite $(X_i)_{i \in I}$ avec I fini ou dénombrable est indépendante si et seulement si la suite de tribus engendrées correspondantes $(\sigma(X_i))_{i \in I}$ est indépendante.

Une famille $(X_i)_{i \in I}$ est dite *indépendante et identiquement distribuée*, noté i.i.d. lorsque la famille est indépendante et lorsque les variables ont toutes la même loi, i.e. $\mathbb{P}_{X_i}(B) = Q(B)$ pour tout $i, j \in I$; on note $X_i \stackrel{\text{id}}{\sim} Q$.

4 Espérance

On considère une v.a.r. X définie sur $(\Omega, \mathcal{F}, \mathbb{P})$. Nous allons définir sous certaines conditions l'*espérance* de X par rapport à \mathbb{P} , notée $\mathbb{E}(X)$, c'est-à-dire l'intégrale de X par rapport à \mathbb{P} :

$$\mathbb{E}(X) \stackrel{\text{d\u00e9f}}{=} \int_{\Omega} X \, d\mathbb{P}$$

également notée $\int_{\Omega} X(\omega) \, d\mathbb{P}(\omega)$ ou $\int_{\Omega} X(\omega) \, \mathbb{P}(d\omega)$. La construction se fait en 3 étapes classiques :

(i) *Cas étagé/positif* : Soit X une v.a.r. étagée positive, $X = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$, on pose :

$$\mathbb{E}(X) \stackrel{\text{d\u00e9f}}{=} \sum_{i=1}^n a_i \mathbb{P}(X \in A_i).$$

(ii) *Cas positif* : Supposons $X \geq 0$, il existe une suite croissante $(X_n)_{n \in \mathbb{N}}$ de v.a.r. étagées t.q. $X_n \uparrow X$ (convergence simple) et on définit :

$$\mathbb{E}(X) \stackrel{\text{d\u00e9f}}{=} \lim_{n \rightarrow \infty} \mathbb{E}(X_n).$$

(iii) *Cas intégrable* : Dans le cas général, $X = X^+ - X^-$ (2) et on définit $\int_{\Omega} X^+ \, d\mathbb{P}$ et $\int_{\Omega} X^- \, d\mathbb{P}$ comme précédemment. Si l'une de ces deux quantités est finie on pose :

$$\mathbb{E}(X) \stackrel{\text{d\u00e9f}}{=} \mathbb{E}(X^+) - \mathbb{E}(X^-).$$

2. On note $X^+ = X \vee 0$, $X^- = (-X) \vee 0$ où $x \vee y = \max(x, y)$ et $x \wedge y = \min(x, y)$.

Donc X admet une espérance (par rapport à \mathbb{P}) lorsque $X \geq 0$ (et l'espérance peut éventuellement être infinie) ou lorsque X est *intégrable* c'est-à-dire lorsque $\mathbb{E}(|X|) = \mathbb{E}(X^+) + \mathbb{E}(X^-) < \infty$.

Naturellement :

$$\mathbb{E}\phi(X) = \int_{\Omega} \phi(X) \, d\mathbb{P} = \int_E \phi(x) \, \mathbb{P}_X(dx).$$

Théorème de Fubini

Si X est une v.a. intégrable définie sur l'espace produit $(\Omega, \mathcal{F}, \mathbb{P}) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \times \mathbb{P}_2)$ alors $\int_{\Omega_i} X(\omega_1, \omega_2) \, d\mathbb{P}_i(\omega_i)$ est \mathcal{F}_j -mesurable et intégrable (avec $i \neq j$ et $i, j \in \{1, 2\}$) et :

$$\begin{aligned} \int_{\Omega} X \, d\mathbb{P} &= \int_{\Omega_2} \left(\int_{\Omega_1} X \, d\mathbb{P}_1 \right) \, d\mathbb{P}_2 \\ &= \int_{\Omega_1} \left(\int_{\Omega_2} X \, d\mathbb{P}_2 \right) \, d\mathbb{P}_1. \end{aligned}$$

Propriétés

- Soit $X = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$ positive³ ou intégrable, alors $\mathbb{E}(X) = \sum_{i=1}^n a_i \mathbb{P}(A_i)$.
- Linéarité : soit X_i des v.a.r. positives et $a_i \geq 0$ ou soit X_i des v.a.r. et $a_i \in \mathbb{R}$, alors $\mathbb{E}(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i \mathbb{E}(X_i)$.
- Pour des v.a.r. X et Y quelconques, si $X \geq Y \geq 0$ p.s. alors $\mathbb{E}(X) \geq \mathbb{E}(Y) \geq 0$. Pour des v.a.r. intégrables si $X \geq Y$ p.s. alors $\mathbb{E}(X) \geq \mathbb{E}(Y)$. Dans les deux cas si $X = Y$ p.s. alors $\mathbb{E}(X) = \mathbb{E}(Y)$.
- Si $X = 0$ p.s. alors $\mathbb{E}(X) = 0$, la réciproque est fautive sauf si $X \geq 0$ p.s. ou $X \leq 0$ si p.s.
- Soit X une v.a.r. intégrable alors : $|\mathbb{E}X| \leq \mathbb{E}|X|$; si $\mathbb{E}(X) < \infty$ alors $X < \infty$ p.s.
- si $X \geq 0$ p.s. et $\mathbb{P}(X > 0) > 0$ alors $\mathbb{E}(X) > 0$.

Inégalités

On dispose de plusieurs inégalités afin de contrôler la probabilité des événements de la forme $\{|X| \geq a\}$ à l'aide des moments de la v.a.r. X . Pour tout $a > 0$ on a :

$$\mathbb{P}(|X| \geq a) \leq \frac{1}{a} \mathbb{E}|X|, \quad (\text{Inégalité de Markov})$$

$$\mathbb{P}(|X| \geq a) \leq \frac{1}{a^p} \mathbb{E}|X|^p,$$

$$\mathbb{P}(|X - \mathbb{E}X| \geq a) \leq \frac{1}{a^2} \text{Var}(X), \quad \text{si } \mathbb{E}(X^2) < \infty, \\ (\text{inégalité de Bienaymé-Tchebychev})$$

$$\mathbb{P}(|X| \geq a) \leq \frac{1}{\phi(a)} \mathbb{E}\phi(|X|)$$

pour tout $a \geq 0$, $p \geq 1$ et toute fonction croissante et positive ϕ . Par ailleurs on considère une v.a. X à valeurs dans \mathbb{R}^n et une fonction convexe $\phi : \mathbb{R}^n \mapsto \mathbb{R}$, si X et $\phi(X)$ sont intégrables, alors :

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X). \quad (\text{inégalité de Jensen})$$

5 Espaces $L^p(\Omega)$

Soit X une v.a.r., on définit son *moment d'ordre p* $\mathbb{E}(X^p)$, sa *variance* $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)^2]$ et son *écart-type* $\sigma(X) = \sqrt{\text{Var}(X)}$.

3. Avec la convention $\infty \times 0 = 0$.

Soit X et Y deux v.a.r., on définit la *covariance* de X et Y par :

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

On a $\text{Cov}(X, X) = \text{Var}(X)$. On définit le *coefficient de corrélation* par :

$$\rho_{X,Y} \stackrel{\text{déf}}{=} \text{Cov}(X, Y) / (\sqrt{\text{Var}X} \sqrt{\text{Var}Y}).$$

Les v.a. X et Y sont dites non-corrélées lorsque $\rho_{X,Y} = 0$. On vérifie aisément que l'indépendance implique la non-corrélation, *la réciproque est fautive* (sauf dans le cas gaussien, mais de façon très précise).

Dans le cas de vecteurs aléatoires réels on adoptera la notation vectorielle $X = (X_1, \dots, X_d)^*$ où "*" est l'opérateur de transposition. On définit alors la *matrice de covariance* par :

$$\text{Cov}(X) \stackrel{\text{déf}}{=} \left[\text{Cov}(X_i, X_j) \right]_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n}} \in \mathbb{R}^{d \times d}.$$

On a $\text{Cov}(X) = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^*] = \mathbb{E}(X X^*) - (\mathbb{E}X)(\mathbb{E}X)^*$. La matrice covariance est une matrice symétrique semi-définie positive. On définit enfin la *matrice de corrélation* par

$$\text{Cor}(X) \stackrel{\text{déf}}{=} \left[\rho_{X_i, X_j} \right]_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n}} \in \mathbb{R}^{d \times d}.$$

Pour tout $p, q > 1$ conjugués (i.e. $\frac{1}{p} + \frac{1}{q} = 1$) :

$$\mathbb{E}|XY| \leq \left\{ \mathbb{E}(|X|^p) \right\}^{1/p} \left\{ \mathbb{E}(|Y|^q) \right\}^{1/q} \quad (\text{inégalité de Hölder})$$

et pour tout $p \geq 1$:

$$\mathbb{E}(|X + Y|^p)^{1/p} \leq \mathbb{E}(|X|^p)^{1/p} + \mathbb{E}(|Y|^p)^{1/p}. \quad (\text{inégalité de Minkowski})$$

Pour $p \geq 1$, on définit $L^p(\Omega, \mathcal{F}, \mathbb{P})$, $L^p(\Omega)$ en abrégé, l'espace vectoriel des (classes de) v.a. p.s. sûrement égales et qui sont L^p intégrables i.e. $\mathbb{E}(|X|^p) < \infty$. D'après l'inégalité de Minkowski :

$$\|X\|_p \stackrel{\text{déf}}{=} (\mathbb{E}(|X|^p))^{1/p}$$

est une norme sur $L^p(\Omega)$. $(L^p(\Omega), \|\cdot\|_p)$ est un espace de Banach, il est en effet complet (i.e. toute suite de Cauchy y est convergente). L'application $p \rightarrow \|X\|_p$ est croissante, ainsi $L^p(\Omega) \subset L^q(\Omega)$ si $p \geq q$.

Lorsque $X \in L^p(\Omega)$, on dira que X est L^p -intégrable ou que $X \in L^p$, l'intégrabilité est équivalente à la L^1 -intégrabilité.

6 Fonctions caractéristiques

À une loi de probabilité on peut associer différentes fonctions qui permettent de caractériser partiellement ou totalement cette loi. La plus importante de ces fonctions est la :

Fonction caractéristique^w

Soit X une v.a. à valeurs dans \mathbb{R}^d , sa *fonction caractéristique* est définie par :

$$\Phi_X(t) \stackrel{\text{déf}}{=} \mathbb{E}e^{it^*X}, \quad t \in \mathbb{R}^d$$

à valeurs complexes, t^*X est le produit scalaire de t par X . Il s'agit donc de la *transformée de Fourier* de \mathbb{P}_X . Cette fonction est *toujours* définie car $e^{it^*X} = \cos(t^*X) + i \sin(t^*X)$ est de module 1 donc Φ_X est de module inférieur à 1. Elle est également continue et de plus $\Phi_X(0) = 1$.

La propriété fondamentale de la transformée de Fourier est que la fonction Φ_X caractérise la loi de la v.a. X :

$$\Phi_X \equiv \Phi_Y \iff \mathbb{P}_X = \mathbb{P}_Y.$$

Par ailleurs les dérivées de $\Phi_X(t)$ en $t = 0$ permettent de calculer les différents moments de X : Si $|X|^m$ est intégrable, alors la fonction $\Phi_X(t)$ est m fois continûment différentiable sur \mathbb{R}^d et

$$\frac{\partial^m \Phi_X(t)}{\partial t_{i_1} \partial t_{i_2} \dots \partial t_{i_m}} = i^m \mathbb{E}\left(e^{it^*X} X_{i_1} X_{i_2} \dots X_{i_m}\right)$$

pour tout $i_1, \dots, i_m \in \{1, \dots, d\}$. Pour $d = 1$: si $X \in L^1$ alors $\mathbb{E}(X) = i \Phi'_X(0)$, si $X \in L^2$ alors $\mathbb{E}(X^2) = -\Phi''_X(0)$ etc. Une autre façon d'exprimer cette propriété est :

$$\Phi_X(t) = 1 + it \Phi'_X(t) \mathbb{E}(X) - \frac{1}{2} \Phi''_X(t) \mathbb{E}(X^2) + o(|t|^2).$$

La propriété d'indépendance s'exprime de façon naturelle grâce à la fonction caractéristique. Les v.a. X_1, \dots, X_n à valeurs dans \mathbb{R}^d sont indépendantes si et seulement si :

$$\Phi_{(X_1, \dots, X_n)}(t_1, \dots, t_n) = \Phi_{X_1}(t_1) \dots \Phi_{X_n}(t_n)$$

pour tout $t_1, \dots, t_m \in \mathbb{R}^d$.

Il existe plusieurs *formules d'inversion de type Fourier*. Par exemple, si Φ_X est intégrable par rapport à la mesure de Lebesgue sur \mathbb{R}^d (mais c'est toujours le cas), alors la fonction de répartition F_X est absolument continue et ainsi la loi de X admet une densité, elle est donnée par :

$$p_X(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-it^*x} \Phi_X(t) dt.$$

Fonction génératrice des moments^w

La *fonction génératrice des moments* (f.g.m.) de X est la *transformée de Laplace* de \mathbb{P}_X définie par :

$$M_X(t) \stackrel{\text{déf}}{=} \mathbb{E}e^{t^*X}$$

pour tout t t.q. e^{t^*X} soit intégrable. Ce concept est moins puissant que celui de fonction caractéristique car il nécessite que la variable soit intégrable.

