PROBABILITÉS: COURS DE LICENCE DE MATHÉMATIQUES APPLIQUÉES

Université PARIS 6

2000/2001

Jean BERTOIN

Table des Matières

(∗) ces parties peuvent être omises en première lecture, et ne feront pas l'objet de question aux examens.

Presentation du cours

Dans le langage commun, la notion de probabilité a deux interprétations différentes.

Interprétation fréquentielle. On fait un grand nombre d'expériences qui font intervenir le hasard et qui se déroulent dans les mêmes conditions (par exemple, on jette une pièce et on regarde si pile sort). La probabilité que l'expérience réussisse est à peu près la fréquence de succès, i.e. nombre de succès divisé par le nombre d'expériences. Ceci correspond en fait à un résultat mathématique rigoureux (et difficile), la loi des grands nombres.

Interprétation subjective. Un patient qui va se faire opérer demande à son médecin la probabilité que l'opération réussise. La réponse du médecin dépend de ce qu'il sait sur le patient et les conditions de l'opération (état de santé du malade, antécédants, qualité de l'équipe chirurgicale etc...). Il n'y a pas de sens à chercher la fréquence de succès dans un grand nombre d'expériences similaires. C'est un problème de probabilités conditionnelles.

Seul le premier de ces deux points de vue sera traité dans ce cours à l'aide d'outils mathématiques rigoureux; le second fait en partie l'objet d'un cours de maitrise. Nous présenterons d'abord les notions de base des probabilités discrètes à l'aide de la notion de famille sommable. On trouve déjà dans ce cadre simple la plupart des notions fondamentales de la théorie (variables aléatoires, moments, lois et indépendance). On introduira ensuite l'axiomatique générale de Kolmogorov, reposant sur la théorie abstraite de la mesure, en faisant le parallèle avec le cas particulier discret traîté précédemment. Puis on s'intéressera aux suites et aux séries de variables aléatoires indépendantes en introduisant diverses notions de convergence. On présentera deux des résultats les plus importants de la théorie: la loi des grands nombres et le théorème central limite. Enfin, on fera une brève introduction à quelques notions de statistique pour présenter des applications naturelles des concepts développés dans ce cours.

Quelques références:

- N. Bouleau: Probabilités de l'ingénieur. Hermann (1986).
- D. Revuz: Probalités. Hermann (1997).
- J. Jacod et P. Protter: Probability essentials. Springer (2000).

Chapitre 1

Espaces de Probabilités Finis

1.1 Equi-probabilité et dénombrement

Historiquement, tout part de la théorie des jeux de hasard (Pascal, Fermat, Bernoulli,...). On peut résoudre certains problèmes simples (en fait, souvent moins simples qu'il n'y parait) par des arguments combinatoires. Assez vite, ce cadre devient trop restrictif; mais il y a une très grande difficulté théorique à mettre sur pied des bases mathématiques à la fois suffisament générales et rigoureuses. Ceci est réalisé par Kolmogorov dans les années 30, et nécéssite la théorie abstraite de la mesure. Formellement, on considère un univers Ω, et on mesure la probabilité d'un évènement (c'est-à-dire une partie de Ω) par un réel compris entre 0 et 1. La difficulté principale est que l'univers Ω est souvent très gros et mal connu. Le formalisme de Kolmogorov a permis un développement très substantiel de la théorie des probabilités, qui intervient aujourd'hui dans de nombreux domaines même déterministes (on peut calculer des intégrales, résoudre des équations différentielles ou étudier des groupes, en utilisant des méthodes probabilistes). Cf. également la physique quantique et le principe d'incertitude d'Heisenberg.

On va voir rapidement comment on résolvait des problemes simples à la préhistoire de la théorie. En fait, la plupart des notions importantes par la suite y figurent déjà. La modernité fera son apparition dans les chapitres ultérieurs.

On considère un univers fini $\Omega = \{\omega_1, \cdots, \omega_n\}$. On appelle parfois Ω l'espace de probabilité, un élément générique ω une éventualité, et une partie Λ de Ω un événement. L'application

$$\mathbb{P} : \mathcal{P}(\Omega) \to [0,1] \qquad \mathbb{P}(\Lambda) = \frac{\sharp(\Lambda)}{\sharp(\Omega)}$$

s'appelle *l'équiprobabilité* sur Ω ($\mathcal{P}(\Omega)$ désigne l'ensemble des parties de Ω, et $\sharp(\Lambda)$ le cardinal de la partie Λ; en particulier $\sharp(\Omega) = n$).

