Département STPI - 3MIC Mathématiques Appliquées

Compléments de Probabilités

Romain Duboscq et Aldéric Joulin

R. Duboscq (duboscq@insa-toulouse.fr) - Bureau 124 - Bâtiment GMM

A. Joulin (ajoulin@insa-toulouse.fr) - Bureau 115 - Bâtiment GMM

Année universitaire 2024-2025

Table des matières

Chapitre 1

Théorie de la mesure probabiliste

Dans ce premier chapitre, nous allons brièvement introduire la théorie des probabilités selon l'axiomatique de Kolmogorov datant des années 1920, en l'inscrivant dans le cadre général de la théorie de la mesure et de l'intégration développée par Lebesgue au début du vingtième siècle. Bien évidemment, il ne s'agit pas de faire un cours exhaustif sur ce trop vaste sujet mais plutôt de voir comment les concepts probabilistes vus en deuxième année se réécrivent dans le langage de la théorie de la mesure.

1.1 Tribus, probabilités, variables aléatoires et lois

Avant de commencer à établir les résultats importants de la théorie des probabilités, reformulons tout d'abord les objets probabilistes d'intérêt à travers le langage de la théorie de la mesure. Pour ce faire, le tableau suivant récapitule les différente éléments de langage.

Langage des probabilités	Théorie de la mesure
espace probabilisé : $(\Omega, \mathcal{A}, \mathbb{P})$ avec de plus $\mathbb{P}(\Omega) = 1$	espace mesuré : $(E, \mathcal{A}, \mu)$
paramètre d'aléa : ω	inconnue : x
variable aléatoire réelle : $X : \Omega \to \mathbb{R}$	fonction mesurable : $f : E \to \mathbb{R}$
espérance : $\mathbb{E}[X] = \int_\Omega X d\mathbb{P}$	intégrale de Lebesgue : $\int_E f d\mu$
le presque sûrement	le presque partout
la convergence presque sûre	la convergence presque partout

Démarrons en introduisant la notion de tribu.

Définition 1.1.1. *Soit Ω un ensemble et $\mathcal{A}$ une famille de parties de Ω. On dit que $\mathcal{A}$ est une tribu (sur Ω) si :*

(i) $\Omega \in \mathcal{A}$.

(ii) Stabilité par passage au complémentaire : pour tout ensemble $A \in \mathcal{A}$, on a $A^c \in \mathcal{A}$, où A^c désigne le complémentaire de A dans Ω.

(iii) Stabilité par union dénombrable : pour toute famille $(A_n)_{n\in\mathbb{N}}$ *de* Ω *satisfaisant* $A_n \in \mathcal{A}$ *pour tout* $n \in \mathbb{N}$*, alors*

$$\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}.$$

En particulier, on appelle tribu engendrée par une classe de parties $\mathcal{C}$ de Ω la plus petite tribu sur Ω contenant $\mathcal{C}$, c'est-à-dire l'intersection de toutes les tribus contenant $\mathcal{C}$ (on vérifiera à titre d'exercice que l'intersection − dénombrable ou non − de tribus est une tribu ; en revanche l'union − même dénombrable − de tribus n'est pas forcément une tribu). On la note $\sigma(\mathcal{C})$. Un exemple classique est la tribu borélienne sur $\Omega = \mathbb{R}$ ou $\mathbb{R}^d$, où $\mathcal{C}$ désigne l'ensemble des ouverts (ou fermés) de $\mathbb{R}^d$. On la note $\mathcal{B}(\mathbb{R})$ ou $\mathcal{B}(\mathbb{R}^d)$.

Les tribus sont le cadre naturel sur lequel sont définies les mesures au sens de Lebesgue.

Définition 1.1.2. *Soit* Ω *un ensemble muni d'une tribu* $\mathcal{A}$*. On appelle mesure (ou mesure positive) toute application* μ *définie sur* $\mathcal{A}$ *et à valeurs dans* $[0, +\infty]$ *telle que :*

1. $\mu(\emptyset) = 0$ *;*
2. μ *est* σ*-additive : pour toute famille dénombrable* $(A_n)_{n\in\mathbb{N}}$ *d'ensembles deux à deux disjoints appartenant à la tribu* $\mathcal{A}$*,*

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

Le triplet $(\Omega, \mathcal{A}, \mu)$ *est alors appelé espace mesuré.*

On classe les mesures en fonction de certaines de leurs propriétés.

Définition 1.1.3. *Soit* $(\Omega, \mathcal{A}, \mu)$ *un espace mesuré.*

1. *On dit que* μ *est finie si* $\mu(\Omega) < +\infty$*.*
2. *Si* $\mu(\Omega) = 1$*, on dit que* μ *est une mesure de probabilité.*
3. *On dit que* μ *est* σ*-finie s'il existe une suite* $(\Omega_k)_{k\in\mathbb{N}}$ *telle que*

$$\mu(\Omega_k) < +\infty \quad \forall k \in \mathbb{N} \quad \text{et} \quad \Omega = \bigcup_{k\in\mathbb{N}} \Omega_k.$$

Par exemple, si Ω est un ensemble fini ou dénombrable muni de la tribu $\mathcal{A} = \mathcal{P}(\Omega)$, l'ensemble des parties de Ω, on définit la mesure de comptage par

$$\mu(A) := \begin{cases} \text{card}(A) & \text{si } A \text{ est fini ;} \\ +\infty & \text{sinon.} \end{cases}$$

En particulier si Ω est fini, alors $\tilde{\mu} := \frac{1}{card(\Omega)}\,\mu$ est une mesure de probabilité : c'est la probabilité uniforme sur l'ensemble fini Ω.

Une autre mesure importante est la mesure de Dirac sur $(\Omega, \mathcal{A})$: pour $x \in \Omega$, la mesure de Dirac en x, notée δ_x, est définie par

$$\delta_x(A) := \begin{cases} 1 & \text{si } x \in A \ ; \\ 0 & \text{sinon.} \end{cases}$$

Il s'agit également d'une mesure de probabilité. Par ailleurs, la mesure de comptage précédente peut s'exprimer comme

$$\mu(A) := \sum_{x \in X} \delta_x(A).$$

Enfin, si $\Omega = \{x_1, \ldots, x_n, \ldots\}$ et $\mathcal{A} = \mathcal{P}(\Omega)$, alors pour toute suite $(\alpha_n)_{n \in \mathbb{N}^*}$ de réels positifs ou nuls,

$$\mu = \sum_{n \in \mathbb{N}} \alpha_n \, \delta_{x_n},$$

est une mesure sur $(\Omega, \mathcal{P}(\Omega))$. Si de surcroît la somme sur tous les α_n vaut 1, alors c'est une mesure de probabilité.

Étudions maintenant les principales propriétés des mesures.

Proposition 1.1.4. *Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré.*

1. *Pour tous $A, B \in \mathcal{A}$ tels que $A \subset B$, on a $\mu(A) \leq \mu(B)$.*
2. *Pour toute famille dénombrable $(A_k)_{k \in \mathbb{N}}$ d'ensembles appartenant à $\mathcal{A}$,*

$$\mu\Big(\bigcup_{k \in \mathbb{N}} A_k \Big) \leq \sum_{k \in \mathbb{N}} \mu(A_k).$$

3. *Pour tous $A, B \in \mathcal{A}$,*

$$\mu(A \cup B) + \mu(A \cap B) = \mu(A) + \mu(B).$$

Démonstration.

1. On écrit $B = A \cup (B \setminus A)$, l'union étant disjointe. On a donc $\mu(B) = \mu(A) + \mu(B \setminus A)$. Comme μ est à valeurs positives, on en déduit que $\mu(B) \geq \mu(A)$.
2. On construit à partir de la famille $(A_k)_{k \in \mathbb{N}}$ une famille $(B_k)_{k \in \mathbb{N}}$ d'ensembles deux à deux disjoints appartenant à $\mathcal{A}$: on pose
   (a) $B_0 = A_0$;
   (b) pour tout $k \geq 1$, $B_k = A_k \setminus \bigcup_{j=0}^{k-1} A_j$.

   Les $(B_k)_{k \in \mathbb{N}}$ vérifient $\bigcup_{k=0}^{n} B_k = \bigcup_{k=0}^{n} A_k$ pour tout $n \in \mathbb{N}$ mais aussi $\bigcup_{k \in \mathbb{N}} B_k = \bigcup_{k \in \mathbb{N}} A_k$. De plus, on a $B_k \subset A_k$ pour tout $k \in \mathbb{N}$. D'où

$$\mu\Big(\bigcup_{k \in \mathbb{N}} A_k \Big) = \mu\Big(\bigcup_{k \in \mathbb{N}} B_k \Big) = \sum_{k \in \mathbb{N}} \mu(B_k) \leq \sum_{k \in \mathbb{N}} \mu(A_k).$$

3. On écrit
$$A = (A \setminus B) \cup (A \cap B),\ B = (B \setminus A) \cup (A \cap B);$$
$$A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A),$$
où les unions sont disjointes. On en déduit
$$\begin{aligned} \mu(A) + \mu(B) &= \mu(A \setminus B) + \mu(A \cap B) + \mu(B \setminus A) + \mu(A \cap B) \\ &= \mu(A \cup B) + \mu(A \cap B). \end{aligned}$$

La démonstration est à présent achevée. ∎

Proposition 1.1.5. *Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré et $(A_k)_{k\in\mathbb{N}}$ une suite d'ensembles appartenant à $\mathcal{A}$.*

1. *Si la suite $(A_k)_{k\in\mathbb{N}}$ est croissante pour l'inclusion (i.e. $A_k \subset A_{k+1}$), alors*
$$\mu\left(\bigcup_{k\in\mathbb{N}} A_k\right) = \lim_{k\to+\infty} \mu(A_k).$$

2. *Si la suite $(A_k)_{k\in\mathbb{N}}$ est décroissante pour l'inclusion (i.e. $A_{k+1} \subset A_k$), et si $\mu(\Omega) < +\infty$, alors*
$$\mu\left(\bigcap_{k\in\mathbb{N}} A_k\right) = \lim_{k\to+\infty} \mu(A_k).$$

Démonstration.

1. En utilisant la famille $(B_k)_{k\in\mathbb{N}}$ d'ensembles deux à deux disjoints définie dans la démonstration de la Proposition 1.1.4, on a
$$\begin{aligned} \lim_{n\to+\infty} \mu(A_n) &= \lim_{n\to+\infty} \mu\left(\bigcup_{k=0}^{n} A_k\right) = \lim_{n\to+\infty} \mu\left(\bigcup_{k=0}^{n} B_k\right) = \lim_{n\to+\infty} \sum_{k=0}^{n} \mu(B_k) \\ &= \sum_{k\in\mathbb{N}} \mu(B_k) = \mu\left(\bigcup_{k\in\mathbb{N}} B_k\right) = \mu\left(\bigcup_{k\in\mathbb{N}} A_k\right). \end{aligned}$$

2. Dans le cas décroissant, il suffit d'appliquer ce qui précède à la famille croissante $(A_k^c)_{k\in\mathbb{N}}$ puis de passer aux complémentaires, la mesure étant supposée finie.

∎

À présent, on suppose que Ω désigne l'univers associé à une expérience aléatoire et est muni d'une tribu $\mathcal{A}$ ainsi que d'une probabilité $\mathbb{P}$, faisant de l'espace mesuré $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité. Les ensembles appartenant à la tribu $\mathcal{A}$ sont appelés évènements.

Définition 1.1.6. *Une application $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ est une variable aléatoire (réelle) si elle est mesurable, c'est-à-dire pour tout $B \in \mathcal{B}(\mathbb{R})$, on a $X^{-1}(B) \in \mathcal{A}$.*

Notons que l'ensemble $X^{-1}(B)$, qui est l'image réciproque de l'ensemble borélien B par la variable aléatoire X, est défini par

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

En théorie des probabilités, l'usage est de ne pas faire apparaître le paramètre de hasard ω dans les évènements (sauf si cela nuit à la compréhension), auquel cas l'évènement $\{\omega \in \Omega : X(\omega) \in B\}$ se réécrit de manière abusive comme l'évènement plus familier $\{X \in B\}$.

Dans la suite de ce cours, toutes nos v.a. seront à valeurs réelles (sauf mention du contraire, notamment dans le chapitre sur les vecteurs gaussiens).

Définition 1.1.7. *La tribu engendrée par une v.a. X, notée $\sigma(X)$, est la sous-tribu de $\mathcal{A}$ engendrée par l'ensemble des images réciproques de X. Autrement dit,*

$$\begin{aligned} \sigma(X) &:= \sigma\left(\{X^{-1}(B) : B \in \mathcal{E}\}\right) \\ &= \sigma\left(\{\{X \in B\} : B \in \mathcal{E}\}\right). \end{aligned}$$

C'est la plus petite tribu sur Ω rendant X mesurable.
De même, si $X_1, \ldots, X_n$ sont des v.a., alors on définit la tribu engendrée par ces v.a. comme

$$\sigma(X_1, \ldots, X_n) := \sigma\left(\left\{\bigcap_{i=1}^{n}\{X_i \in B_i\} : B_1, \ldots, B_n \in \mathcal{E}\right\}\right).$$

Énonçons maintenant un résultat très utile en pratique, le Lemme de Doob, dû à un célèbre probabiliste américain du milieu de vingtième siècle. En particulier, ce lemme nous donne un critère simple pour établir la mesurabilité d'une v.a. Y par rapport à la tribu engendrée par une ou plusieurs v.a. Nous admettrons la démonstration.

Lemme 1.1.8 (Doob)**.** *Étant données des v.a. $X_1, \ldots, X_n$, une v.a. Y est $\sigma(X_1, \ldots, X_n)$-mesurable si et seulement s'il existe une fonction mesurable $h : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ telle que*

$$Y = h(X_1, \ldots, X_n).$$

Afin de pouvoir calculer des quantités intéressantes relatives à une v.a. donnée, il nous faut introduire une structure probabiliste sur l'espace d'arrivée $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$: c'est l'objet des lois (de probabilité).