Si $M_X(t) = M_Y(t)$ pour tout t alors X et Y ont même loi. La propriété principale de la f.g.m. est :

$$\mathbb{E}(X^k) = M_X^{(k)}(0) = \frac{d^k M_X}{dt^k}(0).$$

La f.g.m. peut également être définie pour des v.a. à valeurs dans des espaces généraux. Mais c'est dans le cas des v.a. à valeurs dans un espace fini ou dénombrable qu'elle est la plus utile.

Fonction génératrice des probabilités^w (cas discret)

Dans le cas d'une v.a. X à valeurs dans \mathbb{N} , on peut également faire appel au concept de *fonction génératrice des probabilités* (f.g.p.) définie par :

$$G_X(z) \stackrel{\text{déf}}{=} \mathbb{E}z^X$$

pour tout z t.q. z^X soit intégrable. On a $G_X(e^t) = M_X(t)$. Les propriétés principales de G_X sont :

$$\mathbb{P}(X = k) = \frac{G_X^{(k)}}{k!}, \quad G_X^{(k)}(1^-) = \mathbb{E}\left(\frac{X!}{(X-k)!}\right)$$

où $G_X^{(k)}(z) = d^k G_X(z)/dz^k(z)$ et $G_X^{(k)}(1^-) = \lim_{z \uparrow 1} G_X^{(k)}(z)$, notamment :

$$\mathbb{E}(X) = G_X'(1^-),$$

$$\text{Var}(X) = G_X''(1^-) + G_X'(1^-) - [G_X'(1^-)]^2.$$

7 Convergence de suites de variables aléatoires

Trois modes de convergence

La notion de convergence simple est peu utile. On présente maintenant les 3 modes classiques de convergence. On considère des v.a.r. $(X_n)_{n \in \mathbb{N}}$ et X :

(i) X_n converge *presque sûrement* vers X , noté $X_n \xrightarrow{\text{p.s.}} X$, lorsque :

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1.$$

(ii) X_n converge *en moyenne d'ordre p* vers X ou dans $L^p(\Omega)$, noté $X_n \xrightarrow{L^p} X$, lorsque :

$$\mathbb{E}(|X_n - X|^p) \xrightarrow{n \rightarrow \infty} 0.$$

(iii) X_n converge *en probabilité* vers X , noté $X_n \xrightarrow{\text{proba.}} X$, lorsque :

$$\mathbb{P}(|X_n - X| > \varepsilon) \xrightarrow{n \rightarrow \infty} 0, \quad \forall \varepsilon > 0.$$

Théorèmes de convergence

Nous avons les résultats de convergence suivants :

Lemme de Fatou^w

Si $X_n \geq 0$ p.s., alors $\mathbb{E} \liminf X_n \leq \liminf \mathbb{E}X_n$. Si $X_n \leq 0$ p.s., alors $\mathbb{E} \limsup X_n \geq \limsup \mathbb{E}X_n$.

Théorème de convergence monotone^w

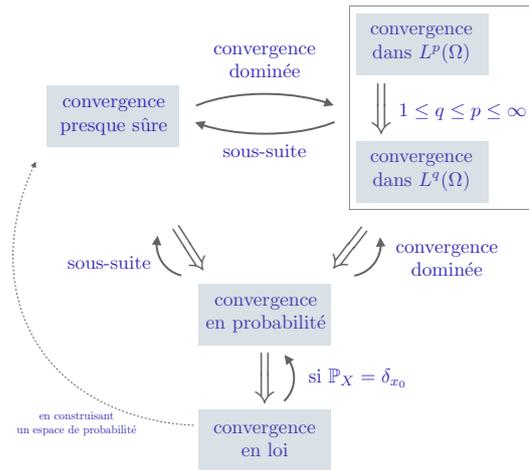
Si $0 \leq X_n \uparrow X$ p.s., alors $\mathbb{E}X_n \uparrow \mathbb{E}X$. De même, si $0 \geq X_n \downarrow X$ p.s., alors $\mathbb{E}X_n \downarrow \mathbb{E}X$.

Ces propositions restent valides lorsque l'on remplace $X_n \geq 0$ (resp. $X_n \leq 0$) p.s. par $X_n \geq Y$ (resp. $X_n \leq Y$) p.s. où Y est intégrable.

Théorème de convergence dominée^w

Si $X_n \xrightarrow{\text{p.s.}} X$ et s'il existe une v.a.r. intégrable Y t.q. $|X_n| \leq Y$ p.s. pour tout n , alors X_n et X sont intégrables et $\mathbb{E}X_n \rightarrow \mathbb{E}X$.

Liens entre les différents types de convergence



- Si $\exists p \geq 1$ t.q. $X_n \xrightarrow{L^p} X$ alors $X_n \xrightarrow{\text{proba.}} X$.
- Si $X_n \xrightarrow{\text{p.s.}} X$ alors $X_n \xrightarrow{\text{proba.}} X$.
- Si $X_n \xrightarrow{\text{proba.}} X$ alors il existe une sous-suite de X_n qui converge p.s. vers X .
- Si $X_n \xrightarrow{L^p} X$ alors il existe une sous-suite de X_n qui converge p.s. vers X .
- $X_n \xrightarrow{\text{proba.}} X$ si et seulement si de toute sous-suite de $(X_n)_{n \in \mathbb{N}}$, on peut extraire une sous-suite qui converge p.s. vers X .
- *Convergence dominée* : Si $X_n \xrightarrow{\text{proba.}} X$ et s'il existe Z intégrable t.q. $|X_n| \leq Z$ p.s. pour un certain $p \geq 1$, alors $X_n \xrightarrow{L^p} X$.
- Soit $(X_n)_{n \in \mathbb{N}}$ et X des v.a. à valeurs dans \mathbb{R}^d , si l'une des conditions suivantes est satisfaite :
 - $\forall \varepsilon > 0, \sum_{n \geq 1} \mathbb{P}(|X_n - X| > \varepsilon) < \infty$;
 - $\exists p \geq 1$ t.q. $\sum_{n \geq 1} \mathbb{E}|X_n - X|^p < \infty$
 alors $X_n \xrightarrow{\text{p.s.}} X$.

8 Convergence en loi

Une suite $(X_n)_{n \in \mathbb{N}}$ de v.a. à valeurs dans \mathbb{R}^d converge *en loi* vers X , noté $X_n \xrightarrow{\text{loi}} X$, lorsque :

$$\mathbb{E}\phi(X_n) \xrightarrow{n \rightarrow \infty} \mathbb{E}\phi(X)$$

pour toute fonction $\phi \in \mathcal{C}_b(\mathbb{R}^d)$ (fonctions continues bornées de \mathbb{R}^d dans \mathbb{R}). Contrairement aux autres modes de convergence, la convergence en loi est une convergence des lois de probabilité et non pas des v.a. elles-mêmes.

Il est clair que la convergence en probabilité implique la convergence en loi et que la réciproque est fautive. Cette dernière est donc très faible mais pourtant :

Théorème de représentation de Skorokhod^w

Sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, on considère des v.a. X_n et X à valeurs dans \mathbb{R}^d , on suppose que $X_n \rightarrow X$ en loi. Alors il existe un espace de probabilité $(\Omega', \mathcal{F}', \mathbb{P}')$ et des v.a. X'_n de même loi que X_n et X' de même loi que X définies sur cet espace et à valeurs dans \mathbb{R}^d telles que $X'_n \rightarrow X'$ presque sûrement.

Ainsi la convergence en loi “enveloppe” les autres modes de convergence. Le théorème est relativement simple à vérifier, en dimension 1 il suffit en effet de poser $F_n(x) = \mathbb{P}(X_n \leq x)$ et $X'_n = F^{-1}(U)$ où $U \sim U[0, 1]$ et $F_n^{-1}(y) = \inf\{x : F_n(x) \geq y\}$.

On considère des mesures de probabilité $Q_n, n \in \mathbb{N}$, et Q sur \mathbb{R}^d . On dit que Q_n converge étroitement vers Q , noté $Q_n \Rightarrow Q$, lorsque :

$$\int \phi(x) Q_n(dx) \xrightarrow{n \rightarrow \infty} \int \phi(x) Q(dx), \quad \forall \phi \in \mathcal{C}_b(\mathbb{R}^d) \quad (4)$$

Ainsi la convergence en loi de v.a. correspond à la convergence étroite des lois de probabilité associées. La convergence en probabilité implique la convergence en loi (la convergence en loi est la plus faible des convergences).

Si $X_n \xrightarrow{\text{loi}} X$, on ne peut pas appliquer le résultat précédent à $f(x) = \mathbf{1}_A(x)$, qui n'est pas continue, et ainsi montrer que $\mathbb{P}(X_n \in A) \rightarrow \mathbb{P}(X \in A)$. Ce résultat est vrai pour les ensembles de \mathbb{P}_X -continuité, i.e. les ensembles $A \in \mathcal{B}(\mathbb{R}^d)$ t.q. $\mathbb{P}_X(\partial A) = 0$ où $\partial A = \bar{A} \setminus \overset{\circ}{A}$. En effet, il y a équivalence entre :

- $X_n \rightarrow X$ en loi ;
- $F_{X_n}(x) \rightarrow F_X(x) \forall x$ où F_X est continue ;
- $\mathbb{P}_{X_n}(A) \rightarrow \mathbb{P}_X(A) \forall A \in \mathcal{B}(\mathbb{R}^d)$ de \mathbb{P}_X -continuité ;
- $\underline{\lim} \mathbb{P}_{X_n}(A) \geq \mathbb{P}_X(A) \forall A \in \mathcal{B}(\mathbb{R}^d)$ ouvert ;
- $\overline{\lim} \mathbb{P}_{X_n}(A) \leq \mathbb{P}_X(A) \forall A \in \mathcal{B}(\mathbb{R}^d)$ fermé ;
- $\Phi_{X_n}(t) \rightarrow \Phi_X(t)$ pour tout $t \in \mathbb{R}^d$.

Loi des grands nombres^w

Soit $(X_n)_{n \in \mathbb{N}}$ une suite de v.a. i.i.d. à valeurs dans \mathbb{R}^d , indépendantes et de même loi. Si $X \in L^1$, alors :

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow[n \rightarrow \infty]{\text{proba.}} \mathbb{E}X_1, \quad (\text{loi faible des grands nombres})$$

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow[n \rightarrow \infty]{\text{p.s.}} \mathbb{E}X_1. \quad (\text{loi forte des grands nombres})$$

Si de plus $X \in L^2$, on a aussi la convergence dans $L^2(\Omega)$.

Ces résultats restent vrais sous des hypothèses bien plus faibles. Par exemple soit $(X_n)_{n \in \mathbb{N}}$ une suite de v.a. à valeurs dans \mathbb{R}^d , de carré intégrable et non corrélées ; on pose $\mu_i = \mathbb{E}X_i$ et $\sigma_i^2 = \text{Var}(X_i)$: si $\frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 \rightarrow 0$, alors $\frac{1}{n} \sum_{i=1}^n (X_i - \mu_i) \rightarrow 0$ en probabilité.

Théorème central limite^w

D'après la loi forte des grands nombres, la moyenne empirique des X_i converge p.s. vers $\mathbb{E}X_1$. La question importante est de savoir à quelle vitesse cette convergence a lieu ? Soit $(X_n)_{n \in \mathbb{N}}$ une suite de v.a. à valeurs dans \mathbb{R}^d , de carré intégrable, indépendants et de même loi, alors :

$$\frac{X_1 + \dots + X_n - n \mathbb{E}X_1}{\sqrt{n}} \xrightarrow[n \rightarrow \infty]{\text{loi}} \mathcal{N}(0, \text{Cov}(X_1)). \quad (\text{théorème central limite})$$

Ce théorème explique l'ubiquité de la loi gaussienne.

4. Si cette convergence est vérifiée $\forall \phi \in \mathcal{C}_0(\mathbb{R}^d)$ (resp. $\forall \phi \in \mathcal{C}_K(\mathbb{R}^d)$), on parlera alors de convergence faible (resp. vague).