Il est immédiat de voir que $\mathbb{P}$ vérifie les deux propriétés suivantes:

$$\begin{aligned} &\mathbb{P}(\Omega) \;=\; 1\,, \\ &\mathbb{P}(\Lambda \cup \Lambda') \;=\; \mathbb{P}(\Lambda) + \mathbb{P}(\Lambda') \quad \text{ si } \Lambda \text{ et } \Lambda' \text{ sont disjoints}\,. \end{aligned}$$

Il en découle que

$$\mathbb{P}(\Lambda^c) = 1 - \mathbb{P}(\Lambda) \quad , \quad \mathbb{P}(\emptyset) = 0$$
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Exemple On jette trois fois une pièce non truquée. On peut représenter l'univers comme l'ensemble des applications de $\{1, 2, 3\}$ (trois jets) dans $\{P, F\}$ (P= pile, F= face). Autrement dit, $\Omega = \{P, F\}^3$, et $\sharp(\Omega) = 2^3 = 8$. Il est alors très facile de voir que

$\mathbb{P}$(on sort exactement une fois P) = 3/8,

$\mathbb{P}$(on sort au moins une fois P) = 1- $\mathbb{P}$(on sort trois F) = 1-1/8 = 7/8.

Exemple On considère l'arrivée d'une course de chevaux, avec dix partants, numérotés de 1 à 10. On note l'ordre d'arrivée. On suppose que les concurrents sont de force égale. L'univers Ω est l'ensemble des injections de $\{1, \cdots, 10\}$ dans lui-même. En particulier $\sharp(\Omega) = 10!$. On a alors

$\mathbb{P}$(le numéro 10 arrive dernier) $= \sharp\left(w \in \Omega : \omega(10) = 10\right)/10! = \frac{1}{10!}$ nombre d'injections de $\{1, \cdots, 9\}$ dans lui même $= \frac{9!}{10!} = 1/10$.

Calculer maintenant la probabilité pour que le numéro 10 arrive dans les trois premiers. (réponse: 3/10).

Exemple On considère une urne contenant dix boules noires et trois rouges. On en tire simultanément deux. L'univers Ω est alors l'ensemble des parties à deux éléments d'un ensemble à treize éléments. Donc $\sharp(\Omega) = \binom{13}{2} = \frac{13!}{2!\times 11!}$. On en déduit par exemple que $\mathbb{P}$(on ne tire aucune boule rouge) $= \binom{10}{2} / \binom{13}{2} = \frac{10!\times 2!\times 11!}{2!\times 8!\times 13!} = \frac{90}{156}$

Ces exemples très simples ne doivent pas cacher la difficulté à trouver le bon modèle mathématique pour traiter un problème donné. Un modèle qui semble être naturel peut se réveller erroné. Voyons quelques raisonnements faussés.

Exemple On met de l'argent dans deux enveloppes indistinguables, l'une contient deux fois plus que l'autre (on ne précise pas les sommes). Un joueur choisit une enveloppe au hasard. Il l'ouvre et compte l'argent. On lui propose alors de laisser l'argent qu'il a eu et de prendre celui qu'il y a dans l'enveloppe encore fermée. A-t-il intéret à le faire? On pourrait faire raisonnement faux suivant: Si on note X la somme dans l'enveloppe ouverte, l'enveloppe fermée contient soit $X/2$, soit $2X$, chaque fois avec probabilité 1/2. L'espérance de son gain s'il change serait donc $\frac{1}{2}(X/2) + \frac{1}{2}(2X) = 5X/4 > X$, et on pourrait conclure que le joueur a intérèt à changer. Bien sûr, le sens commun nous dit que ça ne devrait pas être le cas, que l'espérance de son gain devrait être la même (sinon, il aurait encore intérèt à changer à nouveau, et on reviendrait à la situation de départ). Pour poser rigoureusement le problème, il faut noter x et $2x$ l'argent que contiennent les deux enveloppes. L'univers est tout simplement $\Omega = \{x, 2x\}$ (somme présente dans l'enveloppe que tire le joueur). L'espérance de son gain s'il ne change pas est $\frac{1}{2}x + \frac{1}{2}2x = \frac{3}{2}x$. L'espérance de son gain s'il change est $\frac{1}{2}2x + \frac{1}{2}x = \frac{3}{2}x$; c'est bien la même qu'avant.