Définition 1.1.9. *La loi d'une v.a. X est définie sur $\mathcal{B}(\mathbb{R})$ par*

$$P_X(A) = \mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) = \mathbb{E}[\mathbf{1}_A(X)], \quad A \in \mathcal{B}(\mathbb{R}).$$

Dans le langage de la théorie de la mesure, la loi P_X est la mesure image de la mesure de probabilité $\mathbb{P}$ par la v.a. X. En particulier, il est facile de vérifier qu'elle satisfait les axiomes d'une mesure de probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Par ailleurs, sachant que toute fonction mesurable peut être approchée au sens de la convergence simple par une suite de fonctions étagées, on en déduit que caractériser la loi d'une v.a. X revient à calculer la quantité $\mathbb{E}[f(X)]$ pour toute fonction $f : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mesurable positive ou bornée.

Dans la théorie de la mesure, la mesure de Lebesgue λ est la mesure σ-finie définie sur la tribu borélienne $\mathcal{B}(\mathbb{R})$ qui coïncide sur les intervalles avec leur longueur, c'est-à-dire

$$\lambda([a, b]) = b - a \quad \text{pour tout} \quad a < b \in \mathbb{R}.$$

On peut alors montrer que toute mesure σ-finie (donc toute mesure de probabilité) sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ se décompose en une partie absolument continue et une partie étrangère pour la mesure de Lebesgue. Sans trop rentrer dans les détails, cela signifie dans notre cadre probabiliste que la loi P_X de la v.a. X peut se décomposer de la manière suivante par rapport à la mesure de Lebesgue : il existe f_X une fonction mesurable positive et μ une mesure satisfaisant la propriété

$$\mu(A) \neq 0 \Rightarrow \lambda(A) = 0, \quad A \in \mathcal{B}(\mathbb{R}),$$

telles que

$$P_X(A) = \int_A f_X d\lambda + \mu(A).$$

Si $\mu = 0$ alors on dit que f_X est la densité (de probabilité) de la v.a. X. On dit aussi parfois que X est continue (attention, cela n'a rien à voir avec la notion usuelle de la continuité) et que P_X est absolument continue par rapport à la mesure de Lebesgue. Dans le cas où $f_X = 0$ et μ est une somme dénombrable pondérées de mesures de Dirac δ_{x_n}, alors la v.a. est dite discrète. Une v.a. est donc discrète si et seulement s'il existe des coefficients $(\alpha_n)_{n\in\mathbb{N}}$ positifs ou nuls et une suite de points $(x_n)_{n\in\mathbb{N}}$ tels que

$$P_X = \sum_{n\in\mathbb{N}} \alpha_n \delta_{x_n} \quad \text{avec} \quad \sum_{n\in\mathbb{N}} \alpha_n = 1.$$

Si une v.a. ne prend qu'un nombre dénombrable de valeurs (x_n), alors elle est discrète et

$$P_X = \sum_{n\in\mathbb{N}} \mathbb{P}(X = x_n)\delta_{x_n}.$$

Par exemple, si l'on considère le jeu du pile ou face équilibré, la v.a. X prenant la valeur 1 si l'on obtient pile et 0 sinon, toutes deux avec probabilité $1/2$, suit une loi de Bernoulli de paramètre $1/2$, notée $\mathcal{B}(1/2)$. La loi P_X s'écrit comme une somme de masses de Dirac pondérées :

$$P_X = \frac{1}{2}\left(\delta_0 + \delta_1\right).$$

De même, si X correspond au nombre de résultats pile obtenus lorsqu'on lance à n reprises la même pièce de monnaie, et ce de manière indépendante, alors X est une v.a. binomiale de paramètres n et $p = 1/2$ et sa loi est donnée par

$$P_X = \left(\frac{1}{2}\right)^n \sum_{k=0}^{n} C_n^k \, \delta_k.$$

Bien évidemment, les autres exemples classiques comme la loi uniforme sur un ensemble fini (que nous avons introduite précédemment) ou la loi géométrique rentrent dans ce cadre. Dans le cas des v.a. continues, nous considérons les lois uniforme sur un intervalle borné, normale (ou gaussienne), exponentielle, Gamma, etc...

1.2 Espérance et intégrabilité

À présent nous allons faire le lien entre l'espérance et les espaces L^p. Avant de rentrer dans le vif du sujet, notons que lorsque l'on aura une égalité entre v.a., on sous-entendra qu'elles seront vérifiées presque sûrement (p.s.), c'est-à-dire pour tout ω en dehors d'un ensemble négligeable, i.e. un sous-ensemble (inclus dans un événement de Ω) de probabilité 0. Il s'agit de la version probabiliste du "presque partout" apparaissant en théorie de la mesure, une probabilité étant une mesure de masse totale égale à 1. De même, la convergence d'une suite de v.a. vers une autre sera celle de la convergence presque sûre, c'est-à-dire que la convergence aura lieu pour tout ω en dehors d'un ensemble négligeable. Cette notion de convergence sera étudiée plus précisément dans le chapitre 3.

Étant donné $p \in [1, +\infty[$, on notera pour alléger la notation $L^p(\mathcal{A}) := L^p(\Omega, \mathcal{A}, \mathbb{P})$ (ou simplement L^p s'il n'y a pas de confusion possible) l'ensemble des v.a. pour lesquelles le p-ième moment est fini, à savoir

$$L^p(\Omega, \mathcal{A}, \mathbb{P}) := \left\{ \text{ v.a. } X : \|X\|_{L^p} := \mathbb{E}\left[|X|^p\right]^{1/p} < +\infty \right\}.$$

Ci-dessus, l'espérance est définie de la manière suivante : si $h : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ est une fonction mesurable alors $h(X)$ est une v.a. et

$$\mathbb{E}[h(X)] := \int_\Omega h(X(\omega)) d\mathbb{P}(\omega),$$

sous réserve que l'espérance est bien définie au sens de l'intégrale de Lebesgue, c'est-à-dire que $h(X)$ est positive ou intégrable par rapport à la mesure de probabilité $\mathbb{P}$. Par exemple, toutes les v.a. dont l'ensemble des valeurs est inclus dans un compact sont dans tous les L^p. Même chose pour les v.a. de Poisson, géométrique, exponentielle et normale. En revanche une v.a. X suivant la loi de Cauchy sur $\mathbb{R}$, i.e. dont la densité f_X est donnée par

$$f_X(x) := \frac{1}{\pi(1 + x^2)}, \quad x \in \mathbb{R},$$

n'appartient à aucun des L^p. On peut montrer que cet ensemble est un espace vectoriel et que $\|\cdot\|_{L^p}$ est une norme, donc c'est un espace vectoriel normé. Par le Théorème de Riesz-Fischer, ces espaces L^p sont des espaces de Banach (espaces vectoriels normés complets, la complétude signifiant que toute de Cauchy converge pour la norme associée) et un espace de Hilbert pour $p = 2$, la norme L^2 dérivant du produit scalaire $\langle X, Y\rangle := \mathbb{E}[XY]$.

Le cas $p = +\infty$ est quelque peu différent. On note $L^\infty(\mathcal{A}) := L^\infty(\Omega, \mathcal{A}, \mathbb{P})$ (ou simplement L^∞ s'il n'y a pas de confusion possible) l'ensemble des v.a. bornées p.s., c'est-à-dire

$$L^\infty(\Omega, \mathcal{A}, \mathbb{P}) := \{ \text{ v.a. } X : \|X\|_{L^\infty} := \inf\{c > 0 : \mathbb{P}(|X| \leq c) = 1\} < \infty\}.$$

Par exemple, c'est le cas des v.a. discrètes dont l'ensemble des valeurs est fini, comme une v.a. de Bernoulli, binomiale, uniforme sur un intervalle, tandis que les v.a. de Poisson, géométrique, exponentielle, normale et Cauchy ne le sont pas. Tout comme les espaces L^p, l'espace L^∞ est un espace vectoriel et $\|\cdot\|_{L^\infty}$ est une norme sur cet espace. En revanche ce n'est pas un espace de Banach car il n'est pas complet.

Si $X \in L^1$, elle est dite intégrable et si de surcroît elle est dans l'espace L^2, alors on dit qu'elle est de carré intégrable et sa variance

$$\mathrm{Var}(X) := \mathbb{E}\left[|X - \mathbb{E}[X]|^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2,$$

est positive ou nulle (grâce à l'inégalité de Cauchy-Schwarz que l'on va introduire incessamment sous peu). En pratique, on exprime ces espérances comme intégrales par rapport à la loi P_X. Par exemple, le p-ième moment se réécrit comme

$$\mathbb{E}[|X|^p] = \int_{\mathbb{R}} |x|^p dP_X(x).$$

Plus généralement, si $h : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ est une fonction mesurable alors si $h(X)$ est à valeurs positives ou $h(X) \in L^1$, on a

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x) dP_X(x).$$

Ainsi, on a $h(X) \in L^1$ si et seulement si h est intégrable sur $\mathbb{R}$ par rapport à la loi P_X. Dans le cas des v.a. discrètes, si $P_X = \sum_n \mathbb{P}(X = x_n)\delta_{x_n}$, alors

$$\mathbb{E}[h(X)] = \sum_n h(x_n)\mathbb{P}(X = x_n),$$

tandis que si X admet une densité f_X, on a

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x) f_X(x) dx.$$

À titre d'exemple, une fonction indicatrice étant intégrable par rapport à toute mesure finie, la fonction de répartition F_X de la v.a. X est définie par le choix de la fonction $h = \mathbf{1}_{]-\infty,t]}$:

$$F_X(t) := \mathbb{E}\left[\mathbf{1}_{]-\infty,t]}(X)\right] = \mathbb{P}(X \leq t) = \int_{\mathbb{R}} \mathbf{1}_{]-\infty,t]}(x) dP_X(x), \quad t \in \mathbb{R}.$$

En étendant la notion la notion d'espérance pour des fonctions h mesurables à valeurs dans le plan complexe $\mathbb{C}$, le choix de l'exponentielle complexe $h(x) = e^{itx}$, qui est bien une fonction intégrable par rapport à la loi P_X car de module 1, donne lieu à la fonction caractéristique de X :

$$\varphi_X(t) := \mathbb{E}[e^{itX}] = \int_{\mathbb{R}} e^{itx} dP_X(x), \quad t \in \mathbb{R}.$$

De plus, lorsque X admet une densité f_X, alors la fonction caractéristique est la transformée de Fourier de la densité f_X (ce concept ne sera abordé en détail qu'au second semestre dans le cours de Signal).

Pour terminer cette partie, on mentionnera deux points très importants en pratique : la fonction de répartition et la fonction caractéristique sont deux objets qui caractérisent la loi, c'est-à-dire que si deux v.a. ont même fonction de répartition (respectivement même fonction caractéristique), alors elles ont la même loi.

1.3 Inégalités classiques

En théorie de la mesure, certains inégalités sont à la base de résultats très importants. Les versions probabilistes de ces inégalités sont les suivantes.

1. L'inégalité de Markov : si $X \in L^1$ alors

$$\mathbb{P}(|X| \geq x) \leq \frac{\mathbb{E}[|X|]}{x}, \quad x > 0.$$

2. L'inégalité de Chebyshev : en choisissant dans l'inégalité précédente $X = (Y - \mathbb{E}[Y])^2$, où $Y \in L^2$, on obtient

$$\mathbb{P}\left(|Y - \mathbb{E}[Y]| \geq x\right) \leq \frac{\mathrm{Var}(Y)}{x^2}, \quad x > 0.$$

3. L'inégalité de Hölder : si $X \in L^p$ et $Y \in L^q$, où les exposants $p, q > 1$ sont conjugués, à savoir $p^{-1} + q^{-1} = 1$, alors

$$\mathbb{E}\left[|XY|\right] \leq \mathbb{E}\left[|X|^p\right]^{1/p} \mathbb{E}\left[|Y|^q\right]^{1/q},$$

ou encore en terme de normes,

$$\|XY\|_{L^1} \leq \|X\|_{L^p} \|Y\|_{L^q}.$$

Pour $p = q = 2$, c'est la célèbre inégalité de Cauchy-Schwarz dont nous avons déjà parlé pour montrer que la variance d'une v.a. est toujours positive ou nulle (exercice).

4. L'inégalité de Jensen : si $\varphi : \mathbb{R} \to \mathbb{R}$ est une fonction convexe telle que $\varphi(X) \in L^1$, alors on a l'inégalité
$$\varphi\left(\mathbb{E}[X]\right) \leq \mathbb{E}[\varphi(X)].$$

1.4 Indépendance

L'indépendance est une notion-clé de la théorie des probabilités. Dans ce qui suit, nous ne considérons par souci de simplicité que des couples de v.a., mais ceci se généralise aisément au cas des vecteurs aléatoires, c'est-à-dire des vecteurs dont chaque coordonnée est une v.a. Rappelons que deux v.a. X et Y sont dites indépendantes si pour tous $A, B \in \mathcal{B}(\mathbb{R})$, on a l'égalité
$$\mathbb{P}(X \in A; Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B),$$
le point-virgule représentant l'intersection entre deux évènements. Du point de vue des lois de X et Y, l'égalité précédente se réécrit comme
$$P_{X,Y}(A \times B) = P_X(A)\,P_Y(B).$$
Ci-dessus, la loi $P_{X,Y}$ est la loi jointe du couple (X, Y) vu comme un vecteur aléatoire à valeurs dans l'espace produit $\mathbb{R}^2$. Ainsi, l'indépendance entre deux v.a. se traduit simplement par le fait que la loi jointe $P_{X,Y}$ du couple (X, Y) est le produit tensoriel $P_X \otimes P_Y$ des deux lois marginales P_X et P_Y. Dans le cas des variables admettant une densité, la densité jointe est le produit des densités marginales.