9 Probabilités conditionnelles d'événements

On considère un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et $B \in \mathcal{F}$ t.q. $\mathbb{P}(B) > 0$. La probabilité conditionnelle est définie par :

$$\mathbb{P}(A|B) \stackrel{\text{déf}}{=} \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Ainsi $A \perp B$ si et seulement si $\mathbb{P}(A|B) = \mathbb{P}(A)$ ou $\mathbb{P}(B|A) = \mathbb{P}(B)$. À B fixé, l'application $A \mapsto \mathbb{P}(A|B)$ est une probabilité sur (Ω, \mathcal{F}) . Pour tout $A \in \mathcal{F}$ t.q. $B \subset A$, on a $\mathbb{P}(A|B) = 1$. On peut également écrire :

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

pour tout $B \in \mathcal{F}$. Avec la convention $0 \times \text{“terme non défini”} = 0$ on peut décomposer un événement joint par un produit de probabilités conditionnelles :

$$\begin{aligned} \mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_n) &= \mathbb{P}(A_1) \mathbb{P}(A_2|A_1) \\ &\quad \mathbb{P}(A_3|A_1 \cap A_2) \dots \mathbb{P}(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1}). \end{aligned} \quad (\text{théorème des probabilités composées})$$

Soit B_1, \dots, B_n une suite d'événements formant une partition de Ω (i.e. les B_i sont deux à deux disjoints et $\Omega = \cup_{i=1}^n B_i$) et t.q. $\mathbb{P}(B_i) > 0$ alors :

$$\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A|B_i) \mathbb{P}(B_i), \quad \forall A \in \mathcal{F}. \quad (\text{formule des probabilités totales})$$

Si de plus $\mathbb{P}(A) > 0$, on en déduit que :

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j) \mathbb{P}(B_j)}{\sum_{i=1}^n \mathbb{P}(A|B_i) \mathbb{P}(B_i)}. \quad (\text{formule de Bayes})$$

qui permet de déterminer la probabilité des causes (i.e. étant données les probabilités de l'effet A sachant toutes les causes possibles B_j , quel est la probabilité d'une cause B_j sachant l'effet A).

10 Espérance conditionnelle

On considère un événements B t.q. $\mathbb{P}(B) > 0$. L'espérance conditionnelle d'une v.a.r. X (intégrable ou positive) sachant B , notée $\mathbb{E}(X|B)$, est l'espérance de X sous la probabilité conditionnelle $A \mapsto \mathbb{P}(A|B)$:

$$\mathbb{E}(X|B) = \int_{\Omega} X(\omega) \mathbb{P}(d\omega|B) = \frac{\mathbb{E}(X \mathbf{1}_B)}{\mathbb{P}(B)}$$

Il s'agit de la valeur moyenne de X lorsque B est réalisé ; c'est-à-dire la meilleur estimation de X sachant que B est réalisé. On va définir l'espérance conditionnelle d'une (fonction d'une) v.a. X sachant (i) un événement, (ii) une tribu, (iii) une autre variable aléatoire Y ou encore sachant que (iv) $Y = y$. Dans les cas (i) et (iv) il s'agira d'un réel, dans les cas (ii) et (iii) il s'agira d'une v.a.r.

Nous traitons par ordre de difficulté : le cas discret, le cas à densité et le cas général.

10.1 Cas discret

On considère deux v.a. X et Y à valeurs dans \mathbb{N} (le cas fini se traite de la même façon) de loi :

$$\mathbb{P}(X = i, Y = j), \quad (i, j) \in \mathbb{N}^2$$

et de loi marginale en Y :

$$\mathbb{P}(Y = j) = \sum_{i \in \mathbb{N}} \mathbb{P}(X = i, Y = j), \quad j \in \mathbb{N}.$$

La *probabilité conditionnelle de $X = i$ sachant $Y = j$* , i.e. la probabilité de l'événement $\{X = i\}$ sachant que l'on a observé $Y = j$, est définie par :

$$\begin{aligned} \mathbb{P}(X = i|Y = j) &\stackrel{\text{déf}}{=} \mathbb{P}(X = i, Y = j) / \mathbb{P}(Y = j) \\ &= \mathbb{P}(X = i, Y = j) / \sum_{i' \in \mathbb{N}} \mathbb{P}(X = i', Y = j). \end{aligned}$$

Cette définition est valide lorsque $\mathbb{P}(Y = j) > 0$. Lorsque $\mathbb{P}(Y = j) = 0$ cela signifie que l'événement $\{Y = j\}$ n'est jamais observé et donc la probabilité conditionnelle sachant cet événement n'a pas d'intérêt. En fait lorsque $\mathbb{P}(Y = j) = 0$ on peut poser $\mathbb{P}(X = i|Y = j) \stackrel{\text{déf}}{=} \rho(i)$ où ρ est une loi de probabilité quelconque : on a donc défini la probabilité conditionnelle $\mathbb{P}(X = i|Y = j)$ y compris lorsque $\mathbb{P}(Y = j) = 0$.

On peut aussi définir, avant toute observation spécifique, la *probabilité conditionnelle de X sachant Y* :

$$\mathbb{P}(X = i|Y) \stackrel{\text{déf}}{=} \Psi(i, Y) \text{ où } \Psi(i, j) \stackrel{\text{déf}}{=} \mathbb{P}(X = i|Y = j).$$

Il s'agit donc d'une v.a. puisqu'elle dépend de Y qui est aléatoire. À nouveau $i \mapsto \mathbb{P}(X = i|Y)$ est une mesure de probabilité mais cette fois elle est aléatoire (le concept de mesure aléatoire est a priori un peu "dérangeant", mais il est relativement naturel!).

On définit l'*espérance conditionnelle* d'une fonction $\phi(X)$ sachant respectivement $Y = j$ (à j fixé) et Y comme les espérances associées aux deux mesures de probabilité $i \mapsto \mathbb{P}(X = i|Y = j)$ et $i \mapsto \mathbb{P}(X = i|Y)$ (qui est aléatoire) :

$$\begin{aligned} \mathbb{E}(\phi(X)|Y = j) &\stackrel{\text{déf}}{=} \sum_{i \in \mathbb{N}} \phi(i) \mathbb{P}(X = i|Y = j), \\ \mathbb{E}(\phi(X)|Y) &\stackrel{\text{déf}}{=} \sum_{i \in \mathbb{N}} \phi(i) \mathbb{P}(X = i|Y). \end{aligned}$$

Par ailleurs :

$$i \mapsto \mathbb{P}(X = i|Y = j), \quad i \mapsto \mathbb{P}(X = i|Y)$$

sont appelées les *lois conditionnelles* de X sachant respectivement $Y = j$ et Y .

On peut *décomposer une loi jointe par un produit de lois conditionnelles* :

$$\begin{aligned} \mathbb{P}(X_1 = i_1, \dots, X_n = i_n) &= \mathbb{P}(X_1 = i_1) \\ &\times \mathbb{P}(X_2 = i_2|X_1 = i_1) \\ &\times \mathbb{P}(X_3 = i_3|X_1 = i_1, X_2 = i_2) \times \dots \\ &\dots \times \mathbb{P}(X_n = i_n|X_1 = i_1, \dots, X_{n-1} = i_{n-1}). \end{aligned}$$

10.2 Cas à densité

On considère deux v.a. X et Y à valeurs dans \mathbb{R}^d et $\mathbb{R}^{d'}$ de densité jointe $p(x, y)$. On considère la densité marginale de X :

$$p_Y(y) \stackrel{\text{déf}}{=} \int_{\mathbb{R}^d} p(x, y) dx.$$

La *densité conditionnelle de X sachant $Y = y$* est définie par :

$$p_{X|Y=y}(x) \stackrel{\text{déf}}{=} \frac{p(x, y)}{p_Y(y)} = \frac{p(x, y)}{\int_{\mathbb{R}^d} p(x', y') dx'}.$$

qui est défini pour les y t.q. $p_Y(y) > 0$. Lorsque $p_Y(y) = 0$ on choisit de poser $p_{X|Y=y}(x) = \rho(x)$ où $\rho(x)$ est une densité quelconque sur \mathbb{R}^d . Ainsi $p_{X|Y=y}(x)$ est définie pour tout y . Cette définition s'appuie sur le fait que pour Δ petit :

$$\begin{aligned} \mathbb{P}(X \leq x|y \leq Y \leq y + \Delta) &= \frac{\mathbb{P}(X \leq x, y \leq Y \leq y + \Delta)}{\mathbb{P}(y \leq Y \leq y + \Delta)} \\ &= \frac{\int_{-\infty}^x \int_y^{y+\Delta} p(x', y') dx' dy'}{\int_y^{y+\Delta} p_Y(y') dy'} \\ &\simeq \frac{\int_{-\infty}^x p(x', y) dx' \Delta}{p_Y(y) \Delta} = \int_{-\infty}^x p_{X|Y=y}(x') dx'. \end{aligned}$$

Posons $\Psi(x, y) = p_{X|Y=y}(x)$, on définit la *densité conditionnelle de X sachant Y* par :

$$p_{X|Y}(x) \stackrel{\text{déf}}{=} \Psi(x, Y),$$

il s'agit d'une *densité aléatoire*. On définit les *espérances conditionnelles* associées à ces deux densités conditionnelles :

$$\begin{aligned} \mathbb{E}(\phi(X)|Y = y) &\stackrel{\text{déf}}{=} \int_{\mathbb{R}^d} \phi(x) p_{X|Y=y}(x) dx, \\ \mathbb{E}(\phi(X)|Y) &\stackrel{\text{déf}}{=} \int_{\mathbb{R}^d} \phi(x) p_{X|Y}(x) dx \end{aligned}$$

la première est un réel déterministe, la seconde une v.a.r.

10.3 Cas général

Dans le cas général on introduit usuellement en premier l'espérance conditionnelle d'une v.a.r. X sachant une sous-tribu \mathcal{G} de \mathcal{F} ; on en déduira la probabilité conditionnelle et la loi conditionnelle.

On considère une v.a.r. X intégrable (resp. positive) et une sous-tribu \mathcal{G} de \mathcal{F} . Il existe p.s. une unique v.a.r., appelée *espérance conditionnelle* de X sachant Y et notée $\mathbb{E}(X|\mathcal{G})$, t.q. $\omega \mapsto \mathbb{E}(X|\mathcal{G})(\omega)$ soit \mathcal{G} -mesurable et :

$$\int_B \mathbb{E}(X|\mathcal{G}) d\mathbb{P} = \int_B X d\mathbb{P}, \quad \forall B \in \mathcal{G}.$$

Lorsque \mathcal{G} est engendrée par une variable Y , i.e. $\mathcal{G} = \sigma(Y)$, on note alors $\mathbb{E}(X|Y) = \mathbb{E}(X|\mathcal{G})$. Dans ce cas une définition équivalente est : $\omega \mapsto \mathbb{E}(X|Y)(\omega)$ est $\sigma(Y)$ -mesurable et

$$\mathbb{E}[\mathbb{E}(X|\mathcal{G}) \phi(Y)] = \mathbb{E}[X \phi(Y)]$$

pour toute fonction ϕ réelle, mesurable et bornée. Cette définition est naturellement cohérente avec les définitions précédentes dans le cas discret et le cas à densité.

Propriétés

L'espérance conditionnelle se comporte comme une espérance :

- L'application $X \mapsto \mathbb{E}(X|\mathcal{G})$ est linéaire.
- $|\mathbb{E}(X|\mathcal{G})| \leq \mathbb{E}(|X||\mathcal{G})$ p.s. donc $\mathbb{E}|\mathbb{E}(X|\mathcal{G})| \leq \mathbb{E}|X|$.
- $X \leq X'$ p.s. $\Rightarrow \mathbb{E}(X|\mathcal{G}) \leq \mathbb{E}(X'|\mathcal{G})$ p.s.
- Soit ϕ convexe et une v.a. X intégrable. Supposons soit $\phi \geq 0$, soit $\phi(X) \in L^1$, alors $\phi(\mathbb{E}(X|\mathcal{G})) \leq \mathbb{E}(\phi(X)|\mathcal{G})$ p.s.

On retrouve également des résultats de convergence :

Théorème de convergence monotone conditionnelle

Si $0 \leq X_n \uparrow X$ alors :

$$\mathbb{E}(X|\mathcal{G}) = \lim_{n \rightarrow \infty} \uparrow \mathbb{E}(X_n|\mathcal{G}) \quad \text{p.s.}$$

Lemme de Fatou conditionnel

Si $(X_n)_{n \in \mathbb{N}}$ est une suite de v.a. positives, alors :

$$\mathbb{E}(\liminf_{n \rightarrow \infty} X_n|\mathcal{G}) \leq \liminf_{n \rightarrow \infty} \mathbb{E}(X_n|\mathcal{G}) \quad \text{p.s.}$$

Théorème de convergence dominée conditionnelle

Si $X_n \xrightarrow{\text{p.s.}} X$ et s'il existe Z intégrable t.q. $|X_n| \leq Z$ p.s. pour tout n , alors X et X_n sont intégrables et :

$$\mathbb{E}(X_n|\mathcal{G}) \xrightarrow{\text{p.s.}} \mathbb{E}(X|\mathcal{G}) \quad \text{p.s. et dans } L^1.$$

On a enfin des propriétés spécifiques à l'espérance conditionnelle :

- Si X est \mathcal{G} -mesurable, alors $\mathbb{E}(X|\mathcal{G}) = X$ p.s.
- Si $X \perp \mathcal{G}$, alors $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$ p.s., c'est le cas en particulier lorsque $\mathcal{G} = \{\emptyset, \Omega\}$.
- $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$.
- On considère deux sous-tribus $\mathcal{H} \subset \mathcal{G}$, alors

$$\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(\mathbb{E}(X|\mathcal{H})|\mathcal{G}) = \mathbb{E}(X|\mathcal{H}) \quad \text{p.s.}$$

- Supposons que X et Y sont soit intégrables soit positives, si Y est \mathcal{G} -mesurable, alors :

$$\mathbb{E}(XY|\mathcal{G}) = Y \mathbb{E}(X|\mathcal{G}).$$

- On considère deux v.a. X et Y à valeurs dans E_1 et E_2 . Si $X \perp \mathcal{G}$ et Y \mathcal{G} -mesurable, alors pour toute fonction $\phi : E_1 \times E_2 \mapsto \mathbb{R}$ (bornée ou positive)

$$\mathbb{E}(\phi(X, Y)|\mathcal{G}) = \int_{E_1} \phi(x, Y) \mathbb{P}_X(dx)$$

i.e. dans $\mathbb{E}(\phi(X, Y)|\mathcal{G})$, Y se comporte comme une constante et que \mathcal{G} n'apporte aucune information sur X , ainsi la meilleure approximation de $\phi(X, Y)$ basée sur \mathcal{G} consiste à intégrer $x \rightarrow \phi(x, Y)$ (Y gelé) par rapport à la loi de X .