Exemple Un responsable de jeux dispose de trois enveloppes de couleurs différentes. Il met de l'argent dans l'une, et du papier journal dans les deux autres. Il fait entrer un joueur et lui fait choisir une enveloppe qu'il garde fermée. Parmi les deux enveloppes restantes, il y en a toujours au moins une qui contient du papier journal.

Le responsable ouvre alors une de ces deux enveloppes dont il sait qu'elle contient du papier journal, et propose au joueur de changer l'enveloppe qu'il a en main contre celle qui reste. Le joueur y a-t-il intérèt? Au vu de l'exemple précédent, on pourrait croire hâtivement que ça ne changerait rien. En fait, ce n'est pas le cas. En effet, considérons l'événement Λ que le joueur gagne l'argent s'il décide de ne pas changer l'enveloppe, et l'événement Λ' joueur gagne l'argent s'il décide de changer l'enveloppe. Les événements Λ et Λ' sont complémentaires (ils sont disjoints et leur union, c'est l'univers tout entier). La probabilité de Λ est à l'évidence $1/3$. Celle de Λ' est donc $1 - 1/3 = 2/3$. Ainsi, le joueur a deux fois plus de chances de gagner s'il change son choix que s'il le conserve.

1.2 Probabilités générales sur un espace fini

L'équi-probabilité sur un univers fini est la structure probabiliste la plus simple qu'on peut concevoir, puisque toutes les éventualités ont la même probabilité. Plus généralement, on appelera probabilité sur $\Omega = \{\omega_1, \cdots, \omega_n\}$ toute application $\mathbb{P} : \mathcal{P}(\Omega) \to [0,1]$ qui vérifie les axiomes suivants:

$$\begin{aligned} &\mathbb{P}(\Omega) = 1\,, \\ &\mathbb{P}(\Lambda \cup \Lambda') = \mathbb{P}(\Lambda) + \mathbb{P}(\Lambda') \quad \text{si } \Lambda \text{ et } \Lambda' \text{ sont disjoints}\,. \end{aligned}$$

On déduit immédiatement qu'on a encore

$$\begin{aligned} &\mathbb{P}(\Lambda^c) = 1 - \mathbb{P}(\Lambda) \quad , \quad \mathbb{P}(\emptyset) = 0 \\ &\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \end{aligned}$$

Il est immédiat de voir que qu'une probabilité $\mathbb{P}$ est caractérisée par la donnée des probabilités élémentaires $p(\omega_i) = \mathbb{P}(\{\omega_i\})$, $i = 1, \cdots, n$. En effet on a

$$\mathbb{P}(\Lambda) = \sum_{i:\omega_i \in \Lambda} p(\omega_i)\,, \qquad \forall \Lambda \in \mathcal{P}(\Omega).$$

En particulier $\sum_{i=1}^n p(\omega_i) = 1$.

Exemple Une urne contient dix milliards de boules rouges et cinq milliards de noires, toutes indiscernables au toucher. On en tire quatre simultanément. Estimer la probabilité d'avoir tiré au moins une boule rouge? On pourrait bien sûr traiter cette question avec l'équiprobabilité; cependant ça ferait intervenir des nombres considérables. Comme on ne demande qu'une estimation, on peut raisonner de la façon suivante. Compte tenu du grand nombre de boules dans l'urne, tout se passe presque comme si on effectuait un tirage successif avec remise. La probabilité de tirer une boule noire lors d'un tirage est de $1/3$. La probabilité de tirer quatre boules noires est donc voisine de $(1/3)^4 = 1/81$. La probabilité d'avoir tiré au moins une boule rouge est donc proche de $1 - 1/81 \approx 0,98765$

Exemple On jette deux dés; on prend comme univers les valeurs possibles pour la somme des faces obtenues, c'est-à-dire $\Omega = \{2, 3, \ldots, 12\}$. Calculer $p(\omega)$ pour $\omega \in \Omega$.

Chapitre 2

Espaces de Probabilités Discrets

Dans tout ce chapitre, Ω désignera un ensemble au plus dénombrable, c'est-à-dire qu'il existe une application surjective qui envoie $\mathbb{N}$ sur Ω. On note ω l'élément générique de Ω et $\mathcal{P}(\Omega)$ l'ensemble des parties de Ω.