Notons que l'égalité précédente se réécrit comme
$$\mathbb{E}[\mathbf{1}_A(Y)\mathbf{1}_B(Z)] = \mathbb{E}[\mathbf{1}_{A\times B}(Y, Z)] = \mathbb{E}[\mathbf{1}_A(Y)]\,\mathbb{E}[\mathbf{1}_B(Z)].$$
Autrement dit, deux v.a. sont indépendantes si et seulement si l'espérance du produit d'indicatrices vaut le produit des espérances de ces indicatrices. Un résultat important de la théorie de la mesure nous disant que toute fonction mesurable peut être approchée simplement par une suite de fonctions étagées (i.e. une somme finie de fonctions indicatrices pondérées), on en déduit que l'indépendance peut être reformulée de la manière suivante : deux v.a. X et Y sont indépendantes si et seulement si l'égalité
$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)],$$
a lieu pour toutes fonctions $f, g : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mesurables positives ou bornées. En particulier, l'indépendance est une notion plus forte que la non-corrélation, c'est-à-dire que
$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y].$$
En effet, pour obtenir cette égalité, le choix des deux fonctions f et g est imposé (la fonction identité pour les deux).

Chapitre 2

Espérance conditionnelle

Pour de nombreux problèmes concrets, il est important de pouvoir calculer des quantités relatives à des v.a. pour lesquelles seule une information partielle est disponible. Dès lors, on comprend l'importance de la notion de conditionnement, comme ceci a déjà été abordé dans le passé avec les probabilités conditionnelles. Ainsi, l'espérance par rapport à une loi conditionnelle donnée pourrait correspondre à une bonne définition d'espérance conditionnelle. Néanmoins, l'espérance conditionnelle telle que nous souhaiterions la définir doit être plus générale et prendre en compte le conditionnement par rapport à toute l'information relative à une ou plusieurs v.a., voire au-delà via la notion de tribu. C'est l'objet de ce chapitre.

2.1 Introduction à l'espérance conditionnelle

Démarrons en introduisant la notion d'espérance conditionnelle dans un cas simple, celui des tribus engendrées par une partition.

Définition 2.1.1. *Soit $\mathcal{F}$ une sous-tribu de $\mathcal{A}$, engendrée par une partition $(A_n)_{n\in\mathbb{N}^*}$ de Ω, à savoir*

$$\begin{cases} \bigcup_{n\in\mathbb{N}^*} A_n &= \Omega \\ A_i \cap A_j &= \emptyset \quad \textit{pour} \quad i \neq j. \end{cases}$$

Notons $\mathcal{N} := \{n \in \mathbb{N}^* : \mathbb{P}(A_n) > 0\}$. On appelle probabilité conditionnelle d'un événement $A \in \mathcal{A}$ sachant la tribu $\mathcal{F}$ la quantité

$$\mathbb{P}(A \,|\, \mathcal{F})(\omega) := \sum_{n\in\mathcal{N}} \mathbb{P}(A \,|\, A_n)\, \mathbf{1}_{A_n}(\omega), \quad \omega \in \Omega,$$

où $\mathbf{1}_{A_n}(\omega)$ est l'indicatrice valant 1 si $\omega \in A_n$ et 0 sinon.

Notons que cette probabilité conditionnelle est une v.a., le conditionnement ayant lieu par rapport à une tribu et non un événement. De surcroît, elle est mesurable par rapport à la tribu $\mathcal{F}$ car c'est une somme pondérée de fonctions indicatrices des A_n, ces

derniers événements appartenant à la tribu $\mathcal{F}$. Enfin, elle est constante sur les A_n et vaut alors $\mathbb{P}(A \,|\, A_n)$, la probabilité conditionnelle telle que nous la connaissons.

En passant à l'espérance dans la définition précédente et en permutant somme et espérance (ce qui est permis car en réalité la somme ne comporte qu'un seul terme à cause de la présence de l'indicatrice), on a

$$\begin{aligned}\mathbb{E}\left[\mathbb{P}(A \,|\, \mathcal{F})\right] &= \sum_{n\in\mathcal{N}} \mathbb{P}(A \,|\, A_n)\, \mathbb{E}\left[\mathbf{1}_{A_n}\right] \\ &= \sum_{n\in\mathcal{N}} \mathbb{P}(A \,|\, A_n)\, \mathbb{P}(A_n) \\ &= \mathbb{P}(A),\end{aligned}$$

la dernière égalité étant obtenue par la célèbre formule des probabilités totales. Ainsi, la probabilité conditionnelle par rapport à une tribu vaut, en moyenne, la probabilité initiale.

En utilisant une reformulation avec les indicatrices, la définition de la probabilité conditionnelle devient

$$\mathbb{E}\left[\mathbf{1}_A \,|\, \mathcal{F}\right] := \sum_{n\in\mathcal{N}} \mathbb{E}\left[\mathbf{1}_A \,|\, A_n\right]\, \mathbf{1}_{A_n},$$

la quantité intervenant dans le membre de droite étant comprise comme

$$\mathbb{E}\left[\mathbf{1}_A \,|\, A_n\right] := \mathbb{P}(A \,|\, A_n) = \frac{\mathbb{P}(A \cap A_n)}{\mathbb{P}(A_n)} = \frac{\mathbb{E}\left[\mathbf{1}_{A\cap A_n}\right]}{\mathbb{P}(A_n)} = \frac{\mathbb{E}\left[\mathbf{1}_A\, \mathbf{1}_{A_n}\right]}{\mathbb{P}(A_n)}.$$

On en déduit que cette probabilité conditionnelle est obtenue par "moyennisation" de la v.a. $\mathbf{1}_A$ sur les événements engendrant $\mathcal{F}$. L'étape suivante est donc de remplacer cette indicatrice par une v.a. intégrable, comme on le fait dans le cours de théorie de la mesure pour construire l'intégrale de Lebesgue (passage des fonctions indicatrices aux fonctions étagées puis aux fonctions mesurables positives puis enfin aux fonctions intégrables).

Définition 2.1.2. *Soit $X \in L^1(\mathcal{A})$ et $\mathcal{F}$ une sous-tribu de $\mathcal{A}$ engendrée par une partition $(A_n)_{n\in\mathbb{N}^*}$ de Ω. On appelle espérance conditionnelle de X sachant la tribu $\mathcal{F}$ la v.a.*

$$\mathbb{E}\left[X \,|\, \mathcal{F}\right](\omega) := \sum_{n\in\mathcal{N}} \mathbb{E}\left[X \,|\, A_n\right]\, \mathbf{1}_{A_n}(\omega), \quad \omega \in \Omega,$$

où $\mathbb{E}\left[X \,|\, A_n\right]$ désigne le ratio

$$\mathbb{E}\left[X \,|\, A_n\right] := \frac{\mathbb{E}\left[X\mathbf{1}_{A_n}\right]}{\mathbb{P}(A_n)}.$$

Remarquons plusieurs propriétés. Tout d'abord, l'espérance conditionnelle est clairement $\mathcal{F}$-mesurable, linéaire et p.s. positive si X l'est. De plus, un exercice classique est

de montrer que si $X \in L^2(\mathcal{A})$, alors par l'inégalité de Cauchy-Schwarz l'espérance conditionnelle est aussi de carré integrable. De plus elle vaut en moyenne l'espérance de X. En effet, en permutant de nouveau espérance et somme, on a

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[X \,|\, \mathcal{F}\right]\right] &= \sum_{n\in\mathcal{N}} \mathbb{E}\left[X \,|\, A_n\right] \mathbb{E}\left[\mathbf{1}_{A_n}\right] \\ &= \sum_{n\in\mathcal{N}} \mathbb{E}\left[X\,\mathbf{1}_{A_n}\right] \\ &= \mathbb{E}\left[X \sum_{n\in\mathcal{N}} \mathbf{1}_{A_n}\right] \\ &= \mathbb{E}\left[X\right],\end{aligned}$$

la famille $(A_n)_{n\in\mathbb{N}^*}$ formant une partition de Ω. Par ailleurs, si X s'écrit $X = \sum_{i=1}^{p} \alpha_i \mathbf{1}_{A_i}$ avec $p \in \mathbb{N}^* \cup \{+\infty\}$ et où les α_i sont des constantes, alors X est mesurable par rapport à la tribu $\mathcal{F}$ et

$$\mathbb{E}\left[X \,|\, \mathcal{F}\right] = \sum_{n\in\mathcal{N}} \sum_{i=1}^{p} \alpha_i \,\mathbb{E}\left[\mathbf{1}_{A_i} \,|\, A_n\right] \mathbf{1}_{A_n} = \sum_{i=1}^{p} \alpha_i \,\mathbf{1}_{A_i} = X,$$

en ayant utilisé encore une fois le fait que $(A_n)_{n\in\mathbb{N}^*}$ est une partition de Ω.
Si les tribus $\sigma(X)$ et $\mathcal{F}$ sont indépendantes, c'est-à-dire que tout événement $\sigma(X)$-mesurable est indépendant de tout événement $\mathcal{F}$-mesurable, alors on démontre facilement que

$$\mathbb{E}\left[X \,|\, \mathcal{F}\right] = \mathbb{E}\left[X\right].$$

Enfin, terminons par la propriété clé satisfaite par l'espérance conditionnelle, qui est à la base de la prochaine définition et qu'on laissera en exercice : pour toute v.a. $Z \in L^\infty(\mathcal{F})$, on a

$$\mathbb{E}\left[XZ\right] \,=\, \mathbb{E}\left[\mathbb{E}\left[X \,|\, \mathcal{F}\right]\, Z\right].$$

Ci-dessus on a noté $L^\infty(\mathcal{F})$ le sous-espace de $L^\infty(\mathcal{A})$ constitué des v.a. bornées p.s. et mesurables par rapport à la sous-tribu $\mathcal{F}$. On fera de même pour tous les espaces L^p.

2.2 Le cas général

À présent, on va généraliser l'espérance conditionnelle à une tribu arbitraire, non nécessairement engendrée par une partition. En particulier, on va pouvoir conditionner par une v.a. générale.

Définition et théorème 2.2.1. *Soit $X \in L^2(\mathcal{A})$ (respectivement $L^1(\mathcal{A})$) et soit $\mathcal{F}$ une sous-tribu quelconque de $\mathcal{A}$. Alors il existe une v.a. $Y \in L^2(\mathcal{F})$ (resp. $L^1(\mathcal{F})$) telle que pour toute v.a. $Z \in L^2(\mathcal{F})$ (resp. tout événement $A \in \mathcal{F}$),*

$$\mathbb{E}\left[XZ\right] \,=\, \mathbb{E}\left[YZ\right] \qquad (\textit{resp.} \quad \mathbb{E}\left[X\mathbf{1}_A\right] \,=\, \mathbb{E}\left[Y\mathbf{1}_A\right]).$$

On note $Y = \mathbb{E}\left[X \,|\, \mathcal{F}\right]$: c'est l'espérance conditionnelle de X sachant la tribu $\mathcal{F}$.

Démonstration. Nous n'allons faire que la démonstration dans $L^2(\mathcal{F})$, le cas $L^1(\mathcal{F})$ se traitant ensuite par densité et un passage à la limite. Posons $F = L^2(\mathcal{F})$, qui est un espace de Hilbert donc complet. De plus, F étant inclus dans $L^2(\mathcal{A})$, il est fermé (tout sous-espace complet d'un espace métrique, non nécessairement complet, est fermé). Ainsi, par le théorème du supplémentaire orthogonal d'un fermé dans un espace de Hilbert, on a

$$L^2(\mathcal{A}) = F \oplus F^\perp,$$

c'est-à-dire que l'espace $L^2(\mathcal{A})$ se décompose comme la somme directe de F et de son orthogonal $F^\perp$ défini par

$$F^\perp := \{U \in L^2(\mathcal{A}) : \mathbb{E}[UZ] = 0 \text{ pour tout } Z \in F\}.$$

En d'autres termes, toute v.a. de $L^2(\mathcal{A})$ se décompose de manière unique comme la somme de 2 v.a., l'une dans F et l'autre dans $F^\perp$. Lorsque l'on applique ce résultat à X on obtient l'existence et l'unicité d'une v.a. $Y \in F$ telle que $X = Y + (X - Y)$ où $X - Y \in F^\perp$. Ainsi, on en déduit que pour tout $Z \in F$,

$$\mathbb{E}[(X - Y)Z] = 0,$$

c'est-à-dire le résultat désiré. ∎

Cette définition généralise le cas de l'espérance conditionnelle définie par rapport à une tribu engendrée par une partition. Notons par ailleurs que nous imposons de l'intégrabilité dans la définition de l'espérance conditionnelle afin de s'assurer que cette dernière ait un sens, mais cette hypothèse peut bien évidemment être relaxée en supposant seulement que $X \geq 0$ p.s., auquel cas $\mathbb{E}[X \,|\, \mathcal{F}]$ peut prendre la valeur $+\infty$. De surcroît, on peut montrer que l'on ne change pas la notion d'espérance conditionnelle dans le cas intégrable en remplaçant les indicatrices $\mathbf{1}_A$ par des v.a. $Z \in L^\infty(\mathcal{F})$.

Comme pour le cas de l'espérance conditionnelle par rapport à une tribu engendrée par une partition, l'espérance conditionnelle est clairement linéaire (utiliser dans la définition de l'espérance conditionnelle la linéarité de l'espérance ainsi que l'unicité p.s. de l'espérance conditionnelle) et est p.s. positive si X l'est. En effet, supposons $X \geq 0$ p.s. et montrons que $\mathbb{E}[X \,|\, \mathcal{F}] \geq 0$ p.s. Pour ce faire, choisissons $A = \{\mathbb{E}[X \,|\, \mathcal{F}] < 0\}$ qui est un évènement appartenant bien à la tribu $\mathcal{F}$. Ainsi, on a

$$0 \leq \mathbb{E}[X\mathbf{1}_{\{\mathbb{E}[X \,|\, \mathcal{F}]<0\}}] = \mathbb{E}[\mathbb{E}[X \,|\, \mathcal{F}]\mathbf{1}_{\{\mathbb{E}[X \,|\, \mathcal{F}]<0\}}] \leq 0,$$

d'où le fait que $\mathbb{E}[X\mathbf{1}_{\{\mathbb{E}[X \,|\, \mathcal{F}]<0\}}] = 0$. Comme $X\mathbf{1}_{\{\mathbb{E}[X \,|\, \mathcal{F}]<0\}} \geq 0$ p.s., on en déduit que

$$X\mathbf{1}_{\{\mathbb{E}[X \,|\, \mathcal{F}]<0\}} = 0,$$

et donc que $\mathbb{E}[X \,|\, \mathcal{F}] \geq 0$ p.s.