L'espérance conditionnelle comme projecteur

Dans le cas des variables de carré intégrable, l'espérance conditionnelle peut s'interpréter comme un projecteur. Si \mathcal{G} est une sous-tribu de \mathcal{F} , $L^2(\Omega, \mathcal{G}, \mathbb{P})$ s'identifie au sous-espace fermé de $L^2(\Omega, \mathcal{F}, \mathbb{P})$ des éléments de $L^2(\Omega, \mathcal{F}, \mathbb{P})$ dont un représentant est \mathcal{G} -mesurable.

Si $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ et \mathcal{G} une sous-tribu de \mathcal{F} , alors $\mathbb{E}(X|\mathcal{G})$ est la *projection orthogonale* de X sur $L^2(\Omega, \mathcal{G}, \mathbb{P})$. Cette propriété permet de voir que si X est de

carré intégrable alors $\mathbb{E}(X|\mathcal{G})$ est la meilleure approximation au sens du *risque quadratique* par une v.a. \mathcal{G} -mesurable :

$$\mathbb{E}(|X - \mathbb{E}(X|\mathcal{G})|^2) \leq \mathbb{E}(|X - Z|^2), \quad \forall Z \in L^2(\Omega, \mathcal{G}, \mathbb{P}).$$

Probabilités conditionnelles sachant une tribu \mathcal{G}

Comme nous avons défini l'espérance conditionnelle de X sachant une sous-tribu \mathcal{G} de \mathcal{F} , on peut définir la *probabilité conditionnelle* d'un événement sachant \mathcal{G} par :

$$\mathbb{P}(A|\mathcal{G}) = \mathbb{E}(\mathbf{1}_A|\mathcal{G}) \quad \forall A \in \mathcal{F}$$

mais aussi :

$$\mathbb{P}(X \in B|\mathcal{G}) = \mathbb{E}(\mathbf{1}_{\{X \in B\}}|\mathcal{G}), \quad \forall B \in \mathcal{E}$$

ainsi

$$B \mapsto \mathbb{P}_{X|\mathcal{G}}(B) = \mathbb{P}(X \in B|\mathcal{G})$$

permet de définir la *loi conditionnelle de X sachant \mathcal{G}* et :

$$\mathbb{E}(X|\mathcal{G}) = \int_{\mathbb{R}^d} x \mathbb{P}_{X|\mathcal{G}}(dx)$$

$$\mathbb{E}(\phi(X)|\mathcal{G}) = \int_E \phi(x) \mathbb{P}_{X|\mathcal{G}}(dx)$$

Espérance et probabilités conditionnelles sachant un événement C

On peut définir l'espérance et la probabilité conditionnelles par un événement $C \in \mathcal{F}$ comme étant l'espérance et la probabilité conditionnelles par la sous-tribu $\mathcal{G} = \{C, C^c, \Omega, \emptyset\}$ engendrée par C . Dans ce cas la probabilité conditionnelle s'écrit simplement :

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

qui est naturellement la définition initiale! De même :

$$\mathbb{P}_{X|C}(B) = \mathbb{P}(X \in B|C)$$

$$= \frac{\mathbb{P}(\{X \in B\} \cap C)}{\mathbb{P}(C)} = \frac{\mathbb{E}(\{X \in B\} \cap C)}{\mathbb{P}(C)}$$

$$\mathbb{E}(\phi(X)|B) = \int_E \phi(x) \mathbb{P}_{X|C}(dx) = \mathbb{P}(X \in B|C)$$

$$= \frac{\mathbb{P}(\{X \in B\} \cap C)}{\mathbb{P}(C)} = \frac{\mathbb{E}(\{X \in B\} \cap C)}{\mathbb{P}(C)}.$$

Dans ce cas particulier l'espérance et la probabilité conditionnelles sont déterministes.

Espérance et probabilités conditionnelles sachant un v.a. Y

L'espérance et la probabilité conditionnelles par Y , v.a. à valeurs dans (E', \mathcal{E}') , sont l'espérance et la probabilité conditionnelles par $\mathcal{G} = \sigma(Y)$ la sous-tribu engendrée par Y .

Espérance et probabilités conditionnelles sachant $Y = y$

Il existe une fonction mesurable $\Psi : E' \mapsto \mathbb{R}$ t.q. $\mathbb{E}(X|Y) = \Psi(Y)$ et on pose :

$$\mathbb{E}(X|Y = y) = \Psi(y).$$

On peut aussi définir :

$$\mathbb{P}(X \in B|Y = y) \stackrel{\text{déf}}{=} \mathbb{E}(\mathbf{1}_{\{X \in B\}}|Y = y).$$

À nouveau cette définition est cohérente avec les définitions précédentes notamment avec celle du cas des variables à densité puisque :

$$\mathbb{P}(X \in B|Y = y) = \int_B p_{X|Y=y}(x) dx.$$

Cette “pirouette” pour définir la probabilité conditionnelle sachant $\{Y = y\}$ est nécessaire car dans le cas d’une variable Y de loi diffuse (i.e. n’admettant pas d’atome) l’événement $\{Y = y\}$ est négligeable et la définition classique $\mathbb{P}(X \in B|Y = y) = \mathbb{P}(\{X \in B\} \cap \{Y = y\})/\mathbb{P}(Y = y)$ n’a pas de sens.

On a :

$$\begin{aligned} \mathbb{E}(X) &= \int_{E'} \mathbb{E}(X|Y = y) \mathbb{P}_Y(dy), \\ \mathbb{P}(X \in B) &= \int_{E'} \mathbb{P}(X \in B|Y = y) \mathbb{P}_Y(dy). \end{aligned}$$

On peut décomposer les lois jointes selon le produit de lois conditionnelles de la façon suivante :

$$\begin{aligned} \mathbb{P}_{X,Y}(dx, dy) &= \mathbb{P}_{X|Y=y}(dx) \mathbb{P}_Y(dy) \\ &= \mathbb{P}_{Y|X=x}(dy) \mathbb{P}_X(dx) \end{aligned}$$

et de façon générale :

$$\begin{aligned} \mathbb{P}_{X_1, X_2, \dots, X_n}(dx_1, dx_2, \dots, dx_n) &= \mathbb{P}_{X_1}(dx_1) \\ &\times \mathbb{P}_{X_2|X_1=x_1}(dx_2) \\ &\times \mathbb{P}_{X_3|X_1=x_1, X_2=x_2}(dx_3) \times \dots \\ &\dots \times \mathbb{P}_{X_n|X_1=x_1, \dots, X_{n-1}=x_{n-1}}(dx_n). \end{aligned}$$

11 Quelques lois de probabilité usuelles

Nous traitons successivement du cas fini, du cas discret, du cas continue et enfin de la loi gaussienne.

11.1 Cas fini

Loi de Bernoulli^w

La loi de probabilité d’une v.a. X est dite de *Bernoulli* (ou la variable X est dite de Bernoulli) de paramètre $p \in [0, 1]$, noté $\mathcal{B}(p)$, lorsque X prend deux valeurs, disons 0 et 1, la probabilité d’obtenir 1 est p , la probabilité d’obtenir 0 est $1 - p$. Ainsi :

$$\mathbb{P}(X = x) = p^x (1 - p)^{1-x}, \quad x \in E = \{0, 1\}.$$

Loi binomiale^w

Le modèle binomial de paramètres $n \in \mathbb{N}_*$ et $p \in [0, 1]$ consiste à répéter n tirages indépendants de Bernoulli. On note (X_1, \dots, X_n) le vecteur aléatoire correspondant, ainsi :

$$\begin{aligned} \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) &= \prod_{i=1}^n p^{x_i} (1 - p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i} (1 - p)^{n - \sum_{i=1}^n x_i} \end{aligned}$$

pour tout $x_1, \dots, x_n \in \{0, 1\}$.

La loi de la variable $S_n = X_1 + \dots + X_n$ consistant à compter le nombre de symboles 1 dans (X_1, \dots, X_n) est alors dite *binomiale*, notée $\mathcal{B}(n, p)$ (on dit également que la variable est binomiale). Ainsi S_n prend ses valeurs dans $\{0, \dots, n\}$ et :

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

D’après la formule du binôme de Newton, il s’agit bien d’une probabilité. La moyenne et la variance sont respectivement np et $np(1 - p)$.

En pratique, pour n grand, le calcul de $\binom{n}{k} p^k q^{n-k}$ est impossible, il est alors nécessaire de faire appel aux logarithmes et d’utiliser la *formule de Stirling* ($n! \simeq \sqrt{2\pi n} (n/e)^n$). Lorsque $n \rightarrow \infty$, il existe deux asymptotiques :

- si $p \downarrow 0$ avec $np = a$, alors la loi binomiale converge vers une loi de Poisson de paramètre a . En pratique, on remplace la loi binomiale par une loi de Poisson dès que $n > 30$ et $np < 5$.
- si $p \simeq 1 - p$, la loi binomiale converge vers une loi normale d’espérance np et de variance $np(1 - p)$. En pratique, on remplace une loi binomiale par une loi normale dès que $n > 30$, $np > 5$ et $n(1 - p) > 5$.

Loi multinomiale^w

On peut formuler le modèle de Bernoulli précédent en posant $N_1 = X$, $N_2 = n - X$, $p_1 = p$, $p_2 = 1 - p$. Alors :

$$\mathbb{P}(N_1 = n_1, N_2 = n_2) = \frac{n!}{n_1! n_2!} p_1^{n_1} p_2^{n_2}.$$

On peut généraliser ce modèle en supposant que l’on réalise n tirages indépendants X_i de même loi à valeurs dans $\{1, \dots, r\}$. Notons p_k la probabilité d’obtenir la valeur k . On a donc $p_k \geq 0$ et $\sum_{k=1}^r p_k = 1$, et :

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = p_1^{\nu_1(x)} \dots p_r^{\nu_r(x)}$$

avec $x = (x_1, \dots, x_n) \in E = \{1, \dots, r\}^n$ et $\nu_k(x) = \sum_{i=1}^n \mathbf{1}_{\{x_i=k\}}$.

On définit le nombre de symboles “ k ” dans la suite $X = (X_1, \dots, X_n)$ par :

$$N_k \stackrel{\text{déf}}{=} \sum_{i=1}^n \mathbf{1}_{\{X_i=k\}}$$

ce nombre est compris entre 0 et n . La loi *multinomiale* est la loi du r -uplet $N = (N_1, \dots, N_r)$, c’est-à-dire :

$$\mathbb{P}(N_1 = n_1, \dots, N_r = n_r) = \frac{n!}{n_1! \dots n_r!} p_1^{n_1} \dots p_r^{n_r}$$

pour tout $n_1, \dots, n_r \in \{0, \dots, n\}$ t.q. $\sum_{k=1}^r n_k = n$.

Les variables N_k sont binomiales $\mathcal{B}(n, p_k)$ et non indépendantes. On a $\mathbb{E}(N_k) = n p_k$, $\text{Var}(N_k) = n p_k (1 - p_k)$, $\text{Cov}(N_k, N_\ell) = -n p_k p_\ell$.

D'après le théorème central limite et sous certaines conditions :

$$\frac{N_k - n p_k}{\sqrt{n p_k (1 - p_k)}} \xrightarrow[n \rightarrow \infty]{\text{loi}} \mathcal{N}(0, 1).$$

Du fait de la contrainte linéaire des variables N_k , on montre que :

$$\sum_{k=1}^r \frac{(N_k - n p_k)^2}{n p_k}$$

suit approximativement une loi du χ^2 à $(r - 1)$ degrés de liberté.

Loi hypergéométrique^w

Les deux modèles précédents, binomial et multinomial, correspondent à des tirages *avec remise*. Considérons maintenant un cas *sans remise*. Supposons que l'on tire n boules parmi $N > n$. Dans ces N , M sont noires (associées à 1) et $N - M$ sont blanches (associées à 0). On compte le nombre X de boules noires tirées. Il s'agit naturellement de tirages sans remise. On dit que X suit une loi *hypergéométrique* de paramètres N , M et n et :

$$\mathbb{P}(X = k) = \binom{M}{k} \binom{N-M}{n-k} / \binom{N}{n}$$

pour $\max(0, M + n - N) \leq k \leq \min(n, M)$ et 0 sinon.