2.1 Préliminaires sur les familles sommables.

On rappelle que partie non vide P de $[0,\infty[$ est dite bornée s'il existe un réel M tel que $x \leq M$ pour tout $x \in P$. Un tel réel M s'appelle un majorant de P. L'ensemble des majorants de P admet alors un minimum, qu'on note $\sup P$. Autrement dit, $M = \sup P$ si et seulement si $x \leq M$ pour tout $x \in P$ et $M \leq M'$ pour tout majorant M' de P. Si P n'est pas bornée, on pose $\sup P = \infty$.

On considère une application $a : \Omega \to [0,\infty[$. Pour toute partie finie $F = \{\omega_1, \cdots, \omega_n\}$ de Ω, on pose

$$\sum_F a(\omega) \;=\; \sum_{i=1}^{n} a(\omega_i)\,.$$

Comme l'addition est commutative, cette notion ne dépend pas de la façon dont on a numéroté les éléments de F.

Définition *On pose*

$$\sum_\Omega a(\omega) \;=\; \sup\left\{\sum_F a(\omega) : F \textit{ partie finie de } \Omega\right\}.$$

On dit que la famille $(a(\omega), \omega \in \Omega)$ est sommable si $\sum_\Omega a(\omega) < \infty$.

Par exemple, si Ω est un ensemble fini, toute famille réelle positive indexée par Ω est sommable.

On observe le résultat élémentaire suivant: si $(a(\omega), \omega \in \Omega)$ et $(a'(\omega), \omega \in \Omega)$ sont deux familles positives et si pour tout $\omega \in \Omega$ on a $a(\omega) \leq a'(\omega)$, alors $\sum_\Omega a(\omega) \leq \sum_\Omega a'(\omega)$. En particulier, si la famille $(a'(\omega), \omega \in \Omega)$ est sommable, il en est de même pour la famille $(a(\omega), \omega \in \Omega)$.

Proposition *Si $\Omega = \mathbb{N}$, alors $\sum_\Omega a(\omega) \;=\; \sum_{n=0}^{\infty} a(n)$.*

Preuve: Par définition, $\sum_{n=0}^{\infty} a(n)$ est la limite croissante quand $N \to \infty$ de la suite $\sum_{n=0}^{N} a(n)$. En prenant $F_N = \{0, 1, \cdots, N\}$, on a l'inégalité $\sum_\Omega a(\omega) \geq \sum_{n=0}^{\infty} a(n)$. Pour le sens inverse, toute partie finie F de Ω, on prend $N = \sup F$, de sorte que $F \subseteq F_N$; et il en découle que $\sum_F a(\omega) \leq \sum_{i=0}^{N} a(i) \leq \sum_{i=0}^{\infty} a(i)$. ■

Considérons maintenant une application $b : \Omega \to]-\infty, \infty[$. On dit que la famille $(b(\omega), \omega \in \Omega)$ est sommable si la famille $(|b(\omega)|, \omega \in \Omega)$ l'est. Dans ce cas, on peut encore définir $\sum_\Omega b(\omega)$ de la façon suivante. Rappelons que tout réel x s'écrit comme la différence de sa partie positive x^+ (qui vaut $x^+ = x$ si $x \geq 0$ et $x^- = 0$ sinon) et de sa partie négative x^- (qui coincide avec la partie positive de $-x$). Si la famille $(|b(\omega)|, \omega \in \Omega)$ est sommable, les deux familles positives $(b(\omega)^+, \omega \in \Omega)$ et $(b(\omega)^-, \omega \in \Omega)$ le sont également puisque $b^{\pm}(\omega) \geq |b(\omega)|$, et on définit

$$\sum_\Omega b(\omega) \;=\; \sum_\Omega b(\omega)^+ - \sum_\Omega b(\omega)^- \,.$$

On a alors le résultat suivant.