2.3 Les résultats importants

En pratique, les propriétés de l'espérance conditionnelle que l'on utilise le plus souvent sont les suivantes.

Proposition 2.3.1. *Soit $X \in L^1(\mathcal{A})$. Alors l'espérance conditionnelle vérifie les propriétés suivantes.*

1. $\mathbb{E}\left[\mathbb{E}\left[X \,|\, \mathcal{F}\right]\right] = \mathbb{E}\left[X\right]$.
2. *Si les tribus $\sigma(X)$ et $\mathcal{F}$ sont indépendantes, alors* $\mathbb{E}\left[X \,|\, \mathcal{F}\right] = \mathbb{E}\left[X\right]$.
3. *Si X est $\mathcal{F}$-mesurable, alors* $\mathbb{E}\left[X \,|\, \mathcal{F}\right] = X$.
4. *Si $\mathcal{G} \subset \mathcal{F}$ alors* $\mathbb{E}\left[\mathbb{E}\left[X \,|\, \mathcal{F}\right] \,|\, \mathcal{G}\right] = \mathbb{E}\left[X \,|\, \mathcal{G}\right]$.
5. *Si Y est une v.a. $\mathcal{F}$-mesurable avec de surcroît $XY \in L^1(\mathcal{A})$, alors*
$$\mathbb{E}\left[XY \,|\, \mathcal{F}\right] = Y\,\mathbb{E}\left[X \,|\, \mathcal{F}\right].$$
6. *Supposons que X est $\mathcal{F}$-mesurable et soit Y une v.a. indépendante de $\mathcal{F}$. Soit $h : (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ une application mesurable telle que $h(X,Y) \in L^1(\mathcal{A})$. Alors on a*
$$\mathbb{E}[h(X,Y) \,|\, \mathcal{F}] = H(X), \qquad \text{avec} \qquad H(x) = \mathbb{E}\left[h(x,Y)\right].$$
7. *Si $X \in L^2(\mathcal{A})$, alors $\mathbb{E}\left[X \,|\, \mathcal{F}\right]$ est le projeté orthogonal de X sur le sous-espace fermé $L^2(\mathcal{F})$.*

Démonstration.
1. Prendre $Z = 1$ qui est bien $\mathcal{F}$-mesurable et bornée.
2. Supposons les tribus $\sigma(X)$ et $\mathcal{F}$ indépendantes. Alors pour toute v.a. $Z \in L^\infty(\mathcal{F})$,
$$\mathbb{E}\left[XZ\right] \;=\; \mathbb{E}\left[X\right]\,\mathbb{E}\left[Z\right].$$
Or $\mathbb{E}\left[X\right]$ étant constante, elle est bien $\mathcal{F}$-mesurable et donc on obtient le résultat par unicité de l'espérance conditionnelle.
3. Même raisonnement que précédemment.
4. et 5. Il suffit de l'écrire (un peu pénible tout de même).
6. Démonstration admise.
7. La v.a. $Y := \mathbb{E}[X \,|\, \mathcal{F}]$ est le projeté orthogonal de X sur le fermé $F := L^2(\mathcal{F})$ si et seulement si
$$Y \in F \quad \text{ and } \quad \|X - Y\|_{L^2} \;=\; \inf_{Z \in F} \|X - Z\|_{L^2}.$$
On sait déjà que $Y \in F$. Ainsi, soit $Z \in F$ et posons $U := X - Y \in F^\perp$ et $V := Y - Z \in F$, comme différence de deux éléments de F. Alors $\mathbb{E}\left[UV\right] = 0$ et l'on obtient
$$\mathbb{E}\left[(X-Z)^2\right] \;=\; \mathbb{E}\left[(U+V)^2\right]$$

$$\begin{aligned} &= \mathbb{E}\left[U^2\right] + \mathbb{E}\left[V^2\right] + 2\,\mathbb{E}\left[UV\right] \\ &= \mathbb{E}\left[U^2\right] + \mathbb{E}\left[V^2\right], \end{aligned}$$

quantité qui est minimale pour le choix de $V = 0$, i.e. $Z = Y$. ∎

Mentionnons que l'espérance conditionnelle par rapport à la tribu $\sigma(X)$ se note classiquement $\mathbb{E}[\,\cdot\,|\,X]$. Si l'on considère une v.a. Y intégrable, alors vu que $\mathbb{E}[Y\,|\,X]$ est $\sigma(X)$-mesurable, le Lemme de Doob indique qu'il existe une fonction mesurable $h : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ telle que

$$\mathbb{E}[Y\,|\,X] = h(X).$$

Dans ce cas, on note $\mathbb{E}[Y\,|\,X = x]$ la quantité $h(x)$, pour tout $x \in \mathbb{R}$. Attention, cette notation est quelque peu abusive car si X admet une densité alors $\mathbb{P}(X = x) = 0$ pour tout $x \in \mathbb{R}$ et l'on sait qu'il est interdit de conditionner par un événement de probabilité nulle.

2.4 L'espérance conditionnelle en pratique

Pour terminer ce chapitre, regardons en pratique comment calculer une espérance conditionnelle du type $\mathbb{E}[Y\,|\,X = x]$ lorsque l'on a des informations sur la loi jointe du couple (X, Y).

1. Cas discret: Tout d'abord donnons-nous deux v.a. discrètes X et Y, chacune à valeurs dans des espaces au plus dénombrables E et F respectivement. On suppose de plus que Y est intégrable et que $\mathbb{P}(X = x) > 0$ pour un $x \in E$ donné. Alors l'espérance conditionnelle $\mathbb{E}[Y\,|\,X = x]$ est bien définie et vaut

$$\mathbb{E}[Y\,|\,X = x] = \sum_{y \in F} y\,\mathbb{P}(Y = y\,|\,X = x) = \sum_{y \in F} y\,\frac{\mathbb{P}(X = x; Y = y)}{\mathbb{P}(X = x)}.$$

   Pour obtenir l'expression explicite de l'espérance conditionnelle $\mathbb{E}[Y\,|\,X]$, il reste à remplacer x par la v.a. X dans la formule précédente, la tribu $\sigma(X)$ étant engendrée par la partition $\{X = x\}_{x \in E}$.

2. Cas continu: à présent, si X et Y sont deux v.a. admettant pour densité f_X et f_Y respectivement, alors en supposant de plus que Y est intégrable et que f_X vérifie $f_X(x) > 0$ pour un $x \in \mathbb{R}$ donné, on a

$$\mathbb{E}[Y\,|\,X = x] = \int_{\mathbb{R}} y\, f_Y^{X=x}(y)\, dy = \int_{\mathbb{R}} y\,\frac{f_{X,Y}(x, y)}{f_X(x)}\, dy,$$

   où $f_Y^{X=x}$ désigne la densité conditionnelle de Y sachant $X = x$ et $f_{X,Y}$ est la densité jointe du couple (X, Y). Là aussi, pour en déduire l'expression explicite de l'espérance conditionnelle $\mathbb{E}[Y\,|\,X]$, on remplace x par la v.a. X dans le résultat obtenu et le tour est joué.

Chapitre 3

Vecteurs gaussiens

L'objectif de ce chapitre est de généraliser au cadre mutidimensionnel la notion de v.a. gaussienne : on parle alors de vecteurs gaussiens. Comme en dimension 1, ces vecteurs aléatoires apparaissent naturellement comme des objets limites et par conséquent jouent un rôle important en théorie des Probabilités et en Statistique. En particulier, le caractère multidimensionnel de ces vecteurs aléatoires rend leur étude à la fois plus intéressante et plus délicate car il faut prendre en compte la structure de dépendance entre les coordonnées.

3.1 Définition et premières propriétés

Si X est un vecteur aléatoire en dimension d (i.e. une v.a. à valeurs dans $\mathbb{R}^d$ vue comme un vecteur colonne) tel que chacune de ses coordonnées X_i, $i \in \{1, \dots, d\}$, soit dans L^2, alors on définit dans la suite son vecteur espérance et sa matrice de covariance par

$$\mathbb{E}[X] = m := \begin{pmatrix} \mathbb{E}[X_1] \\ \cdot \\ \cdot \\ \cdot \\ \mathbb{E}[X_d] \end{pmatrix} \qquad \text{et} \qquad \mathrm{Var}(X) = \Gamma := \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,d} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,d} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sigma_{d,1} & \sigma_{d,2} & \cdots & \sigma_{d,d} \end{pmatrix},$$

où $\sigma_{i,j}$ désigne la covariance entre les variables X_i et X_j:

$$\sigma_{i,j} \;=\; \mathrm{Cov}(X_i, X_j) \;:=\; \mathbb{E}[(X_i - m_i)(X_j - m_j)] \;=\; \mathbb{E}[X_i X_j] - m_i\, m_j,$$

où $m_i := \mathbb{E}[X_i]$ pour tout $i \in \{1, \dots, d\}$. En particulier, cette covariance est bien définie par l'inégalité de Cauchy-Schwarz. De plus, on remarque que la matrice réelle Γ est symétrique et qu'elle peut s'écrire comme

$$\Gamma \;=\; \mathbb{E}\left[(X - m)(X - m)^T\right] \;=\; \mathbb{E}\left[X X^T\right] - m\, m^T,$$

où le symbole T désigne la transposition et l'espérance d'une matrice est définie comme la matrice des espérances de chacun de ses éléments. On en déduit qu'elle est semi-définie positive au sens où pour tout $x \in \mathbb{R}^d$,

$$\langle x, \Gamma x \rangle \;=\; \mathbb{E}\left[\langle x, X - m\rangle^2\right] \;\geq\; 0,$$

où $\langle a, b\rangle = a^T b$ est le produit scalaire euclidien dans $\mathbb{R}^d$ entre deux vecteurs donnés $a, b \in \mathbb{R}^d$.

Dans la suite de ce chapitre, on supposera que les vecteurs gaussiens considérés sont non dégénérés, c'est-à-dire que $\det \Gamma \neq 0$ (la matrice de covariance est donc définie positive, i.e. $\langle x, \Gamma x\rangle > 0$ pour tout $x \neq 0$ dans $\mathbb{R}^d$).

Définition 3.1.1. *Soit X un vecteur aléatoire en dimension d de vecteur espérance m et de matrice de covariance Γ. Il est dit gaussien si la densité jointe est donnée par*

$$f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Gamma}} \exp\left(-\frac{1}{2}\langle x - m, \Gamma^{-1}(x-m)\rangle\right), \quad x \in \mathbb{R}^d.$$

On note alors $X \sim \mathcal{N}_d(m, \Gamma)$ (et $\mathcal{N}(m, \Gamma)$ si $d = 1$).

Ainsi, comme en dimension 1, la donnée du vecteur espérance et de la matrice de covariance caractérise la loi d'un vecteur gaussien. De surcroît, on a la caractérisation suivante de la loi d'un vecteur gaussien en exprimant sa fonction caractéristique, une situation identique à celle du cas unidimensionnel.

Proposition 3.1.2. *Un vecteur aléatoire X est gaussien de vecteur espérance m et de matrice de covariance Γ si et seulement si sa fonction caractéristique (multidimensionnelle) est donnée par*

$$\varphi_X(\theta) := \mathbb{E}[e^{i\langle \theta, X\rangle}] = \exp\left(i\langle \theta, m\rangle - \frac{1}{2}\langle \theta, \Gamma\theta\rangle\right), \quad \theta \in \mathbb{R}^d.$$

À présent, posons-nous la question suivante : un vecteur gaussien a-t-il toutes ses composantes unidimensionnelles gaussiennes ? Et réciproquement, suffit-il d'avoir toutes les coordonnées gaussiennes pour que le vecteur associé soit gaussien ? La proposition suivante permet de répondre à la question.

Proposition 3.1.3. *Un vecteur aléatoire X est gaussien si et seulement si $\langle \theta, X\rangle$ est une v.a. gaussienne unidimensionnelle pour tout $\theta \in \mathbb{R}^d$ différent du vecteur nul, i.e. toute combinaison linéaire non nulle de ses coordonnées est gaussienne.*

Démonstration. Notons respectivement m et Γ le vecteur espérance et la matrice de covariance du vecteur X. Supposons le gaussien. Alors pour tout $\theta \in \mathbb{R}^d$ différent du vecteur nul, la fonction caractéristique de la v.a. $\langle \theta, X\rangle$ s'écrit pour tout $u \in \mathbb{R}$:

$$\varphi_{\langle \theta, X\rangle}(u) \;=\; \mathbb{E}\left[e^{iu\langle \theta, X\rangle}\right]$$

$$\begin{aligned} &= \varphi_X(u\theta) \\ &= \exp\left(iu\langle\theta, m\rangle - \frac{u^2}{2}\langle\theta, \Gamma\theta\rangle\right). \end{aligned}$$

Ainsi, on en déduit que pour tout $\theta \in \mathbb{R}^d$ non nul, la v.a. $\langle\theta, X\rangle$ suit une loi gaussienne d'espérance $\langle\theta, m\rangle$ et de variance $\langle\theta, \Gamma\theta\rangle$, qui est strictement positive car Γ est définie positive.
Réciproquement, si $\langle\theta, X\rangle$ suit une loi gaussienne pour tout $\theta \in \mathbb{R}^d$ non nul, alors

$$\begin{cases} \mathbb{E}\left[\langle\theta, X\rangle\right] &= \mathbb{E}\left[\sum_{i=1}^d \theta_i X_i\right] &= \langle\theta, m\rangle; \\ \mathrm{Var}\left(\langle\theta, X\rangle\right) &= \mathrm{Var}\left(\sum_{i=1}^d \theta_i X_i\right) &= \sum_{i,j=1}^d \theta_i\theta_j \,\mathrm{Cov}(X_i, X_j) \;=\; \langle\theta, \Gamma\theta\rangle. \end{cases}$$

Alors on a pour tout $u \in \mathbb{R}$

$$\varphi_{\langle\theta,X\rangle}(u) \;=\; \exp\left(iu\langle\theta, m\rangle - \frac{u^2}{2}\langle\theta, \Gamma\theta\rangle\right),$$

et en choisissant $u = 1$ on obtient que pour tout $\theta \in \mathbb{R}^d$ non nul,

$$\varphi_X(\theta) \;=\; \exp\left(i\langle\theta, m\rangle - \frac{1}{2}\langle\theta, \Gamma\theta\rangle\right),$$

c'est-à-dire que $X \sim \mathcal{N}_d(m, \Gamma)$. ∎

Ainsi, chacune des coordonnées d'un vecteur gaussien suit forcément une loi gaussienne. En revanche, la réciproque est fausse : il se peut que 2 v.a. Y et Z soient gaussiennes sans que le vecteur $(Y, Z)^T$ soit un vecteur gaussien. En effet, considérons une v.a. $Y \sim \mathcal{N}(0, 1)$, indépendante d'une variable ε dite de Rademacher, de loi donnée par

$$\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = \frac{1}{2}.$$

On peut montrer que $Z = \varepsilon Y$ est une v.a. gaussienne alors que le vecteur $(Y, Z)^T$ ne l'est pas. Ainsi, demander à ce qu'un vecteur aléatoire soit gaussien requiert plus que le caractère gaussien des coordonnées.