On peut comprendre cette formule de la façon suivante : il y a $\binom{N}{n}$ tirages sans remise possibles. Il y a $\binom{M}{k}$ façons d'obtenir k boules noires et il y a $\binom{N-M}{n-k}$ façons de compléter le reste de l'échantillon avec des boules blanches. On a :

$$\mathbb{E}X = \frac{nM}{N}, \quad \text{Var}(X) = \frac{n \frac{M}{N} (1 - \frac{M}{N})(N - n)}{N - 1}.$$

Lorsque $N \uparrow \infty$ à $p = M/N$ fixe, alors X converge en loi vers la binomiale $\mathcal{B}(n, p)$. En pratique, on remplace la loi hypergéométrique de paramètres (n, p, N) par une loi binomiale de paramètres (n, p) dès que $10n < N$.

11.2 Cas dénombrable

Loi géométrique^w

On considère une suite d'épreuves de Bernoulli de probabilité de succès p . La loi géométrique est la loi du "temps" qu'il faut pour atteindre un succès. Il y a en fait deux définitions : le plus souvent il s'agira du nombre X d'essais pour atteindre un succès, alors $\mathbb{P}(X = k) = (1 - p)^{k-1} p$ avec $k \in \{1, 2, \dots\}$. Mais parfois il s'agit du nombre X' d'échecs pour atteindre un succès, dans ce cas $\mathbb{P}(X = k) = (1 - p)^k p$ avec $k \in \{1, 2, \dots\}$.

Il s'agit de l'équivalent en temps discret de la loi exponentielle qui présente comme elle une propriété *d'absence de mémoire* :

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

La moyenne et la variance sont $1/p$ et $(1 - p)/p^2$. La fonction caractéristique, les f.g.m. et f.g.p. sont respectivement : $\Phi_X(t) = p e^{it} / [1 - (1 - p) e^{it}]$ pour tout $t \in \mathbb{R}$; $M_X(t) = p e^t / [1 - (1 - p) e^t]$ pour $t < \log(1 - p)$; $G_X(t) = t p / [1 - t(1 - p)]$ pour $|t| < 1/(1 - p)$.

Loi de Poisson^w

Il s'agit d'une loi de probabilité sur \mathbb{N} , définie par :

$$p(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \in \mathbb{N}$$

paramétrée par $\lambda > 0$. Cette loi représente le nombre k d'occurrences d'un événement sur un intervalle de temps donnée et λ est le nombre moyen d'occurrence de cet événement sur cet intervalle. Si par exemple, un événement survient en moyenne toutes les 4 minutes et si vous souhaitez étudier le nombre d'occurrences de cet événement dans un intervalle de 10 minutes, alors $\lambda = 10/4$.

Les moyenne et variance sont λ , le mode est $\lfloor \lambda \rfloor$. La fonction caractéristique, les f.g.p. et f.g.m. sont : $\Phi(t) = \exp(\lambda(e^{it} - 1))$; $G(t) = e^{\lambda(z-1)}$; $M(t) = e^{\lambda(e^t - 1)}$.

Loi binomiale négative^w

Soit X le nombre d'essais de Bernoulli (de paramètre p) nécessaires pour obtenir r succès. X est à valeurs dans $\{r, r + 1, \dots\}$. Soit $Y = X - r$ le nombre d'échecs avant le r -ième succès. Alors :

$$\mathbb{P}(X = k) = \binom{k-1}{r-1} p^r (1 - p)^{k-r}$$

pour $k \geq r$, et

$$\mathbb{P}(Y = k) = \binom{k+r-1}{r-1} p^r (1 - p)^k$$

pour $k \geq 0$. La f.g.p. de X est $G_X(z) = \left(\frac{p}{1 - (1-p)z}\right)^r$ sa f.g.m. est $M_X(t) = \left(\frac{p}{1 - (1-p)e^t}\right)^r$.

11.3 Cas continu

On considère une v.a.r. X de loi de probabilité \mathbb{P}_X sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, de fonction de répartition F_X et, lorsqu'elle existe, de densité p . La *médiane* est un point $m \in \mathbb{R}$ t.q. $\mathbb{P}(X \leq m) \geq \frac{1}{2} \leq \mathbb{P}(X \geq m)$, ou bien $F_X(m) = \int_{-\infty}^m p(x) dx = \frac{1}{2}$. Le *mode* est un point où la densité atteint son maximum.

Loi exponentielle^w

Une v.a.r. X est dite exponentielle de paramètre λ , noté $X \sim \mathcal{E}(\lambda)$, si elle admet la densité :

$$p(x) = \lambda e^{-\lambda x} \mathbf{1}_{[0, \infty[}(x), \quad x \in \mathbb{R}.$$

Sa moyenne est λ^{-1} , sa variance λ^{-2} et sa fonction de répartition :

$$F_X(t) = \mathbb{P}(X \leq t) = 1 - e^{-\lambda t}.$$

Sa fonction caractéristique est $\Phi_X(t) = (1 - \frac{it}{\lambda})^{-1}$.

La loi exponentielle permet de modéliser, en temps continu, les intervalles de temps entre deux événements, par exemple dans le cas du processus de Poisson.

Une propriété importante de la loi exponentielle est son absence de mémoire qui jouera un rôle fondamental dans la théorie des processus de Markov. Cette absence de mémoire est en fait une caractéristique de cette loi :

Absence de mémoire

Soit X une v.a.r. positive t.q. $\mathbb{P}(X > s) > 0$ pour tout $s \in \mathbb{R}^+$, alors X suit une loi exponentielle si et seulement si :

$$\mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s), \quad \forall s, t > 0.$$

Loi uniforme

Soit $a, b \in \mathbb{R}$, $a < b$. On considère une v.a.r. X t.q. $\mathbb{P}(s \leq X \leq t)$ soit proportionnel à $(t - s)$ pour tous $a < s < t < b$. Si X admet une densité p , elle vérifie alors $\int_s^t p(x) dx = \text{cte}(t - s)$, on en déduit que $x \mapsto p(x)$ est une fonction constante et, puisque $\int_a^b p(x) dx = 1$ alors :

$$p(x) = \frac{1}{b - a} \mathbf{1}_{[a, b]}(x) \quad x \in \mathbb{R}.$$

On dit alors que la loi de X est uniforme, notée $X \sim U[a, b]$, i.e. toutes les valeurs entre a et b sont "équiprobables". On calcule la moyenne et la variance de X : $\mathbb{E}X = (a + b)/2$ et $\text{Var}(X) = (b - a)^2/12$; la fonction caractéristique est $\Phi_X(t) = [e^{itb} - e^{ita}]/[it(b - a)]$.

Loi de Cauchy^w

Une v.a.r. X est dite de Cauchy, noté $X \sim \mathcal{C}(x_0, \gamma)$ si elle admet la densité :

$$p(x) = \frac{1}{\pi} \left[\frac{\gamma}{(x - x_0)^2 + \gamma^2} \right], \quad x \in \mathbb{R}$$

où $x_0 \in \mathbb{R}$ est le paramètre de centrage et $\gamma > 0$ celui d'échelle. *La loi de Cauchy n'admet pas de moments!* Sa médiane et son mode sont x_0 . Sa fonction caractéristique est $\Phi_X(t) = \exp(x_0 it - \gamma |t|)$.

Loi Gamma^w

La loi gamma de paramètres $a > 0$ et $\lambda > 0$, notée $G(a, \lambda)$, a pour densité :

$$p(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x} \mathbf{1}_{]0, \infty[}(x), \quad x \in \mathbb{R}$$

où $\Gamma(a) = \int_0^{+\infty} e^{-x} x^{a-1} dx$ est la fonction gamma. Pour n entier, $G(n/2, 1/2)$ est la loi du $\chi^2(k)$. La loi $G(1, \lambda)$ est la loi exponentielle $\mathcal{E}(\lambda)$.

11.4 La loi gaussienne^w

Un vecteur aléatoire réel X à valeurs dans \mathbb{R}^d suit une *loi gaussienne*, noté $X \sim \mathcal{N}(\mu, R)$, lorsque sa fonction caractéristique est de la forme :

$$\Phi_X(\xi) = \mathbb{E} \exp(i \xi^* X) = \exp \left(i \xi^* \mu - \frac{1}{2} \xi^* R \xi \right)$$

où $\xi \in \mathbb{R}^d$, $\mu \in \mathbb{R}^d$ et $R \in \mathbb{R}^{d \times d}$ est une matrice symétrique définie positive⁵; μ est la *moyenne* de X et R sa

5. Une matrice symétrique R est dite définie positive, noté $R \geq 0$, lorsque $x^* R x \geq 0$ pour tout $x \in \mathbb{R}^d$; elle est dite strictement définie positive, noté $R > 0$, lorsque $x^* R x > 0$ pour tout $x \neq 0$.

matrice de covariance. Lorsque $R = I$, la loi gaussienne est dite *standard*. Si R est définie strictement positive, noté $R > 0$, alors X admet la densité suivante :

$$p(x) = \frac{1}{\sqrt{(2\pi)^d |R|}} \exp \left(-\frac{1}{2} (x - \mu)^* R^{-1} (x - \mu) \right)$$

où $|R|$ est le déterminant de la matrice R ; dans ce cas en dimension 1, la densité s'écrit :

$$p(x) = \frac{1}{\sqrt{2\pi R}} \exp \left(-\frac{1}{2R} (x - \mu)^2 \right).$$

Si $R > 0$ n'est pas vérifié la loi est dite dégénérée. En particulier lorsque $R = 0$ alors $\mathbb{P}_X = \mathcal{N}(\mu, 0) = \delta_\mu$ est la mesure de Dirac au point μ , i.e. $\mathbb{E}\phi(X) = \phi(\mu)$ pour toute fonction ϕ mesurable et bornée.

Propriétés :

- Un vecteur aléatoire X à valeurs dans \mathbb{R}^d est gaussien si et seulement s'il existe un vecteur aléatoire gaussien Y à valeurs dans \mathbb{R}^d dont les composantes sont centrées/réduites et $(\mu, \Gamma) \in \mathbb{R}^d \times \mathbb{R}^{d \times d}$ t.q. $X = \mu + \Gamma Y$. Dans ce cas $X \sim \mathcal{N}(\mu, \Gamma \Gamma^*)$.
- Si X est un vecteur aléatoire gaussien alors toutes ses composantes sont des v.a.r. gaussiennes, *la réciproque est fautive*.

Conditionnement de v.a. gaussiennes

On considère des v.a.r. X, Y_1, \dots, Y_n de carré intégrable. Il existe deux méthodes d'approximation de X en fonction de Y_1, \dots, Y_n :

Par conditionnement : l'espérance conditionnelle $\mathbb{E}(X | Y_{1:n})$ est la projection orthogonale de X sur $L^2(\Omega, \sigma(Y_{1:n}), \mathbb{P})$, c'est la meilleure approximation de X , au sens du risque quadratique, de la forme $\phi(Y_{1:n})$.

Par régression linéaire : on peut déterminer la meilleure approximation de X comme fonction affine de $Y_{1:n}$. C'est la projection orthogonale dans l'espace de dimension finie engendré par $\{1, Y_1, \dots, Y_n\}$

Ces deux projections ne coïncident en général pas. Le plus souvent la projection (i) donne une bien meilleure approximation que (ii), mais cette dernière est toujours "calculable" alors que la première nécessite le plus souvent une approximation numérique. *Dans le cas gaussien ces notions coïncident.*

On considère un vecteur aléatoire gaussien (X, Y) de dimension \mathbb{R}^{n+m} :

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \\ Y_1 \\ \vdots \\ Y_m \end{pmatrix} \sim \mathcal{N}(\mu, Q)$$

avec

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad Q = \begin{pmatrix} Q_X & Q_{XY} \\ Q_{YX} & Q_Y \end{pmatrix}$$

et

$$\mu_X = \begin{pmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_n \end{pmatrix}, \quad \mu_Y = \begin{pmatrix} \mathbb{E}Y_1 \\ \vdots \\ \mathbb{E}Y_m \end{pmatrix}$$

ainsi que $[Q_X]_{i,j} = \text{Cov}(X_i, X_j)$, $[Q_Y]_{i,j} = \text{Cov}(Y_i, Y_j)$ et $[Q_{XY}]_{i,j} = [Q_{YX}^*]_{i,j} = \text{Cov}(X_i, Y_j)$. On a :

$$X \perp\!\!\!\perp Y \iff Q_{XY} = 0.$$

i.e. X et Y sont indépendants si et seulement si les termes $\text{Cov}(X_i, Y_j)$ sont tous nuls.

On a des formules explicites pour le calcul de la loi conditionnelle de X sachant $Y = y$:

$$\mathbb{P}_{X|Y=y} = \mathcal{N}(\hat{X}(y), R)$$

avec $\hat{X}(y) \stackrel{\text{def}}{=} \mu_X + Q_{XY} Q_Y^{-1} (y - \mu_Y)$ et $R \stackrel{\text{def}}{=} Q_X - Q_{XY} Q_X^{-1} Q_{YX}$.

Ces deux merveilleux résultats ne sont valables que lorsque le couple (X, Y) est gaussien. Lorsque X et Y sont gaussiens alors le couple n'est pas nécessairement gaussien !

12 Martingales à temps discret

Processus et filtrations

Un processus aléatoire (à temps discret) est une suite $(X_n)_{n \in \mathbb{N}}$ de v.a. définies sur un même espace $(\Omega, \mathcal{F}, \mathbb{P})$ et prenant leurs valeurs dans un même espace d'états (E, \mathcal{E}) .

Une filtration est une famille croissante $(\mathcal{F}_n)_{n \in \mathbb{N}}$ de sous-tribus \mathcal{F}_n de \mathcal{F} ($\mathcal{F}_m \subset \mathcal{F}_n$, pour $n \leq m$). Elle représente une information qui s'accumule au cours du temps : \mathcal{F}_n est l'ensemble des événements survenus avant l'instant n . Le quadruplet $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n \in \mathbb{N}}, \mathbb{P})$ est appelé *espace de probabilité filtré*. La filtration naturelle du processus $(X_n)_{n \in \mathbb{N}}$ est définie par :

$$\mathcal{F}_n^X \stackrel{\text{def}}{=} \sigma(X_k ; 0 \leq k \leq n).$$

Un processus $(X_n)_{n \in \mathbb{N}}$ est dit \mathcal{F}_n -adapté lorsque X_n est \mathcal{F}_n -mesurable pour tout $n \in \mathbb{N}$. Un processus est "naturellement" adapté à sa filtration naturelle.

La loi d'un processus aléatoire $(X_n)_{n \in \mathbb{N}}$ est la donnée de la loi de (X_0, \dots, X_n) pour tout $n \in \mathbb{N}$. La loi de ce $(n+1)$ -uple se décompose selon les marginales conditionnelles :

$$\begin{aligned} \mathbb{P}_{X_0, \dots, X_n}(dx_0, \dots, dx_n) &= \mathbb{P}_{X_0}(dx_0) \mathbb{P}_{X_1|X_0=x_0}(dx_1) \\ &\quad \cdots \mathbb{P}_{X_n|X_0:n-1=x_0:n-1}(dx_n) \end{aligned}$$

où $X_{0:n-1} = x_{0:n-1}$ est une notation compacte pour $X_0 = x_0, \dots, X_{n-1} = x_{n-1}$.

(Sur|Sous)-Martingales^w

Soit $(X_n)_{n \in \mathbb{N}}$ un processus \mathcal{F}_n -adapté t.q. X_n soit intégrable, ce processus est appelé :

- (i) une *martingale* lorsque $\mathbb{E}(X_n | \mathcal{F}_m) = X_m$ p.s.
 - (ii) une *sur-martingale* lorsque $\mathbb{E}(X_n | \mathcal{F}_m) \leq X_m$ p.s.
 - (iii) une *sous-martingale* lorsque $\mathbb{E}(X_n | \mathcal{F}_m) \geq X_m$ p.s.
- pour tous $0 \leq m \leq n$.

Si X_n est une sur-martingale alors $-X_n$ est une sous-martingale. Si X_n est une martingale (resp. sur, resp.

sous) alors $n \mapsto \mathbb{E}(X_n)$ est constante (resp. croissante, resp. décroissante). Cette notion provient des jeux, mais contrairement aux jeux où une martingale est censée rapporter un gain, ici en moyenne la martingale ne rapporte rien mais ne coûte rien non plus.

Exemples :

- Si X est intégrable, alors $X_n = \mathbb{E}(X | \mathcal{F}_n)$ est une martingale.
- Soit Z_n une suite de v.a.r. indépendantes alors $X_n = \sum_{m \leq n} Z_m$ est une \mathcal{F}_n^Z -martingale appelée *marche aléatoire*.

Décomposition de Doob^w

En moyenne, une sous-martingale (resp. sur-martingale) est croissante (resp. décroissante). En fait si X_n est une sous-martingale, alors :

$$X_n = X_0 + M_n + A_n \quad (\text{décomposition de Doob})$$

où :

- (i) M_n est une \mathcal{F}_n -martingale ;
- (ii) A_n est un processus croissant ($A_n \geq A_m$ p.s. pour $n \geq m$), \mathcal{F}_{n-1} -adapté

avec $M_0 = A_0 = 0$, cette décomposition est p.s. unique.

Il est important de noter que A_n est \mathcal{F}_{n-1} -adapté, avec la convention $\mathcal{F}_{-1} = \{\emptyset, \Omega\}$. Cela signifie que A_n est connu à l'instant $n-1$, il est dit *prévisible*. En pratique on calcule $A_n = \mathbb{E}(X_n | \mathcal{F}_{n-1}) - X_{n-1} + A_{n-1}$.

Si X_n est une martingale de carré intégrable alors X_n^2 est une sous-martingale et on définit le *crochet* $\langle X \rangle_n$ de la martingale X_n comme étant le processus croissant prévisible de la décomposition de Doob de X_n^2 ; en particulier $\mathbb{E}\langle X \rangle_n = \mathbb{E}(X_n^2)$.

Temps d'arrêt^w

Une v.a. T à valeurs dans $\mathbb{N} \cup \{\infty\}$ est appelée \mathcal{F}_n -temps d'arrêt (noté t.a.) si $\{T \leq n\} \in \mathcal{F}_n$ pour tout n . On appelle *tribu des événements antérieurs* à T la tribu :

$$\mathcal{F}_T \stackrel{\text{def}}{=} \{A \in \mathcal{F} | A \cap \{T \leq n\} \in \mathcal{F}_n, \forall n \in \mathbb{N}\}.$$

On pose $X_T(\omega) = X_{T(\omega)}(\omega)$ (si $T(\omega) < \infty$ et une valeur quelconque sinon), les v.a. T et X_T sont \mathcal{F}_T -mesurables.

Exemples :

- Si X_n est un processus adapté à valeurs dans \mathbb{N} alors le *premier temps d'atteinte d'un état* $i \in \mathbb{N}$ est le t.a. défini par $T \stackrel{\text{def}}{=} \min\{n \in \mathbb{N} ; X_n = i\}$ avec la convention $\min\{\emptyset\} = +\infty$.
- Attention le premier temps de sortie d'un état, c'est-à-dire $\max\{n \in \mathbb{N} ; X_n = i\}$, n'est pas un t.a. car pour déterminer $\{T \leq n\}$ il faut connaître le futur !
- Le *premier temps d'atteinte d'un ensemble* $B \in \mathcal{E}$ est le t.a. $T \stackrel{\text{def}}{=} \min\{n \in \mathbb{N} ; X_n \in B\}$.
- Les temps d'atteinte de B après T , c'est-à-dire $\min\{n \geq T ; X_n \in B\}$ et $\min\{n > T ; X_n \in B\}$ sont aussi des t.a. par rapport à la même filtration.

Si T et S sont deux t.a. alors $S \vee T$ et $S \wedge T$ sont également des t.a. Si $S \leq T$ p.s. alors $\mathcal{F}_S \subset \mathcal{F}_T$. Une constante c est un t.a. et donc $T \wedge c$ aussi.

Soit $S_n = X_1 + \dots + X_n$ où X_n est une suite i.i.d. et intégrable, alors pour tout \mathcal{F}_n^S -t.a. T :

$$\mathbb{E}(X_1 + \dots + X_T) = \mathbb{E}(T) \mathbb{E}(X_1). \quad (\text{équation de Wald})$$

Martingale arrêtée et théorème d'arrêt de Doob^w

On considère une martingale X_n t.q. $\sup_{n \in \mathbb{N}} \mathbb{E}|X_n| < \infty$ et un t.a. T fini p.s. (i.e. $\mathbb{P}(T < \infty) = 1$) alors la martingale arrêtée X_T est L^1 et $\mathbb{E}|X_T| \leq \sup_{n \in \mathbb{N}} \mathbb{E}|X_n|$.

3 versions du théorème d'arrêt :

- Si X_n est une martingale (resp. sous, resp. sur) et si T_m est une suite croissante et bornée de t.a. alors X_{T_m} est une martingale (resp. sous, resp. sur).
- Si X_n est une martingale (resp. sous, resp. sur), si S et T sont deux t.a. bornés t.q. $S \leq T$ alors $\mathbb{E}(X_T | \mathcal{F}_S) = X_S$ p.s. (resp. \geq , resp. \leq).
- Soit X_n est une martingale (resp. sous, resp. sur) et T est un t.a., si l'une des conditions suivantes est satisfaite :
 - (i) $\exists c < \infty$ t.q. $T \leq c$ p.s. (i.e. T est p.s. bornée) ;
 - (ii) $\exists c < \infty$ t.q. $|X_{T \wedge n}| \leq c$ p.s. pour tout n ;
alors $\mathbb{E}(X_T) = \mathbb{E}(X_0)$ (resp. \geq , resp. \leq).

Inégalités et convergence de martingales

Le théorème d'arrêt permet de démontrer des *inégalités de martingales*, également appelées *inégalités de Doob* ou *maximales*. Par exemple si X_n est une sous-martingale alors :

$$\mathbb{P}\left(\max_{1 \leq k \leq n} X_k \geq c\right) \leq \frac{\mathbb{E}(X_n^+)}{c}, \quad \forall c > 0.$$

Si X_n est une martingale alors $|X_n|$ est une sous-martingale et donc :

$$\mathbb{P}\left(\max_{1 \leq k \leq n} |X_k| \geq c\right) \leq \frac{\mathbb{E}|X_n|}{c}, \quad \forall c > 0.$$

Si de plus X_n est de carré intégrable alors X_n^2 est une sous-martingale et :

$$\mathbb{P}\left(\max_{1 \leq k \leq n} |X_k| \geq c\right) \leq \frac{\mathbb{E}(X_n^2)}{c^2} = \frac{\mathbb{E}\langle X \rangle_n}{c^2}, \quad \forall c > 0.$$

Soit X_n une sous-martingale positive ou une martingale et $p \in]1, \infty[$ alors :

$$\mathbb{E}\left(\max_{1 \leq k \leq n} |X_k|^p\right) \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}(|X_n|^p)$$

que l'on peut aussi noter :

$$\left\| \max_{1 \leq k \leq n} X_k \right\|_p \leq \frac{p}{p-1} \|X_n\|_p.$$

Ces inégalités servent à démontrer des résultats de *convergence de martingales*, comme par exemple la loi des grands nombres puisque $X_n = Y_1 + \dots + Y_n$ est une martingale lorsque les Y_n sont indépendants et intégrables. Elles permettent aussi de démontrer le :

Théorème de convergence des martingales

Si X_n est une martingale (ou une sous, ou une sur) t.q. $\sup_{n \in \mathbb{N}} \mathbb{E}|X_n| < \infty$ alors $\lim_{n \rightarrow \infty} X_n$ existe p.s.

13 Chaînes de Markov à espace d'états dénombrable

13.1 Chaînes de Markov

Un processus aléatoire $(X_n)_{n \in \mathbb{N}}$ défini sur $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans $(E, \mathcal{P}(E))$, E fini ou dénombrable, est appelé *chaîne de Markov* s'il vérifie la *propriété de Markov* suivante :

$$\begin{aligned} \mathbb{P}(X_{n+1} = i_{n+1} | X_{0:n} = i_{0:n}) \\ = \mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n) \end{aligned}$$

pour tout $n \geq 0$ et $i_1, \dots, i_{n+1} \in E$ t.q. $\mathbb{P}(X_{0:n} = i_{0:n}) > 0$, avec la notation : $i_{k:\ell} = (i_k, \dots, i_\ell)$ lorsque $k \leq \ell$ et $i_{k:\ell} = \emptyset$ sinon. E est appelé *espace d'états*.

La probabilité $\mu = \mathbb{P}_{X_0}$ est appelée la *loi initiale* et :

$$P_{ij}(n) \stackrel{\text{déf}}{=} \mathbb{P}(X_{n+1} = j | X_n = i), \quad i, j \in E$$

la *probabilité de transition* de la chaîne de Markov. Lorsque $\mathbb{P}(X_n = i) = 0$, nous posons $P_{ij}(n) = \rho(j)$ où ρ est une probabilité quelconque sur E .

Lorsque la condition initiale est déterministe, $X_0 = i$, alors la loi initiale est $\delta_i(A) = \mathbf{1}_A(i)$, il s'agit d'un équivalent discret de la mesure de Dirac.

Interprétation

La propriété de Markov signifie aussi que :

$$\mathbb{P}_{X_{n+1} | X_{0:n}} \stackrel{\text{p.s.}}{=} \mathbb{P}_{X_{n+1} | X_n}$$

c'est-à-dire :

Le futur ne dépend du passé que par l'intermédiaire du présent.

De façon générale, un processus X_n sera une chaîne de Markov par rapport à une filtration \mathcal{F}_n lorsque :

$$\mathbb{P}_{X_{n+1} | \mathcal{F}_n} \stackrel{\text{p.s.}}{=} \mathbb{P}_{X_{n+1} | X_n},$$

c'est-à-dire :

$$\mathbb{P}(X_{n+1} \in A | \mathcal{F}_n) \stackrel{\text{p.s.}}{=} \mathbb{P}(X_{n+1} \in A | X_n)$$

pour tout $A \subset E$. Cette dernière définition s'étend aux chaînes de Markov à valeurs dans des espaces non nécessairement finis ou dénombrables. Nous nous limitons ici aux chaînes de Markov par rapport à leur tribu naturelle.