Le résultat qui suit est en quelque sorte un "miracle" du cadre gaussien, au sens où il suffit de démontrer la nullité des différentes covariances afin d'obtenir l'indépendance des coordonnées.

Proposition 3.1.4 (Miracle gaussien)**.** *Soit X un vecteur gaussien. Alors ses coordonnées sont indépendantes si et seulement si sa matrice de covariance est diagonale.*

Démonstration. Les coordonnées du vecteur gaussien X sont indépendantes si et seulement si sa densité f_X est le produit des densités des coordonnées. Ceci équivaut à dire que la forme quadratique $\langle x - m, \Gamma^{-1}(x - m)\rangle$ apparaissant dans f_X s'écrit de la forme $\sum_{i=1}^d \alpha_i (x_i - m_i)^2$. En d'autres termes, il n'y a aucun terme croisé du type $(x_i - m_i)(x_j - m_j)$ pour $i \neq j$, c'est-à-dire que la matrice Γ^{-1} (donc Γ) est diagonale. ∎

3.2 Autres propriétés importantes

Donnons dans ce bref paragraphe quelques propriétés que l'on rencontre usuellement lorsque l'on s'attaque à un problème faisant intervenir des vecteurs gaussiens.

Proposition 3.2.1 (Transformation linéaire)**.** *Soit A une matrice inversible $d \times d$ et $b \in \mathbb{R}^d$. Considérons le vecteur gaussien $X \sim \mathcal{N}_d(m, \Gamma)$. Alors le vecteur $AX + b$ est lui aussi gaussien :*

$$AX + b \sim \mathcal{N}_d(Am + b, A\Gamma A^T).$$

En particulier, si $X \sim \mathcal{N}_d(0, \sigma^2 I_d)$ et la matrice A est orthogonale, i.e. $AA^T = I_d$, alors les vecteurs X et AX ont même loi.

Démonstration. Très simple, il suffit de faire les calculs soit en utilisant la caractérisation des vecteurs gaussiens à l'aide des combinaisons linéaires (cf. TD), soit en manipulant la fonction caractéristique. On remarquera au passage que la matrice $A\Gamma A^T$ est bien semi-définie positive, et même définie positive car A est supposée inversible. ∎

Proposition 3.2.2. *Un vecteur aléatoire d-dimensionnel X est gaussien de vecteur espérance m et de matrice de covariance Γ si et seulement si X peut s'écrire*

$$X = AU + m,$$

où $U \sim \mathcal{N}_d(0, I_d)$ et A est une matrice carrée $d \times d$ inversible et vérifiant $AA^T = \Gamma$.

Démonstration. La matrice Γ est réelle symétrique donc diagonalisable dans une base orthonormale. De plus, étant définie positive, toutes ses valeurs propres λ_i sont strictement positives. Notons P la matrice de passage que l'on prend orthogonale, i.e. $PP^T = I_d$. Alors la matrice carrée $d \times d$ donnée par $A := PD$, où D est la matrice diagonale formée par les $\sqrt{\lambda_i}$, est inversible et satisfait $AA^T = \Gamma$. Enfin, la proposition précédente entraîne que

$$U := A^{-1}(X - m),$$

est un vecteur gaussien centré et de matrice de covariance l'identité. ∎

Proposition 3.2.3 (Projection gaussienne)**.** *Notons $X = (X_1, \ldots, X_d)^T$ et $Y = (Y_1, \cdots, Y_{d'})^T$ et supposons que le vecteur aléatoire $(X, Y)^T$ soit gaussien dans $\mathbb{R}^{d+d'}$. Alors la loi conditionnelle de Y sachant $X = x$ est elle-aussi gaussienne et il existe des constantes $a, b_1, \ldots, b_d \in \mathbb{R}^{d'}$ telles que*

$$\mathbb{E}[Y|X] = a + \sum_{i=1}^{d} b_i X_i.$$

De plus, le vecteur aléatoire $Y - \mathbb{E}[Y|X]$ est gaussien centré et indépendant de X.

Démonstration. Nous n'allons démontrer que le cas $d = d' = 1$, le cas général en étant une adaptation immédiate. Ainsi, soit $(X, Y)^T$ un vecteur gaussien dans $\mathbb{R}^2$. On montre tout d'abord que le vecteur $(X, Y - a - bX)^T$ est gaussien dans $\mathbb{R}^2$ pour tous $a, b \in \mathbb{R}$.

Soit (a_0, b_0) l'unique couple de valeurs pour lesquelles la quantité $\mathbb{E}\left[(Y - a - bX)^2\right]$ est minimale (ce couple peut être calculé explicitement). En d'autres termes, $a_0 + b_0 X$ est le projeté orthogonal de Y sur l'espace $\mathrm{Vect}\{1, X\} \subset L^2$. La v.a. gaussienne $\varepsilon_0 := Y - a_0 - b_0 X$ étant par conséquent dans l'espace $\mathrm{Vect}\{1, X\}^\perp$, on a alors que $\mathbb{E}[\varepsilon_0] = \mathbb{E}[\varepsilon_0 X] = 0$, c'est-à-dire que ε_0 est centrée et indépendante de X, le vecteur $(X, \varepsilon_0)^T$ étant gaussien dans $\mathbb{R}^2$ d'après ce qui précède. Il en résulte enfin que

$$\mathbb{E}[Y \mid X] = \mathbb{E}\left[\varepsilon_0 + a_0 + b_0 X \mid X\right] = a_0 + b_0 X,$$

d'où le résultat. ∎

Théorème 3.2.4 (Théorème Central Limite multidimensionnel)**.** *Soit $(X_n)_{n\geq 1}$ une suite de vecteurs aléatoires i.i.d. à valeurs dans $\mathbb{R}^d$, et dont les coordonnées sont de carré intégrable. Notons $m := \mathbb{E}[X_1]$ et $\Gamma := \mathrm{Var}(X_1)$. Alors on a la convergence en loi suivante :*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - m) \underset{n\to\infty}{\longrightarrow} \mathcal{N}_d(0, \Gamma).$$

Démonstration. Il s'agit d'une généralisation immédiate de celle du TCL unidimensionnel (remplacer les produits de nombres par des produits scalaires euclidiens). ∎

Théorème 3.2.5 (Cochran)**.** *Soit $X \sim \mathcal{N}_d(0, \sigma^2)$ avec $\sigma^2 > 0$ et $L_1 \oplus \cdots \oplus L_p$ une décomposition de $\mathbb{R}^d$ en sous-espaces orthogonaux de dimensions respectives $d_1, \ldots, d_p$. Si pour tout $i = 1, \ldots, p$, P_{L_i} désigne la matrice de la projection orthogonale sur L_i, alors les vecteurs $(P_{L_i} X)_{i=1,\ldots,p}$ sont des vecteurs gaussiens indépendants et pour tout $i = 1, \ldots, p$,*

$$\frac{\|P_{L_i} X\|^2}{\sigma^2} \sim \chi_2(d_i),$$

où $\|\cdot\|$ est la norme euclidienne sur $\mathbb{R}^d$ et $\chi_2(d_i)$ désigne la loi du χ_2 à d_i degrés de liberté, de densité donnée par

$$f(x) = \frac{1}{2^{d_i/2}\Gamma(d_i/2)}\, x^{d_i/2-1}\, e^{-x/2}, \quad x > 0.$$

Démonstration. Soit $(e_j^i)_{i,j}$ une base orthonormée de $\mathbb{R}^d$ telle que pour chaque $i = 1, \ldots, p$, $(e_j^i)_{j=1,\ldots,d_i}$ est une base orthonormée du sous-espace L_i. Alors pour chaque $i = 1, \ldots, p$, par définition de la matrice de la projection orthogonale sur L_i, on a pour tout $x \in \mathbb{R}^d$,

$$P_{L_i} x \;=\; \sum_{j=1}^{d_i} \frac{\langle x, e_j^i \rangle}{\|e_j^i\|^2}\, e_j^i \;=\; \sum_{j=1}^{d_i} \langle x, e_j^i \rangle\, e_j^i.$$

En particulier chaque $P_{L_i} X$ est un vecteur gaussien de dimension d_i (car toute combinaison linéaire non nulle des v.a. $(\langle X, e_j^i \rangle)_{j=1,\ldots,d_i}$ est une v.a. gaussienne). Les vecteurs $(e_j^i)_{i,j}$ étant orthogonaux, on a pour tous $i \neq k$,

$$\mathrm{Cov}(P_{L_i} X, P_{L_k} X) = 0.$$

Comme $(P_{L_1}X, \dots, P_{L_p}X)^T$ est un vecteur gaussien dans $\mathbb{R}^d$ (car toute combinaison linéaire non nulle des v.a. $\langle X, e_j^i \rangle$, $i = 1, \dots, p$ et $j = 1, \dots, d_i$, est une v.a. gaussienne), les $P_{L_i}X$ sont donc indépendants. Enfin, pour tout $i = 1, \dots, p$, les v.a. $(\langle X, e_j^i \rangle)_{j=1,\dots,d_i}$ sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Ainsi, on obtient

$$\frac{\|P_{L_i}X\|^2}{\sigma^2} = \sum_{j=1}^{d_i} \left(\frac{\langle X, e_j^i \rangle}{\sigma} \right)^2 \sim \chi_2(d_i),$$

où l'on a utilisé le fait qu'une somme de k v.a. i.i.d. de loi $\mathcal{N}(0,1)$ suit la loi du χ_2 à k degrés de liberté. ∎

En Statistique, le Théorème de Cochran donne des informations sur les estimateurs de la moyenne ou de la variance dans un échantillon gaussien. Pour un échantillon $(X_1, \dots, X_n)$ de v.a. i.i.d. de loi $\mathcal{N}(m, \sigma^2)$ avec $m \in \mathbb{R}$ et $\sigma^2 > 0$, on note

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{et} \quad S_n^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2,$$

les moyenne et variance empiriques, respectivement. Alors on peut montrer grâce au Théorème de Cochran que les v.a. $\overline{X}_n$ et S_n^2 sont indépendantes avec de surcroît

$$\frac{(n-1)}{\sigma^2} S_n^2 \sim \chi_2(n-1).$$

En effet, on peut supposer sans perte de généralité que $(m, \sigma^2) = (0, 1)$. Considérons le vecteur gaussien $X = (X_1, \dots, X_n)^T$ et L le sous-espace vectoriel de $\mathbb{R}^n$ engendré par le vecteur $e := (1, \dots, 1)^T$, qui est donc de dimension 1. Alors

$$P_L X \;=\; \frac{\langle X, e \rangle}{\|e\|^2}\, e \;=\; \overline{X}_n\, e.$$

De plus,

$$(I - P_L)X = (X_1 - \overline{X}_n, \dots, X_n - \overline{X}_n)^T.$$

Comme $(I - P_L)$ est la matrice de la projection orthogonale sur l'orthogonal de L, qui est de dimension $n-1$, on déduit du Théorème de Cochran que les vecteurs $P_L X$ et $(I - P_L)X$ sont indépendants, et en particulier l'indépendance de $\overline{X}_n$ et S_n^2. Enfin, toujours d'après le Théorème de Cochran, on a

$$(n-1)\, S_n^2 \;=\; \|(I - P_L)X\|^2 \;\sim\; \chi_2(n-1).$$

Ces propriétés permettent de démontrer que la v.a. centrée et correctement renormalisée $(\sqrt{n}\,(\overline{X}_n - m))/S_n$ suit pour tout entier $n \geq 2$ la loi dite de Student à $n-1$ degrés de liberté, menant ainsi à des intervalles de confiance utilisables en pratique lorsque la variance σ^2 des v.a. X_i est inconnue. Néanmoins, nous allons nous arrêter ici car ces considérations statistiques ô combien intéressantes feront l'objet d'un cours dédié au second semestre.

Chapitre 4

Convergence de variables aléatoires

En théorie des probabilités, il existe différents modes de convergence menant à des résultats fondamentaux tels que la Loi des Grands Nombres ou le Théorème Central Limite. C'est ce que nous nous proposons d'étudier dans ce chapitre. Nous y rappelons les différents modes de convergence et les liens entre eux. Les résultats établis le sont pour des v.a. réelles, mais ils restent vrais (dans leur grande majorité) dans le cadre multidimensionnel.

4.1 Convergence presque sûre

Les v.a. étant des fonctions mesurables définies sur l'espace de probabilités $(\Omega, \mathcal{A}, \mathbb{P})$, la convergence de v.a. correspond à la convergence de fonctions. Pour les fonctions, la convergence la plus faible est la convergence simple qui s'énonce pour des v.a. de la façon suivante : pour tout $\omega \in \Omega$, $\lim_{n\to+\infty} X_n(\omega) = X(\omega)$. En théorie des probabilités (et plus généralement en théorie de la mesure), il est trop restrictif de demander la convergence pour tous les $\omega \in \Omega$. A la place, on considère la convergence presque sûre, que nous avons brièvement évoquée dans le chapitre 1.