La propriété de Markov se comprend aussi de la façon suivante : *conditionnellement à $\{X_{n_0} = i\}$, $(X_n)_{n \geq n_0}$ est une chaîne de Markov (de loi initiale δ_i) indépendante de $X_{0:n_0-1}$.* Ainsi :

Conditionnellement au présent, future et passé sont indépendants.

La propriété de Markov est dite *faible* par contraste avec le fait qu'une chaîne de Markov X_n vérifie également la : *Propriété de Markov forte*

Pour tout t.a. T fini p.s., conditionnellement à $\{X_T = i\}$, $(X_n)_{n \geq T}$ est une chaîne de Markov (de loi initiale δ_i) indépendante de $X_{0:T-1}$.

13.2 Chaînes de Markov homogènes

Lorsque la probabilité de transition $P_{ij}(n)$ ne dépend pas de n , ce que nous supposons dans la suite, la chaîne est dite *homogène* (en temps). Nous définissons alors la *matrice de transition* $P = (P_{ij})_{i,j \in E}$ par :

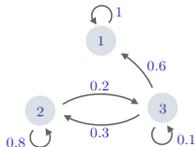
$$P_{ij} \stackrel{\text{d\u00e9f}}{=} \mathbb{P}(X_{n+1} = j | X_n = i).$$

La matrice de transition est \u00e9ventuellement de taille infinie quand E est infini-d\u00e9nombrable. Les propri\u00e9t\u00e9s d'une matrice de transition sont : $P_{ij} \geq 0$ pour tout $i, j \in E$ et $\sum_{j \in E} P_{ij} = 1$ pour tout $i \in E$, i.e. chaque ligne est une probabilit\u00e9 sur E .

Graphes associ\u00e9 \u00e0 une matrice de transition

On peut associer \u00e0 P un graphe orient\u00e9 et pond\u00e9r\u00e9 dont les n\u00f4uds sont les \u00e9tats de E . Un n\u00f4ud i est reli\u00e9 \u00e0 un n\u00f4ud j par une arr\u00eate orient\u00e9e (i, j) lorsque $P_{ij} > 0$ et cette arr\u00eate est pond\u00e9r\u00e9e par P_{ij} .

Lorsque E est fini et de cardinal relativement petit, il peut \u00eatre int\u00e9ressant de visualiser ce graphe. Voici un exemple avec une matrice 3×3 o\u00f9 les \u00e9tats sont class\u00e9s dans l'ordre $(1, 2, 3)$:

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.8 & 0.2 \\ 0.6 & 0.3 & 0.1 \end{pmatrix}$$


Loi d'une cha\u00eene de Markov

La loi d'une cha\u00eene de Markov homog\u00e8ne X_n est caract\u00e9ris\u00e9e par sa loi initiale et sa matrice de transition. En effet \u00e0 l'aide de la d\u00e9composition de la loi conjointe par les lois conditionnelles et par la propri\u00e9t\u00e9 de Markov :

$$\mathbb{P}(X_{0:n} = i_{0:n}) = \mu(\{i_0\}) P_{i_0 i_1} \cdots P_{i_{n-1} i_n}.$$

R\u00e9ciproquement si un processus v\u00e9rifie ce dernier r\u00e9sultat pour tout n et i_0, \dots, i_n , il s'agit alors d'une cha\u00eene de Markov homog\u00e8ne. On a \u00e9galement :

$$\begin{aligned} \mathbb{P}(X_{n+1:n+m} = i_{n+1:n+m} | X_{0:n} = i_{0:n}) \\ = P_{i_n i_{n+1}} \cdots P_{i_{n+m-1} i_{n+m}} \end{aligned}$$

ainsi en sommant sur $i_{n+1:n+m-1}$, on \u00e9tablit :

$$\mathbb{P}(X_{n+m} = j | X_n = i) = P_{ij}^m, \quad \forall i, j \in E$$

(\u00e9quation de Chapman-Kolmogorov)

o\u00f9 $P^m = P \times \cdots \times P$ (m fois) est la puissance m de la matrice P avec $P^0 = I$ (identit\u00e9). Attention : P_{ij}^m signifie $[P^m]_{ij}$ et non pas $[P_{ij}]^m$.

On consid\u00e8re un ensemble E au plus d\u00e9nombrable, une mesure μ et une matrice de transition P sur E , il est possible de d\u00e9finir de fa\u00e7on canonique un espace $(\Omega, \mathcal{F}, \mathbb{P})$ et une cha\u00eene de Markov X_n \u00e0 valeurs dans E , de loi initiale μ et de matrice de transition P . La probabilit\u00e9 \mathbb{P} ainsi construite est souvent not\u00e9e \mathbb{P}_μ lorsqu'il n'y a pas d'ambigu\u00eft\u00e9 sur P . Lorsque $\mu = \delta_i$ on note $\mathbb{P}_\mu = \mathbb{P}_i$, i.e. $\mathbb{P}_i(\cdot) = \mathbb{P}(\cdot | X_0 = i)$ est la probabilit\u00e9 conditionnelle sachant $X_0 = i$, on note \mathbb{E}_i l'esp\u00e9rance conditionnelle associ\u00e9e.

Notation matricielle

Toute mesure de probabilit\u00e9 μ sur E peut s'identifier \u00e0 une matrice ligne de termes $\mu(\{i\})$ que l'on notera aussi μ_i . Ainsi si $\mathbb{P}_{X_0} = \mu$, alors $\mathbb{P}_{X_1} = \mu P$, $\mathbb{P}_{X_2} = \mu P P = \mu P^2$, en fait :

$$\mathbb{P}_{X_n} = \mu P^n, \quad \mathbb{P}_{X_n | X_0 = i} = P_{i,\cdot}^n.$$

Une fonction $f : E \mapsto \mathbb{R}$ peut se mettre sous la forme d'un vecteur colonne et :

$$\mathbb{E}(f(X_n) | X_0 = i) = \sum_{j \in E} P_{ij}^n f(j) = [P^n f](i),$$

$$\mathbb{E}(f(X_n)) = \sum_{j \in E} \mathbb{P}(X_n = j) f(j) = \mu P^n f$$

\u00e0 condition par exemple que f soit positive ou born\u00e9e. De m\u00eame $\mu f = \sum_{i \in E} f(i) \mu_i$ est l'int\u00e9grale de f contre μ .

13.3 Comportement asymptotique

Un syst\u00e8me \u00e9voluant selon une dynamique markovienne (homog\u00e8ne) converge-t-il vers un \u00e9quilibre et que signifie "'\u00e9quilibre" dans ce contexte ?

Probabilit\u00e9 asymptotique et mesure invariante

Une mesure μ^* sur E est *asymptotique* pour la cha\u00eene X_n s'il existe une loi initiale μ_0 t.q. X_n converge en loi vers μ^* , c'est-\u00e0-dire :

$$\mathbb{P}_{\mu_0}(X_n = i) \xrightarrow[n \rightarrow \infty]{} \mu_i^*, \quad \forall i \in E,$$

μ^* est alors n\u00e9cessairement une mesure de probabilit\u00e9.

Une mesure positive λ sur E sera dite *invariante* ou *stationnaire* pour P ou pour la cha\u00eene de Markov X_n , lorsque :

$$\lambda P = \lambda.$$

Les mesures invariantes sont donc les vecteurs propres (\u00e0 gauche) de la matrice P associ\u00e9es \u00e0 la valeur propre 1.

Ici λ n'est pas n\u00e9cessairement une probabilit\u00e9, mais lorsque $\lambda(E) < \infty$ alors $\bar{\lambda} = \frac{1}{\lambda(E)} \lambda$ est une mesure de probabilit\u00e9. L'invariance signifie que si $\mathbb{P}_{X_0} = \bar{\lambda}$ alors $\mathbb{P}_{X_1} = \bar{\lambda}$, plus pr\u00e9cis\u00e9ment $\mathbb{P}_{\bar{\lambda}}(X_1 = i) = \bar{\lambda}_i$. En d'autres termes, si \u00e0 un instant n donn\u00e9 la loi de X_n est $\bar{\lambda}$, alors X_m conserve la m\u00eame loi pour tout $m \geq n$. En fait si μ est une mesure de probabilit\u00e9 sur E alors on a l'\u00e9quivalence entre les assertions suivantes :

- (i) μ est une mesure invariante ;
- (ii) μ est une probabilit\u00e9 asymptotique.

Existence d'une mesure invariante dans le cas E fini

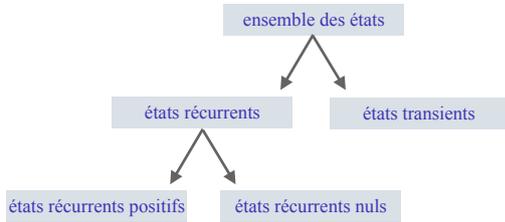
Le cas fini est tr\u00e8s particulier, on peut s'appuyer sur le classique :

Th\u00e9or\u00e8me de Perron-Frobenius^w

Toute matrice de transition P de dimension finie admet 1 comme valeur propre (\u00e0 gauche) et c'est la valeur propre de plus grand module (i.e. $|\gamma| \leq 1$ pour tout autre valeur propre γ). De plus cette valeur propre admet un vecteur propre dont tous les termes sont positifs.

Ainsi il existe une mesure de probabilité invariante. Il faut simplement comprendre que toute chaîne de Markov à valeurs dans un espace d'états fini "y est à l'étroit" ! L'unicité d'une mesure invariante ainsi que la convergence vers cette mesure invariante demande un peu plus d'attention. Par ailleurs le cas dénombrable est bien plus complexe.

Classification des états



On considère le temps d'atteinte de l'état i :

$$T_i \stackrel{\text{def}}{=} \inf\{n; X_n = i\}.$$

Un état i est dit :

- *transient* si $\mathbb{P}_i(T_i < \infty) < 1$;
- *récurrent* si $\mathbb{P}_i(T_i < \infty) = 1$, on distinguera alors entre :
 - *récurrent nul* si de plus $\mathbb{E}_i(T_i) = \infty$;
 - *récurrent positif* si de plus $\mathbb{E}_i(T_i) < \infty$.

Par exemple, i vérifiant $P_{ii} = 1$ est un état récurrent dit *absorbant*.

Soit :

$$N_j \stackrel{\text{def}}{=} \sum_{n>0} \mathbf{1}_{\{X_n=j\}}$$

le nombre de passages en j après l'instant 0 et :

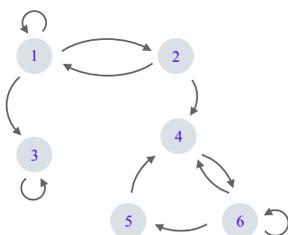
$$\rho_{ij} \stackrel{\text{def}}{=} \mathbb{P}_i(T_j < \infty) = \mathbb{P}_i(N_j \geq 1)$$

la probabilité d'atteindre j à partir de i en temps fini.

On a la dichotomie suivante :

- si i est récurrent alors X_n revient \mathbb{P}_i -p.s. une infinité de fois en i , i.e. $\mathbb{P}_i(N_i = \infty) = 1$.
- si i est transient, le nombre de passages en i est p.s. fini, i.e. $\mathbb{P}_i(N_i = \infty) = 0$, et suit une loi géométrique : $\mathbb{P}_i(N_i = k) = (1 - \rho_{ii}) \rho_{ii}^{k-1}$ pour $k \in \mathbb{N}_*$ (noter la "loi 0-1" : $\mathbb{P}_i(N_i = \infty)$ ne peut être que 0 ou 1).

Il est clair que $\rho_{ij} > 0$ ssi il existe un chemin reliant i à j dans le graphe associé à la matrice P . Dans ce cas on dit que j est *accessible* depuis i , noté $i \rightarrow j$. Cette propriété s'écrit également : $\exists m \in \mathbb{N}$ t.q. $P_{ij}^m > 0$. Lorsque $i \rightarrow j$ et $j \rightarrow i$, on dira que les états i et j *communiquent*, noté $i \leftrightarrow j$; il s'agit d'une relation d'équivalence sur E . Lorsque deux états communiquent, ils sont alors tous deux soit récurrents, soit transients.



Les classes de communication sont :
 $\{1, 2\}$ (transiente),
 $\{3\}$ (récurrente),
 $\{4, 5, 6\}$ (récurrente);
 3 est un état absorbant.

Chaîne récurrente et irréductible

Une chaîne de Markov, ou sa matrice de transition, sera dite *irréductible* lorsqu'il n'existe qu'une seule classe d'équivalence pour la relation de communication, c'est-à-dire lorsque tous les états communiquent. Dans ce cas :

- si E est fini : alors tous les états sont récurrents ;
- si E est infini-dénombrable : alors les états sont soit tous récurrents, soit tous transients.

La chaîne de Markov, ou sa matrice de transition, sera dite *récurrente et irréductible* lorsqu'il n'existe qu'une seule classe et que tous les états sont récurrents.

Si la chaîne X_n est irréductible alors :

$$\frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_{\{X_k=i\}} \xrightarrow[n \rightarrow \infty]{\text{p.s.}} \frac{1}{\mathbb{E}_i(T_i)} \quad (\Delta)$$

pour toute loi initiale. Par ailleurs :

Existence et unicité de la mesure invariante

Si la chaîne est récurrente irréductible alors il existe une mesure invariante λ strictement positive et elle est unique à une constante multiplicative près.