Définition 4.1.1. *Une suite de v.a. $(X_n)_{n\in\mathbb{N}}$ converge presque sûrement vers une v.a. X si la convergence est vraie avec probabilité 1 :*

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to+\infty} X_n(\omega) = X(\omega)\right\}\right) = 1,$$

c'est-à-dire pour tout ω en dehors d'un ensemble négligeable (on dit "pour presque tout $\omega \in \Omega$"), $\lim_{n\to+\infty} X_n(\omega) = X(\omega)$.
On note cette convergence $X_n \xrightarrow[n\to+\infty]{p.s.} X$.

Notons que l'ensemble $A := \{\lim_{n\to+\infty} X_n = X\}$ est bien un évènement appartenant à la tribu $\mathcal{A}$ car il s'écrit

$$A = \bigcap_{k\in\mathbb{N}^*} \bigcup_{N\in\mathbb{N}} \bigcap_{n\geq N} \left\{|X_n - X| < \frac{1}{k}\right\},$$

qui ne fait intervenir que des opérations ensemblistes dénombrables.

On montre aisément que la limite p.s. est unique au sens où si une suite de v.a. $(X_n)_{n\in\mathbb{N}}$ converge p.s. vers deux limites X et Y, alors nécessairement $X \overset{p.s.}{=} Y$.

En pratique la convergence p.s. est le mode de convergence le plus difficile à établir, les autres modes de convergence que l'on verra plus tard, plus faibles en général, donnant lieu à des critères simples d'utilisation qui leur sont équivalents. Néanmoins il existe quelques résultats reliés à la convergence p.s. comme les deux suivants. La démonstration du premier étant une application directe du Lemme de Borel-Cantelli que nous n'avons pas introduit (car il fait appel aux notionx désagréables de limites supérieure et inférieure d'évènements), nous l'admettrons.

Proposition 4.1.2. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a.*

1. *Si pour tout $\varepsilon > 0$, $\sum_{n\in\mathbb{N}} \mathbb{P}(|X_n - X| \geq \varepsilon) < +\infty$, alors $X_n \xrightarrow[n\to+\infty]{p.s.} X$.*
2. *On suppose les v.a. $(X_n)_{n\in\mathbb{N}}$ indépendantes. Alors $X_n \xrightarrow[n\to+\infty]{p.s.} 0$ si et seulement si pour tout $\varepsilon > 0$, $\sum_{n\in\mathbb{N}} \mathbb{P}(|X_n| \geq \varepsilon) < +\infty$.*

Le second résultat permet d'affirmer que la convergence p.s. est stable par composition avec une fonction continue. La démonstration est immédiate.

Proposition 4.1.3. *Si $X_n \xrightarrow[n\to+\infty]{p.s.} X$ et $f : \mathbb{R} \to \mathbb{R}$ est une fonction continue, alors $f(X_n) \xrightarrow[n\to+\infty]{p.s.} f(X)$.*

4.2 Convergence en probabilité

Attaquons à présent un deuxième mode de convergence que l'on a déjà abordé dans le passé, la convergence en probabilité.

Définition 4.2.1. *Une suite de v.a. $(X_n)_{n\in\mathbb{N}}$ converge en probabilité vers une v.a. X si pour tout $\varepsilon > 0$,*

$$\lim_{n\to+\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0.$$

On note cette convergence $X_n \xrightarrow[n\to+\infty]{\mathbb{P}} X$.

Contrairement à la convergence p.s., l'unicité de la limite en probabilité n'est pas immédiate. Pour la démontrer, considérons une suite de v.a. $(X_n)_{n\in\mathbb{N}}$ convergeant en probabilité vers deux limites X et Y. Alors pour tout $\varepsilon > 0$,

$$\{|X - Y| > \varepsilon\} \subset \{|X_n - X| > \varepsilon/2\} \cup \{|X_n - Y| > \varepsilon/2\},$$

et en passant à la probabilité, il vient

$$\mathbb{P}(|X - Y| > \varepsilon) \leq \mathbb{P}(|X_n - X| > \varepsilon/2) + \mathbb{P}(|X_n - Y| > \varepsilon/2),$$

ces deux probabilités tendant vers 0 lorsque n tend vers l'infini. Ainsi on a pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(|X - Y| > \varepsilon\right) = 0.$$

Enfin on conclut en observant que

$$\mathbb{P}\left(\{X \neq Y\}\right) = \mathbb{P}\left(\bigcup_{k\geq 1}\left\{|X-Y| > \frac{1}{k}\right\}\right) \leq \sum_{k\geq 1}\mathbb{P}\left(|X-Y| > \frac{1}{k}\right) = 0,$$

c'est-à-dire que $X \overset{p.s.}{=} Y$.

La Proposition 4.1.2 ne nous permet pas directement de comparer les convergences p.s. et en probabilité. Néanmoins, le résultat suivant permet d'en dégager une hiérarchie.

Proposition 4.2.2. *La convergence p.s. entraîne la convergence en probabilité.*

Démonstration. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. convergeant p.s. vers une v.a. X. Soit $\varepsilon > 0$ fixé. Alors pour presque tout $\omega \in \Omega$, il existe $n_0 \in \mathbb{N}$ tel que pour tout $n \geq n_0$,

$$|X_n - X| < \varepsilon.$$

Posons $Y_n = 1_{\{|X_n - X| \geq \varepsilon\}}$ qui sont des v.a. positives et majorées par 1. Alors on a $Y_n(\omega) = 0$ dès que $n \geq n_0$, donc la suite $(Y_n)_{n\in\mathbb{N}}$ converge p.s. vers 0. Ainsi, par le théorème de convergence dominée (voir le Théorème 4.3.1 à venir),

$$\lim_{n\to+\infty}\mathbb{P}\left(|X_n - X| \geq \varepsilon\right) = \lim_{n\to+\infty}\mathbb{E}[Y_n] = \mathbb{E}\left[\lim_{n\to+\infty} Y_n\right] = \mathbb{E}\left[0\right] = 0.$$

■

Attention, la réciproque est fausse, comme on peut le voir avec des v.a. indépendantes $X_n \sim \mathcal{B}(1/n)$ suivant la loi de Bernoulli de paramètre $1/n$, $n \in \mathbb{N}^*$ (cf. TD).

Une autre façon de démontrer la Proposition 4.2.2 sans utiliser le théorème de convergence dominée est la suivante. Tout d'abord si $X_n \xrightarrow[n\to+\infty]{p.s.} X$, alors on a

$$\mathbb{P}\left(\bigcap_{k\in\mathbb{N}^*}\bigcup_{N\in\mathbb{N}}\bigcap_{n\geq N}\left\{|X_n - X| < \frac{1}{k}\right\}\right) = 1.$$

Or pour toute famille d'évènements $(A_k)_{k\in\mathbb{N}^*}$, on a $\mathbb{P}(\bigcap_{k\in\mathbb{N}^*} A_k) = 1$ si et seulement si $\mathbb{P}(A_k) = 1$ pour tout $k \in \mathbb{N}^*$ (exercice), donc l'égalité précédente est équivalente à

$$\mathbb{P}\left(\bigcup_{N\in\mathbb{N}}\bigcap_{n\geq N}\left\{|X_n - X| < \frac{1}{k}\right\}\right) = 1,$$

pour tout $k \in \mathbb{N}^*$. En d'autres termes, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\bigcup_{N\in\mathbb{N}}\bigcap_{n\geq N}\left\{|X_n - X| < \varepsilon\right\}\right) = 1.$$

En passant au complémentaire, on a pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\bigcap_{N\in\mathbb{N}}\bigcup_{n\geq N}\{|X_n - X| \geq \varepsilon\}\right) = 0.$$

Si l'on note $A_N := \bigcup_{n\geq N}\{|X_n - X| \geq \varepsilon\}$, alors la famille dénombrable $(A_N)_{N\in\mathbb{N}}$ est décroissante pour l'inclusion, auquel cas pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\bigcap_{N\in\mathbb{N}} A_N\right) = \lim_{N\to+\infty}\mathbb{P}(A_N).$$

Ainsi, la convergence p.s. est équivalente à

$$\lim_{N\to+\infty}\mathbb{P}\left(\sup_{n\geq N}|X_n - X| \geq \varepsilon\right) = 0,$$

c'est-à-dire à la convergence en probabilité de la suite de v.a. supremum $(\sup_{n\geq N}|X_n - X|)_{N\in\mathbb{N}}$ vers 0. Enfin, en observant que pour tout $\varepsilon > 0$ et tout $N \in \mathbb{N}$, on a l'inégalité

$$\mathbb{P}\left(|X_N - X| \geq \varepsilon\right) \leq \mathbb{P}\left(\sup_{n\geq N}|X_n - X| \geq \varepsilon\right),$$

on en déduit en passant à la limite $N \to +\infty$ que la convergence p.s. entraîne la convergence en probabilité.

Le résultat suivant, similaire à celui pour la convergence p.s., établit des propriétés de stabilité de la convergence en probabilité.

Proposition 4.2.3. *On a les résultats de stabilité suivants pour la convergence en probabilité :*

1. *Si* $X_n \xrightarrow[n\to+\infty]{\mathbb{P}} X$ *et* $Y_n \xrightarrow[n\to+\infty]{\mathbb{P}} Y$ *alors pour tous* $\alpha, \beta \in \mathbb{R}$,

$$\alpha X_n + \beta Y_n \xrightarrow[n\to+\infty]{\mathbb{P}} \alpha X + \beta Y.$$

2. *Si* $X_n \xrightarrow[n\to+\infty]{\mathbb{P}} X$ *et* $f : \mathbb{R} \to \mathbb{R}$ *est une fonction continue, alors* $f(X_n) \xrightarrow[n\to+\infty]{\mathbb{P}} f(X)$.

Démonstration. La première propriété est immédiate. En effet, en supposant sans perte de généralité que $\alpha, \beta \neq 0$, on a pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(|\alpha X_n + \beta Y_n - (\alpha X + \beta Y)| \geq \varepsilon\right) \leq \mathbb{P}\left(|X_n - X| \geq \varepsilon/2|\alpha|\right) + \mathbb{P}\left(|Y_n - Y| \geq \varepsilon/2|\beta|\right),$$

et le membre de droite tend vers 0 lorsque $n \to +\infty$ car les suites $(X_n)_{n\in\mathbb{N}}$ et $(Y_n)_{n\in\mathbb{N}}$ convergent en probabilité vers X et Y, respectivement.

Pour la deuxième propriété, supposons tout d'abord f uniformément continue. Alors pour tout $\varepsilon > 0$ fixé, il existe $\eta > 0$ tel quel que

$$|X_n - X| < \eta \quad \Longrightarrow \quad |f(X_n) - f(X)| < \varepsilon.$$

Ainsi, on obtient l'inclusion $\{|f(X_n) - f(X)| \geq \varepsilon\} \subset \{|X_n - X| \geq \eta\}$ et donc en passant à la probabilité,

$$\mathbb{P}\left(|f(X_n) - f(X)| \geq \varepsilon\right) \leq \mathbb{P}\left(|X_n - X| \geq \eta\right).$$

D'où la conclusion en faisant tendre n vers $+\infty$, la suite $(X_n)_{n\in\mathbb{N}}$ convergeant en probabilité vers X.
La démonstration dans le cas général consiste à localiser sur un compact la fonction f puis utiliser le Théorème de Heine, auquel cas la fonction f restreinte à ce compact est uniformément continue. Ce passage étant quelque peu technique, on l'admettra. ∎

4.3 Convergence dans l'espace L^p

Ce troisième mode de convergence est très souvent utilisé en probabilités. Avant de rentrer dans le vif du sujet, rappelons les résultats classiques issus de la théorie de la mesure, les théorèmes de convergences monotone et dominée.

Théorème 4.3.1. *L'espérance vérifie les propriétés suivantes.*

1. *(Convergence monotone) Soit $(X_n)_{n\in\mathbb{N}}$ une suite croissante de v.a. positives convergeant p.s. vers une v.a. X. Alors*

$$\lim_{n\to+\infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

2. *(Convergence dominée) Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. convergeant p.s. vers une v.a. X et telles que l'hypothèse de domination suivante soit satisfaite : il existe $V \in L^1$ telle que pour tout $n \in \mathbb{N}$, $|X_n| \leq V$. Alors*

$$\lim_{n\to+\infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

À présent, introduisons la notion de convergence dans l'espace L^p, $p \in [1, +\infty[$. Dans la suite du chapitre, le paramètre p appartiendra toujours à l'intervalle $[1, +\infty[$, le cas $p = +\infty$ étant exclu pour simplifier la présentation.

Définition 4.3.2. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. dont chaque élément appartient à l'espace L^p et soit $X \in L^p$. On dit que la suite $(X_n)_{n\in\mathbb{N}}$ converge vers X dans l'espace L^p si*

$$\lim_{n\to+\infty} \mathbb{E}\left[|X_n - X|^p\right] = 0,$$

ou encore en terme de norme,

$$\lim_{n\to+\infty} \|X_n - X\|_{L^p} = 0.$$

On note cette convergence $X_n \xrightarrow[n\to+\infty]{L^p} X$.

Lorsque l'on fait varier le paramètre p, nous sommes en mesure de hiérarchiser la convergence. Il s'agit du même argument reposant sur l'inégalité de Hölder qui permet de montrer que $L^q \subset L^p$ dès que $1 \leq p < q < +\infty$.

Proposition 4.3.3. *Soient* $1 \leq p \leq q < +\infty$. *Si* $X_n \xrightarrow[n\to+\infty]{L^q} X$ *alors* $X_n \xrightarrow[n\to+\infty]{L^p} X$.

Démonstration. Posons $\alpha := q/p > 1$ et β son exposant conjugué de sorte que $\alpha^{-1}+\beta^{-1} = 1$, à savoir $\beta := \alpha/(\alpha-1) = q/(q-p) > 1$. D'après l'inégalité de Hölder appliquée aux v.a. $|Y|^p$ et 1, où Y désigne n'importe quelle v.a. appartenant à L^q, on a

$$\mathbb{E}\left[|Y|^p\right] \leq \mathbb{E}\left[|Y|^{p\alpha}\right]^{1/\alpha} \mathbb{E}\left[1^\beta\right]^{1/\beta} = \mathbb{E}\left[|Y|^q\right]^{p/q},$$

c'est-à-dire en terme de normes,

$$\|Y\|_{L^p}^p \leq \|Y\|_{L^q}^p.$$

Autrement dit, l'application $\|\cdot\|_{L^r}$ est croissante en $r \in [1, +\infty[$. Appliquée à la v.a. $Y = X_n - X$, on en déduit le résultat désiré en passant à la limite $n \to +\infty$. ∎

En particulier, la convergence dans l'espace L^p entraîne toujours la convergence dans L^1.

Tentons de relier ce mode de convergence avec les deux autres déjà étudiés, ceux de la convergence p.s. et de la convergence en probabilité. Tout d'abord il n'y a pas de hiérarchie particulière entre la convergence dans l'espace L^p et la convergence p.s. Néanmoins en ajoutant une hypothèse de domination appropriée quelque peu similaire à celle apparaissnat dans le théorème de convergence dominée, la convergence p.s. entraîne la convergence dans l'espace L^p.

Proposition 4.3.4. *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de v.a. convergeant p.s. vers une v.a.* X *et telles que l'hypothèse de domination suivante soit satisfaite : il existe* $V \in L^1$ *telle que pour tout* $n \in \mathbb{N}$, $|X_n|^p \leq V$. *Alors* $X_n \xrightarrow[n\to+\infty]{L^p} X$.

Démonstration. La démonstration repose sur l'utilisation du théorème de convergence dominée. Tout d'abord, on a évidemment la convergence p.s. vers 0 de la suite $(|X_n - X|^p)_{n\in\mathbb{N}}$. Par ailleurs, l'hypothèse de domination indique que $X_n \in L^p$, mais c'est aussi le cas pour la v.a. X en passant à la limite dans l'hypothèse de domination (auquel cas $|X|^p \leq V$), d'où le fait que $|X_n - X| \in L^p$ pour tout $n \in \mathbb{N}$, l'espace L^p étant un espace vectoriel. De plus, en utilisant l'inégalité triviale $|a-b|^p \leq 2^{p-1}(|a|^p + |b|^p)$, $a, b \in \mathbb{R}$, on a

$$|X_n - X|^p \leq 2^{p-1}\left(|X_n|^p + |X|^p\right) \leq 2^p V,$$

i.e., chaque élément de la suite $(|X_n - X|^p)_{n\in\mathbb{N}}$ satisfait une hypothèse de domination. Ainsi, on obtient par le théorème de convergence dominée appliqué à cette suite que

$$\lim_{n\to+\infty} \mathbb{E}\left[|X_n - X|^p\right] = \mathbb{E}\left[\lim_{n\to+\infty} |X_n - X|^p\right] = \mathbb{E}\left[0\right] = 0,$$

c'est-à-dire $X_n \xrightarrow[n\to+\infty]{L^p} X$. ∎

À présent, intéressons-nous à la convergence en probabilité. On va montrer que la convergence dans l'espace L^p entraîne la convergence en probabilité.

Proposition 4.3.5. *La convergence dans l'espace L^p entraîne la convergence en probabilité.*

Démonstration. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. dont chaque élément appartient à l'espace L^p et soit $X \in L^p$. Soit $\varepsilon > 0$. Alors par l'inégalité de Markov, on a

$$\mathbb{P}\left(|X_n - X| \geq \varepsilon\right) = \mathbb{P}\left(|X_n - X|^p \geq \varepsilon^p\right) \leq \frac{\mathbb{E}\left[|X_n - X|^p\right]}{\varepsilon^p},$$

et le résultat désiré en découle en faisant tendre n vers l'infini. ∎

Attention, la réciproque est fausse, comme on peut le voir avec la suite de v.a. $(X_n)_{n\in\mathbb{N}^*}$ dont chaque X_n ne prend que deux valeurs : a_n avec probabilité $1/n$ et 0 avec probabilité restante, où $(a_n)_{n\in\mathbb{N}^*}$ est une suite de nombres strictement positifs convenablement choisis. On étudiera ce cas en TD. En fait, il faut ajouter une hypothèse supplémentaire pour obtenir l'équivalence entre ces deux modes de convergence : l'uniforme intégrabilité, donnant lieu ci-dessous au Théorème de Vitali.

Définition 4.3.6. *Une suite de v.a. $(X_n)_{n\in\mathbb{N}}$ est uniformément intégrable si*

$$\lim_{a\to+\infty} \sup_{n\in\mathbb{N}} \mathbb{E}\left[|X_n|\, 1_{\{|X_n|>a\}}\right] = 0.$$

Bien qu'en apparence difficile à manipuler, l'uniforme intégrabilité est une notion importante car elle est plus faible que l'hypothèse de domination apparaissant dans la Proposition 4.3.4 et permet par conséquent de traiter plus de situations intéressantes. Pour le voir, fixons pour simplifier $p = 1$ et supposons donc que $|X_n| \leq V \in L^1$ pour tout $n \in \mathbb{N}$, auquel cas on a l'inégalité suivante : pour tout $a > 0$ et tout $n \in \mathbb{N}$,

$$|X_n|\, 1_{\{|X_n|>a\}} \leq V\, 1_{\{V>a\}},$$

donc en passant à l'espérance puis au supremum sur $n \in \mathbb{N}$, il vient

$$\sup_{n\in\mathbb{N}} \mathbb{E}\left[|X_n|\, 1_{\{|X_n|>a\}}\right] \leq \mathbb{E}\left[V\, 1_{\{V>a\}}\right].$$

Enfin, le théorème de convergence dominée nous permet d'affirmer que le membre de droite tend vers 0 lorsque $a \to +\infty$, d'où l'uniforme intégrabilité de la suite $(X_n)_{n\in\mathbb{N}}$.

Par ailleurs, l'uniforme intégrabilité entraîne la bornitude dans l'espace L^1 au sens suivant.

Proposition 4.3.7. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite uniformément intégrable. Alors on a*

$$\sup_{n\in\mathbb{N}} \mathbb{E}\left[|X_n|\right] < +\infty.$$

Démonstration. L'uniforme intégrabilité de la suite $(X_n)_{n\in\mathbb{N}}$ se traduit de la manière suivante : pour tout $\varepsilon > 0$, il existe $A > 0$ tel que pour tout $a \geq A$ et tout $n \in \mathbb{N}$,

$$\mathbb{E}\left[|X_n|\, 1_{\{|X_n|>a\}}\right] < \varepsilon.$$

Ainsi, on obtient

$$\mathbb{E}\left[|X_n|\right] \leq \varepsilon + \mathbb{E}\left[|X_n|\, 1_{\{|X_n|\leq a\}}\right] \leq \varepsilon + a.$$

Les paramètres ε et a étant indépendants de n, le passage au supremum sur $n \in \mathbb{N}$ achève la preuve. ∎

Attention, en général le supremum sur n de l'espérance ne coïncide pas avec l'espérance du supremum (cette dernière majore la première), c'est-à-dire que supremum sur n et espérance ne commutent pas !

Nous sommes maintenant en mesure d'énoncer le Théorème de Vitali reliant la convergence en probabilité à la convergence dans l'espace L^p. La démonstration, reposant sur la proposition précédente, sur la notion d'équi-continuité et utilisant une propriété que nous n'avons pas vue, à savoir que la convergence en probabilité entraîne la convergence p.s. à une sous-suite près, est admise.

Théorème 4.3.8 (Vitali)**.** *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. dont chaque élément appartient à l'espace L^p. Il y a équivalence entre les deux assertions suivantes :*

1. *La suite $(X_n)_{n\in\mathbb{N}}$ converge dans l'espace L^p.*
2. *La suite $(X_n)_{n\in\mathbb{N}}$ converge en probabilité et $(X_n^p)_{n\in\mathbb{N}}$ est uniformément intégrable.*

4.4 Convergence en loi

Terminons ce chapitre en introduisant le mode de convergence sans doute le plus utilisé en pratique, celui de la convergence en loi. Dans la suite, étant donnée une v.a. X, on note F_X la fonction de répartition associée,

$$F_X(t) = \mathbb{P}(X \leq t), \quad t \in \mathbb{R}.$$

On rappelle que F_X est une fonction croissante, continue à droite et ayant une limite à gauche et telle que

$$\lim_{t\to-\infty} F_X(t) = 0, \quad \lim_{t\to+\infty} F_X(t) = 1.$$

De plus elle admet au plus un nombre dénombrable de points de discontinuité. Enfin, elle caractérise la loi au sens où Y est une autre v.a. telle que $F_X = F_Y$, alors X et Y ont même loi.

Définition 4.4.1. *On dit qu'une suite $(X_n)_{n\in\mathbb{N}}$ de v.a. converge en loi vers une v.a. X si la suite $(F_{X_n})_{n\in\mathbb{N}}$ de fonctions de répartition converge simplement vers F_X en tout point où F_X est continue (c'est-à-dire en les points $t \in \mathbb{R}$ pour lesquels $\mathbb{P}(X = t) = 0$).*
On note cette convergence $X_n \xrightarrow[n\to+\infty]{loi} X$.

Notons que si X est une v.a. admettant une densité, sa fonction de répartition est continue et la convergence en loi correspond à la convergence simple sur tout $\mathbb{R}$ des fonctions de répartition.

Lorsqu'en pratique l'expression de la fonction de répartition n'est pas connue, il peut être intéressant d'utiliser les deux formulations équivalentes de la convergence en loi. La première, que l'on a déjà vu dans le passé, fait appel à la notion de fonction caractéristique. On rappelle que la fonction caractéristique d'une v.a. X est définie par

$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right], \quad t \in \mathbb{R}.$$

La démonstration du résultat suivant reposant sur l'analyse de Fourier qui ne sera abordée qu'au second semestre, nous l'admettrons.

Théorème 4.4.2. *La convergence $X_n \xrightarrow[n\to+\infty]{loi} X$ est équivalente à la convergence simple des fonctions caractéristiques :*

$$\lim_{n\to+\infty} \varphi_{X_n}(t) = \varphi_X(t), \quad t \in \mathbb{R}.$$

La seconde formulation équivalente, faisant intervenir les fonctions continues et bornées, est la suivante. La démonstration utilisant des arguments fins de théorie de la mesure, nous l'admettrons également.

Théorème 4.4.3. *La convergence $X_n \xrightarrow[n\to+\infty]{loi} X$ est équivalente à l'assertion suivante : pour toute fonction $f : \mathbb{R} \to \mathbb{R}$ continue et bornée,*

$$\lim_{n\to+\infty} \mathbb{E}\left[f(X_n)\right] = \mathbb{E}\left[f(X)\right].$$

En particulier, ce résultat permet d'établir la stabilité de la convergence en loi par composition avec une fonction continue. La démonstration, immédiate, est laissée en exercice.

Proposition 4.4.4. *Si $X_n \xrightarrow[n\to+\infty]{loi} X$ et $g : \mathbb{R} \to \mathbb{R}$ est une fonction continue, alors $g(X_n) \xrightarrow[n\to+\infty]{loi} g(X)$.*

Contrairement à la convergence p.s. voire à la convergence en probabilité, la convergence en loi ne vérifie pas de bonnes propriétés arithmétiques : par exemple la convergence en loi d'une suite de v.a. $(X_n)_{n\in\mathbb{N}}$ vers une v.a. X n'assure pas forcément que $(X_n - X)_{n\in\mathbb{N}}$ converge en loi vers 0. Pour le voir, on se donne une suite de nombres strictement positifs

$(\sigma_n^2)_{n\in\mathbb{N}}$ qui tend vers 1 lorsque $n \to +\infty$. Considérons alors une suite $(X_n)_{n\in\mathbb{N}}$ de v.a. de loi $\mathcal{N}(0,\sigma_n^2)$ indépendante d'une v.a. $X \sim \mathcal{N}(0,1)$. On verra en TD que $X_n \xrightarrow[n\to+\infty]{loi} X$ sans que $X_n - X \xrightarrow[n\to+\infty]{loi} 0$. Cependant, dans certains cas, le Théorème de Slutsky nous fournit un cadre pour lequel il existe certains résultats positifs. Avant de l'énoncer, commençons par une version plus faible de ce théorème, mais qui contient l'essentiel de la difficulté et à laquelle on va se ramener.

Proposition 4.4.5. *Si* $X_n \xrightarrow[n\to+\infty]{loi} X$ *et* $Y_n \xrightarrow[n\to+\infty]{loi} 0$ *alors* $X_n + Y_n \xrightarrow[n\to+\infty]{loi} X$.