Il reste maintenant à distinguer les deux cas :

- (i) *Irréductible récurrent positif* : Dans ce cas tous les états sont récurrents positifs et $\lambda(E) < \infty$; on peut donc définir une probabilité invariante $\bar{\lambda} = \lambda(E)^{-1} \lambda$ et cette probabilité est *unique*. On a $\bar{\lambda}_i = 1/\mathbb{E}_i(T_i)$ et d'après (Δ) :

$$\frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_{\{X_k=i\}} \xrightarrow[n \rightarrow \infty]{\text{p.s.}} \bar{\lambda}_i. \quad (\Delta\Delta)$$

Plus généralement on a aussi :

$$\frac{1}{n} \sum_{k=0}^{n-1} \varphi(X_k) \xrightarrow[n \rightarrow \infty]{\text{p.s.}} \sum_{i \in E} \lambda_i \varphi(i)$$

(théorème ergodique)

pour toute fonction $\varphi : E \rightarrow \mathbb{R}$ bornée.

- (ii) *Irréductible récurrent nul* : Dans ce cas $\lambda(E) = \infty$ et tous les états sont récurrents nuls, i.e. $\mathbb{E}_i(T_i) = \infty$ pour tout $i \in E$.

Le théorème ergodique est important en pratique : il permet par exemple de calculer les moments de la loi asymptotique de X_n à partir de moyennes temporelles d'une seule trajectoire $n \mapsto X_n(\omega)$.

En conclusion une chaîne irréductible ne peut être que dans un des cas suivants :

- (i) *transiente* : tous les états sont transients,
- (ii) *récurrente nulle* : tous les états sont récurrents nuls,
- (iii) *récurrente positive* : tous les états sont récurrents positif,

uniquement dans ce dernier cas il existe une unique probabilité invariante. Dans le cas fini une chaîne irréductible est nécessairement récurrente positive.

Dans le cas non-irréductible on peut associer une mesure invariante à chaque classe d'équivalence récurrente positive dont le support sera exactement cette classe, et les combinaisons linéaires de ces mesures invariantes sont aussi des mesures invariantes.

Périodicité et convergence vers la mesure invariante

On considère une chaîne X_n irréductible et récurrente positive de probabilité invariante $\bar{\lambda}$. En prenant l'espérance par rapport à \mathbb{E}_j dans $(\Delta\Delta)$ (et en inversant i et j) on obtient la convergence de $\frac{1}{n} \sum_{k=0}^{n-1} \mathbb{P}_i(X_k = j)$ vers $\bar{\lambda}_j$. Mais $P_{ij}^n = \mathbb{P}_i(X_n = j)$ converge-t-elle vers $\bar{\lambda}_j$? On a déjà ce résultat :

Convergence vers la probabilité invariante 1

S'il existe k t.q. :

$$P_{ij}^k > 0, \quad \forall i, j \in E$$

alors la X_n est irréductible, récurrente positive et admet une unique mesure de probabilité invariante $\bar{\lambda}$, de plus :

$$\mathbb{P}(X_n = j) \xrightarrow[n \rightarrow \infty]{} \bar{\lambda}_j, \quad \forall j \in E \quad (\Delta\Delta\Delta)$$

(convergence de la loi marginale)

pour toute loi initiale, i.e. la mesure de probabilité asymptotique de la chaîne X_n est $\bar{\lambda}$; en particulier $P_{ij}^n \rightarrow \bar{\lambda}_j$. Dans ce cas la chaîne est dite *ergodique*.

Mais pour répondre de façon plus précise à cette question il va falloir écartier une difficulté (une de plus, mais c'est la dernière).

Un état i est dit *apériodique* lorsqu'il existe k t.q. :

$$P_{ii}^n > 0 \text{ pour tout } n \geq k.$$

Si X_n est irréductible et s'il existe un état apériodique alors tous les états sont apériodiques, on dit que X_n ou P sont apériodiques.

Convergence vers la probabilité invariante 2

Une chaîne irréductible, récurrente positive et apériodique est ergodique, i.e. elle vérifie $(\Delta\Delta\Delta)$.

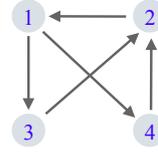
La *période* d'un état i est définie par :

$$d(i) \stackrel{\text{déf}}{=} \text{p.g.c.d.} \{n \geq 1 : P_{ii}^n > 0\}.$$

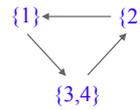
si deux états communiquent, ils ont alors même période. Donc lorsque la chaîne est irréductible, les états partagent la même période d qui sera appelée la *période de*

la chaîne ou de la matrice de transition. Lorsque $d(i) = 1$ (resp. $d = 1$), l'état (resp. la chaîne ou la matrice de transition) sera dit(e) apériodique..

Étant donnée une classe irréductible C de période d , il existe alors une partition $C = \mathcal{P}_0 \cup \dots \cup \mathcal{P}_{d-1}$ telle que les seules transitions possibles dans la classe soient de \mathcal{P}_ℓ dans $\mathcal{P}_{\ell+1}$ (ℓ modulo d).



La chaîne est de période 3 et les classes de périodicité sont $\mathcal{P}_0 = \{1\}$, $\mathcal{P}_1 = \{2\}$ et $\mathcal{P}_2 = \{3, 4\}$.



La structure dynamique des classes de périodicité est représentée dans le second schéma.

Convergence d'une chaîne irréductible périodique

On considère une chaîne irréductible périodique de période d de matrice de transition P , alors pour tout $i, j \in E$ il existe $r \in \{0, \dots, d-1\}$ t.q. :

$$P_{ij}^{nd+r} \xrightarrow[n \rightarrow \infty]{} \frac{d}{\mathbb{E}_j(T_j)}$$

tandis que $P_{ij}^{nd+r'} = 0$ pour $0 \leq r' < d$ avec $r' \neq r$.

Références

- [1] P. Barbe and M. Ledoux. *Probabilité*. EDP Sciences, 2007.
- [2] M. Cottrell, V. Genon-Catalot, C. Duhamel, and T. Meyre. *Exercices de Probabilités*. Cassini, Paris, 1999.
- [3] J. Jacod and P. Protter. *L'essentiel en théorie des probabilités*. Cassini, 2003.
- [4] J. R. Norris. *Markov chains*. Cambridge University Press, 1998.

Index

- $\{0 : n\}$ suite des indices 0 à n , 13
- A_n^p nombre d'arrangements, 2
- $C_0(\mathbb{R}^d)$, espace des fonctions
 - continues nulles à l'infini de \mathbb{R}^d
 - dans \mathbb{R} , 7
- $C_K(\mathbb{R}^d)$, espace des fonctions continues
 - bornées de \mathbb{R}^d
 - dans \mathbb{R} , 6
- C_n^p , voir $\binom{n}{p}$
- $\binom{n}{p}$ nombre de combinaisons, 2
- $\langle X \rangle_n$, crochet de la martingale X_n , 13
- $X_n \xrightarrow{L^p} X$, voir convergence en moyenne d'ordre p
- $X_n \xrightarrow{\text{proba.}} X$, voir convergence en probabilité
- $X_n \xrightarrow{p.s.} X$, voir convergence presque sûre
- $\delta_i(A) = \mathbf{1}_A(i)$ "mesure de Dirac" discrète., 14
- $\mathbb{E}(X|\mathcal{G})$, 1
- $\overset{\text{i.i.d.}}{\sim}$, voir indépendant et identiquement distribué
- i.i.d., voir indépendant et identiquement distribué
- $A \perp B$, événements indépendants, 2
- $\mathbf{1}_A(x)$, fonction indicatrice de A , $\mathbf{1}_A(x) = 1$ si $x \in A$, 0 sinon, 2
- $L^p(\Omega)$, 5
- $x \vee y = \max(x, y)$, 4
- $x \wedge y = \min(x, y)$, 4
- $\mathcal{P}(\Omega)$, ensemble des parties de Ω , 1
- p.s., voir presque sûrement
- σ -algèbre, voir tribu
- $\sigma(X)$ tribu engendrée par la variable aléatoire X , 2
- t.a., voir temps d'arrêt
- v.a., voir variable aléatoire
- v.a.r., voir variable aléatoire réelle
- absence de mémoire
 - de la loi exponentielle, 12
 - de la loi géométrique, 11
- arrangements, 2
- atome, 3
- borélien, 1
- chaîne de Markov, 14
 - apériodique, 17
 - ergodique, 17
 - homogène, 15
 - irréductible, 16
 - loi initiale d'une, 14
 - matrice de transition d'une, 15
 - périodique, 17
 - probabilités de transition d'une, 14
 - récurrente et irréductible, 16
- coefficient de corrélation, 5
- combinaisons, 2
- convergence (de mesures) étroite, 7
 - faible, 7
 - vague, 7
- convergence (de variables aléatoires) dans $L^p(\Omega)$, voir convergence en moyenne d'ordre p
 - en loi, 6
 - en moyenne d'ordre p , 6
 - en probabilité, 6
 - presque sûre, 6
 - simple, 3
- covariance, 5
- crochet d'une martingale, 13
- décomposition
 - d'un événement joint par un produit de probabilités conditionnelles, 7
 - d'une loi jointe par un produit de lois conditionnelles, 8, 10
 - de Doob, 13
- densité, 3
 - conditionnelle, 8
 - marginale, 3
- écart-type d'une variable aléatoire réelle, 4
- équation
 - de Chapman-Kolmogorov, 15
 - de Wald, 14
- ergodicité, 17
- espace
 - d'états, 14
 - de probabilité, 1
 - filtré, 13
 - produit, 2
 - probabilisable, 1
- espérance, 4
 - conditionnelle, 7, 8
- états
 - absorbants, 16
 - accessibles, 16
 - apériodiques, 17
 - communicants, 16
 - récurrents, 16
 - nuls, 16
 - positifs, 16
 - transients, 16
- événement, 1
 - négligeable, 1
- filtration, 13
 - naturelle, 13
- fonction
 - caractéristique, 5
 - de répartition, 3
 - génératrice des moments, 5
 - génératrice des probabilités, 6
- formule
 - d'inversion de Fourier, 5
 - de Bayes, 7
 - de Stirling, 10
 - des probabilités totales, 7
- indépendant et identiquement distribué, 4
- indépendance
 - d'événements, 2
 - de tribus, 2
 - de variables aléatoires, 4
- inégalité
 - de Bienaymé-Tchebychev, 4
 - de Doob, 14
 - de Hölder, 5
 - de Jensen, 4
 - de Markov, 4
 - de martingales, 14
 - de Minkowski, 5
- intégrabilité, 4
- irréductibilité, 16
- lemme
 - de Fatou, 6
 - conditionnel, 9
- loi
 - binomiale, 10
 - binomiale négative, 11
 - conditionnelle, 8, 9
 - de Bernoulli, 10
 - de Cauchy, 12
 - de Poisson, 11
 - de probabilité, 3
 - des grands nombres, 7
 - exponentielle, 11
 - Gamma, 12
 - gaussienne, 12
 - géométrique, 11
 - hypergéométrique, 11
 - marginale, 3
 - multinomiale, 10
 - uniforme, 12
- médiane, 11
- marche aléatoire, 13
- martingale, 13
 - arrêtée, 14
- matrice
 - de covariance, 5
 - de transition, 15
 - irréductible, 16
 - récurrente et irréductible, 16
- mesurabilité, 2
- mesure
 - de probabilité, 1
 - finie positive, 1
 - invariante, 15
 - existence et unicité, 16
- stationnaire, 15
- mode d'une densité, 11
- moment d'une variable aléatoire réelle, 4
- période
 - d'un état, 17
 - d'une chaîne de Markov, 17
 - d'une matrice de transition, 17
- premier temps d'atteinte d'un état, 13
- d'un ensemble, 13
- presque sûrement, 2
- probabilité
 - asymptotique, 15
 - conditionnelle, 7–9
 - des causes, 7
 - sur un ensemble fini, 2, 10
- processus aléatoire, 13
 - adapté, 13
 - prévisible, 13
- propriété de Markov, 14
 - faible, 14
 - forte, 14
- risque quadratique, 9, 12
- sous-martingale, 13
- sur-martingale, 13
- temps d'arrêt, 13
- théorème
 - central limite, 7
 - d'arrêt, 14
 - de convergence des martingales, 14
 - de convergence dominée, 6
 - conditionnelle, 9
 - de convergence monotone, 6
 - conditionnelle, 9
 - de Fubini, 4
 - de Perron-Frobenius, 15
 - de représentation de Skorokhod, 6
 - des classes monotones, 2
 - des probabilités composées, 7
 - ergodique, 16
- transformée
 - de Fourier, 5
 - de Laplace, 5
- tribu, 1
 - borélienne, 1
 - complète, 1
 - des événements antérieurs, 13
 - engendrée, 1, 2
- variable aléatoire, 2
 - discrète, 2
 - étagée, 2
 - réelle, 2
 - simple, voir variable aléatoire étagée
- variance d'une variable aléatoire réelle, 4
- vecteur aléatoire
 - gaussien, 12
 - réel, 2