Démonstration. La démonstration est un tantinet technique mais néanmoins accessible. Soit $t \in \mathbb{R}$ un point de continuité de F_X et soit $\alpha > 0$. Pour tout $n \in \mathbb{N}$, on a l'inclusion

$$\{X_n + Y_n \leq t\} \subset \{X_n \leq t + \alpha\} \cup \{Y_n \leq -\alpha\},$$

auquel cas en passant à la probabilité, il vient

$$F_{X_n+Y_n}(t) \leq F_{X_n}(t+\alpha) + F_{Y_n}(-\alpha).$$

De la même manière, on a

$$\{X_n \leq t - \alpha\} \subset \{X_n + Y_n \leq t\} \cup \{-Y_n \leq -\alpha\},$$

et donc

$$F_{X_n}(t-\alpha) \leq F_{X_n+Y_n}(t) + F_{-Y_n}(-\alpha),$$

c'est-à-dire

$$F_{X_n+Y_n}(t) \geq F_{X_n}(t-\alpha) - 1 + F_{Y_n}(\alpha).$$

Ainsi, on a l'encadrement

$$F_{X_n}(t-\alpha) - 1 + F_{Y_n}(\alpha) \leq F_{X_n+Y_n}(t) \leq F_{X_n}(t+\alpha) + F_{Y_n}(-\alpha).$$

À présent fixons $\varepsilon > 0$, par continuité de F_X au point t, il existe $\alpha > 0$ tel que

$$F_X(t+\alpha) \leq F_X(t) + \frac{\varepsilon}{3}.$$

On peut s'arranger pour choisir α tel que F_X soit continue au point $t+\alpha$ (une fonction de répartition ayant au plus une infinité dénombrable de points de discontinuité, ses points de continuité sont partout denses). Ainsi, la convergence en loi de la suite $(X_n)_{n\in\mathbb{N}}$ vers X indique que

$$\lim_{n\to+\infty} F_{X_n}(t+\alpha) = F_X(t+\alpha).$$

Il existe donc $n_0 \in \mathbb{N}$ tel que pour tout $n \geq n_0$,

$$F_{X_n}(t+\alpha) \leq F_X(t+\alpha) + \frac{\varepsilon}{3}.$$

Par ailleurs, la convergence en loi de la suite $(Y_n)_{n\in\mathbb{N}}$ vers 0 entraîne que

$$\lim_{n\to+\infty} F_{Y_n}(-\alpha) = \mathbb{P}(0 \leq -\alpha) = 0,$$

la fonction de répartition de la v.a. nulle étant continue en $-\alpha$. Il existe donc également $n_1 \in \mathbb{N}$ tel que pour tout $n \geq n_1$,

$$F_{Y_n}(-\alpha) \leq \frac{\varepsilon}{3}.$$

Ainsi, en recollant les morceaux, pour tout $n \geq N_0 := \max\{n_0, n_1\}$, on obtient

$$F_{X_n+Y_n}(t) \leq F_{X_n}(t+\alpha) + F_{Y_n}(-\alpha) \leq F_X(t) + \varepsilon.$$

On montre de même qu'il existe $N_1 \in \mathbb{N}$ tel que pour tout $n \geq N_1$,

$$F_{X_n+Y_n}(t) \geq F_{X_n}(t-\alpha) - 1 + F_{Y_n}(\alpha) \geq F_X(t) - \varepsilon.$$

Enfin, cela donne pour tout $n \geq \max\{N_0, N_1\}$ en tout point de continuité de F_X :

$$F_X(t) - \varepsilon \leq F_{X_n+Y_n}(t) \leq F_X(t) + \varepsilon,$$

c'est-à-dire

$$|F_{X_n+Y_n}(t) - F_X(t)| \leq \varepsilon.$$

Ceci achève la démonstration de la proposition. ∎

Il est temps maintenant d'énoncer le Théorème de Slutsky. Pour ce faire, il faut adapter les définitions des convergences en probabilité et en loi au cadre multidimensionnel. Pour la convergence en probabilité, on va remplacer la valeur absolue par n'importe quelle norme en dimension supérieure, toutes les normes étant équivalentes en dimension finie. Pour ce qui est de la convergence en loi, on utilisera comme définition la version multidimentionnelle de la caractérisation apparaissant dans le Théorème 4.4.3.

Théorème 4.4.6 (Slutsky). *Soit $X_n \xrightarrow[n\to+\infty]{\star} X$ et $Y_n \xrightarrow[n\to+\infty]{loi} c$ où c est une constante réelle. Alors on a la convergence*

$$(X_n, Y_n) \xrightarrow[n\to+\infty]{\star} (X, c),$$

la convergence $\xrightarrow{\star}$ désignant la convergence en loi ou en probabilité selon le même mode de convergence que la suite $(X_n)_{n\in\mathbb{N}}$.

Démonstration. Établissons tout d'abord le résultat pour la convergence en loi. Pour ce faire, on se donne une fonction continue et bornée $f : \mathbb{R}^2 \to \mathbb{R}$ et l'on veut montrer que

$$\lim_{n\to+\infty} \mathbb{E}\left[f(X_n, Y_n)\right] = \mathbb{E}\left[f(X, c)\right].$$

La fonction $g := f(\cdot, c)$ de la première variable étant continue et bornée, on a par hypothèse que

$$\lim_{n\to+\infty} \mathbb{E}\left[g(X_n)\right] = \mathbb{E}\left[g(X)\right],$$

c'est-à-dire

$$\lim_{n\to+\infty} \mathbb{E}\left[f(X_n, c)\right] = \mathbb{E}\left[f(X, c)\right],$$

ou encore que $(X_n, c) \xrightarrow[n\to+\infty]{loi} (X, c)$. Par ailleurs, on a

$$(X_n, Y_n) - (X_n, c) = (0, Y_n - c) \xrightarrow[n\to+\infty]{loi} (0, 0),$$

car $Y_n \xrightarrow[n\to+\infty]{loi} c$ entraîne trivialement que $Y_n - c \xrightarrow[n\to+\infty]{loi} 0$. Ainsi, la Proposition 4.4.5 entraîne alors que

$$(X_n, Y_n) \xrightarrow[n\to+\infty]{loi} (X, c).$$

Pour le résultat concernant la convergence en probabilité, la démonstration est encore plus rapide. En effet, si $\|\cdot\|$ désigne par exemple la norme 1 sur $\mathbb{R}^2$, à savoir $\|(x, y)\| := |x|+|y|$ pour tout $(x, y) \in \mathbb{R}^2$, alors pour tout $\varepsilon > 0$ et tout $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{P}\left(\|(X_n, Y_n) - (X, c)\| \geq \varepsilon\right) &= \mathbb{P}\left(|X_n - X| + |Y_n - c| \geq \varepsilon\right) \\ &\leq \mathbb{P}\left(|X_n - X| \geq \frac{\varepsilon}{2}\right) + \mathbb{P}\left(|Y_n - c| \geq \frac{\varepsilon}{2}\right).\end{aligned}$$

Le premier terme dans le membre de droite tend vers 0 par hypothèse, tout comme le second dès lors que la suite $(Y_n)_{n\in\mathbb{N}}$ converge en probabilité vers la constante c. C'est l'objet de la Proposition 4.4.9 à venir, énonçant que les convergences en loi et en probabilité sont en réalité équivalentes lorsque la limite est une constante. ■

Le Théorème de Slutsky est très important, notamment en Statistique. Considérons par exemple une suite $(X_n)_{n\in\mathbb{N}^*}$ de v.a. i.i.d. et de carré intégrable dont on suppose l'espérance m inconnue. On souhaite estimer ce paramètre en proposant un intervalle de confiance. Le TCL nous dit que la moyenne empirique correctement renormalisée converge en loi vers une v.a. de loi $\mathcal{N}(0, 1)$, i.e.,

$$\sqrt{n}\left(\frac{\overline{X}_n - m}{\sigma}\right) \xrightarrow[n\to+\infty]{loi} X \sim \mathcal{N}(0, 1),$$

où $\sigma^2 := \mathrm{Var}(X_1)$. Néanmoins il se peut que cette variance soit elle-aussi inconnue, auquel cas l'intervalle de confiance (asymptotique) pour l'estimation du paramètre inconnu m fait intervenir la variance inconnue. Pour remédier à ce problème, on remplace la variance par la variance empirique S_n^2 qui est un estimateur consistant de σ^2 (c'est-à-dire convergeant en probabilité vers la variance). Ainsi, en appliquant le Théorème de Slutsky ainsi que la stabilité de la convergence en loi par composition avec la fonction continue sur $\mathbb{R}\times]0, +\infty[$ définie par $f(x, y) = x/\sqrt{y}$, il vient

$$\sqrt{n}\left(\frac{\overline{X}_n - m}{S_n}\right) = f\left(\sqrt{n}(\overline{X}_n - m), S_n^2\right) \xrightarrow[n\to+\infty]{loi} f(\sigma X, \sigma^2) = X \sim \mathcal{N}(0, 1).$$

Terminons ce chapitre en évoquant les liens avec les autres modes de convergence. Commençons par la convergence p.s.

Proposition 4.4.7. *La convergence p.s. entraîne la convergence en loi.*

Démonstration. On pourrait justifier ce résultat en utilisant le fait que la convergence p.s. entraîne la convergence en probabilité, qui entraîne à son tour la convergence en loi (cf. le résultat ci-dessous). Proposons une autre démonstration toute aussi rapide. Soit $X_n \xrightarrow[n\to+\infty]{p.s.} X$ et $f : \mathbb{R} \to \mathbb{R}$ une fonction continue et bornée. Alors $f(X_n) \xrightarrow[n\to+\infty]{p.s.} f(X)$. De surcroît, comme les v.a. $f(X_n)$ sont bornées (par $\|f|_\infty$), le théorème de convergence dominée entraîne que

$$\lim_{n\to+\infty} \mathbb{E}\left[f(X_n)\right] = \mathbb{E}\left[f(X)\right],$$

donc la convergence en loi désirée. ∎

À présent, étudions le cas de la convergence en probabilité.

Proposition 4.4.8. *La convergence en probabilité implique la convergence en loi.*

Démonstration. La preuve est quelque peu similaire à celle de la Proposition 4.4.5. Soit $X_n \xrightarrow[n\to+\infty]{\mathbb{P}} X$, on veut montrer que F_{X_n} converge simplement vers F_X en tout point de continuité de cette dernière. Soit $x \in \mathbb{R}$ un tel point et soit $\varepsilon > 0$. On a facilement que

$$F_{X_n}(x) \leq F_X(x+\varepsilon) + F_{X_n - X}(-\varepsilon) \leq F_X(x+\varepsilon) + \mathbb{P}\left(|X_n - X| \geq \varepsilon\right).$$

Soit $\eta > 0$, par continuité de F_X au point x, il existe $\varepsilon_0 > 0$ tel que

$$F_X(x+\varepsilon_0) \leq F_X(x) + \frac{\eta}{2}.$$

Pour ce choix de ε_0, la suite $(X_n)_{n\in\mathbb{N}}$ convergeant en probabilité vers X, il existe $n_0 \in \mathbb{N}$ tel que pour tout $n \geq n_0$,

$$\mathbb{P}\left(|X_n - X| \geq \varepsilon_0\right) \leq \frac{\eta}{2}.$$

En rassemblant ces inégalités, il vient pour tout $n \geq n_0$,

$$F_{X_n}(x) \leq F_X(x) + \eta.$$

De surcroît, un raisonnement analogue permettant d'obtenir l'inégalité

$$F_X(x-\varepsilon) \leq F_{X_n}(x) + F_{X-X_n}(-\varepsilon),$$

donne lieu à l'inégalité

$$F_{X_n}(x) \geq F_X(x-\varepsilon) - \mathbb{P}\left(|X_n - X| \geq \varepsilon\right),$$

et pour n assez grand, on a

$$F_{X_n}(x) \geq F_X(x) - \eta.$$

Ceci achève la démonstration de la convergence en loi. ∎

Attention, la réciproque est fausse, comme on peut le voir avec des v.a. i.i.d. $X_n \sim \mathcal{B}(1/2)$ (cf. TD). En revanche lorsque la limite est une constante, alors ces deux modes de convergence sont en réalité équivalents. C'est cette propriété qui a été utilisée dans la démonstration du Théorème de Slutsky.

Proposition 4.4.9. *Si* $X_n \xrightarrow[n\to+\infty]{loi} c$*, où* c *est une constante réelle, alors* $X_n \xrightarrow[n\to+\infty]{\mathbb{P}} c$*.*

Démonstration. Soit $\varepsilon > 0$ alors

$$\mathbb{P}\left(|X_n - c| \geq \varepsilon\right) = \mathbb{E}\left[h(X_n)\right],$$

où h est la fonction indicatrice $h(x) := 1_{\{|x-c|\geq\varepsilon\}}$. Par ailleurs, soit g la fonction continue et bornée sur $\mathbb{R}$ qui vaut 1 si $|x-c| \geq \varepsilon$, 0 en c et affine entre $c-\varepsilon$ et c et entre c et $c+\varepsilon$. En faisant un dessin, on remarque que $h \leq g$. D'où

$$\mathbb{P}\left(|X_n - c| \geq \varepsilon\right) = \mathbb{E}\left[h(X_n)\right] \leq \mathbb{E}\left[g(X_n)\right].$$

Enfin, la convergence en loi vers la constante c de la suite $(X_n)_{n\in\mathbb{N}}$ nous permet d'affirmer que

$$\lim_{n\to+\infty} \mathbb{E}\left[g(X_n)\right] = \mathbb{E}\left[g(c)\right] = g(c) = 0,$$

d'où la convergence en probabilité vers la constante c de la suite $(X_n)_{n\in\mathbb{N}}$. ∎

La convergence dans l'espace L^p implique la convergence en loi (puisqu'elle implique la convergence en probabilité qui implique celle en loi). La réciproque est fausse comme on peut le voir avec le même contre-exemple suivant la Proposition 4.3.5.

Résumons les liens entre ces différents modes de convergence. Le schéma général des implications entre convergences est le suivant :

Convergence p.s.

$\searrow$

Convergence en probabilité $\Longrightarrow$ Convergence en loi.

$\nearrow$

Convergence dans l'espace L^p

Par ailleurs, on a aussi les résultats suivants :

1. La convergence en loi n'implique pas la convergence en probabilité, sauf lorsque la limite est une constante.

2. La convergence en probabilité n'implique pas la convergence p.s.

3. $X_n \xrightarrow[n\to+\infty]{p.s.} X$ si et seulement si $\sup_{n\geq N} |X_n - X| \xrightarrow[N\to+\infty]{\mathbb{P}} 0$.

4. La convergence en probabilité n'implique pas la convergence dans l'espace L^p, sauf si $(X_n^p)_{n\in\mathbb{N}}$ est uniformément intégrable.

5. Les convergences p.s. et dans l'espace L^p ne sont pas comparables en général. En revanche, si $X_n \xrightarrow[n\to+\infty]{p.s.} X$ et s'il existe une variable aléatoire positive et intégrable V telle que $|X_n|^p \leq V$ pour tout $n \in \mathbb{N}$, alors $X \in L^p$ et $X_n \xrightarrow[n\to+\infty]{L^p} X$.

6. Les convergences p.s., en probabilité et en loi sont stables par composition avec une fonction continue. En revanche ce n'est pas le cas pour la convergence dans l'espace L^p.

Bibliographie

[1] P. Barbe et M. Ledoux. Probabilités. EDP Sciences, 2007.

[2] J.-C. Breton. Fondements des Probabilités. Cours de L3 Mathématiques, Université de Rennes 1, 2014.

[3] J.-P. Delmas. Introduction aux probabilités. Ellipses, 2000.

[4] J.-Y. Ouvrard. Probabilités. Tomes 1 et 2. Cassini, 2008.