Proposition *Supposons que la famille $(b(\omega), \omega \in \Omega)$ est sommable, et soit $(F_n, n \in \mathbb{N})$ une suite croissante de parties finies de Ω telle que $\bigcup_{n\in\mathbb{N}} F_n = \Omega$. Alors la suite*

$$\sum_{F_n} b(\omega)\,, \quad n \in \mathbb{N}$$

converge quand $n \to \infty$ et sa limite vaut $\sum_\Omega b(\omega)$. Enfin on a l'inégalité

$$|\sum_\Omega b(\omega)| \;\leq\; \sum_\Omega |b(\omega)|\,.$$

Preuve: La première assertion découle de la commutativité de l'addition et de la proposition précédente. Plus précisément, F_n étant fini, on a

$$\sum_{F_n} b(\omega) \;=\; \sum_{F_n} b(\omega)^+ - \sum_{F_n} b(\omega)^-$$

et on sait que les deux termes du membre de droite convergent vers $\sum_\Omega b(\omega)^+$ et $\sum_\Omega b(\omega)^-$, respectivement. L'inégalité dans la seconde assertion est évidente de la définition même de $\sum_\Omega b(\omega)$. ■

Il découle immédiatement de cette proposition que quand $\Omega = \mathbb{N}$, la famille $(b(\omega), \omega \in \Omega)$ est sommable si et seulement si la série $\sum b(n)$ est absolument convergente, et alors $\sum_\Omega b(\omega) \;=\; \sum_0^\infty b(n)$.

Nous allons maintenant conclure cette section en démontrant deux propriétés fondamentales de la somme d'une famille sommable.

Linéarité *Si $(b(\omega), \omega \in \Omega)$ et $(b'(\omega), \omega \in \Omega)$ sont deux familles sommables, et si α, α' sont deux réels, alors la famille $(\alpha b(\omega) + \alpha' b'(\omega), \omega \in \Omega)$ est elle aussi sommable et*

$$\sum_\Omega (\alpha b(\omega) + \alpha' b'(\omega)) \;=\; \alpha \left(\sum_\Omega b(\omega)\right) + \alpha' \left(\sum_\Omega b'(\omega)\right)\,.$$

Preuve: Pour toute partie finie $F \subseteq \Omega$, on a

$$|\sum_F (\alpha b(\omega) + \alpha' b'(\omega))| \leq |\alpha| \left(\sum_F |b(\omega)|\right) + |\alpha'| \left(\sum_F |b'(\omega)|\right),$$

et comme $(b(\omega), \omega \in \Omega)$ et $(b'(\omega), \omega \in \Omega)$ sont sommables, il en est de même pour $(\alpha b(\omega) + \alpha' b'(\omega), \omega \in \Omega)$. De plus, on a l'égalité

$$\sum_F (\alpha b(\omega) + \alpha' b'(\omega)) = \alpha \left(\sum_F b(\omega)\right) + \alpha' \left(\sum_F b'(\omega)\right)$$

pour toute partie finie F. En considérant une suite croissante de parties finies $(F_n, n \in \mathbb{N})$ telle que $\Omega = \bigcup F_n$, on conclut qu'on a bien

$$\sum_\Omega (\alpha b(\omega) + \alpha' b'(\omega)) = \alpha \left(\sum_\Omega b(\omega)\right) + \alpha' \left(\sum_\Omega b'(\omega)\right). \quad \diamond$$

Sommation par paquets *Soit I un ensemble fini ou dénombrable, et $(\Omega_i, i \in I)$ une partition de Ω, c'est-a-dire que $\Omega_i \in \mathcal{P}(\Omega)$, $\Omega_i \cap \Omega_j = \emptyset$ si $i \neq j$, et $\Omega = \bigcup_I \Omega_i$. Soit $(b(\omega), \omega \in \Omega)$ une famille sommable. Alors pour tout $i \in I$, la sous-famille $(b(\omega), \omega \in \Omega_i)$ est sommable. Si on note $\sigma(i) = \sum_{\Omega_i} b(\omega)$, alors la famille $(\sigma(i), i \in I)$ est elle aussi sommable, et*

$$\sum_I \sigma(i) = \sum_I \left(\sum_{\Omega_i} b(\omega)\right) = \sum_\Omega b(\omega).$$

Preuve: Pour tout $i \in I$ et toute partie finie $F_i \subseteq \Omega_i$, F_i est a fortiori une partie finie de Ω, et il est alors immédiat que la sous-famille $(b(\omega), \omega \in \Omega_i)$ est sommable.

Pour montrer la formule de sommation par paquets, supposons tout d'abord que l'ensemble I est fini. Pour chaque indice i, considérons une suite croissante de parties finies de Ω_i, $(F_n^i, n \in \mathbb{N})$ telle que $\bigcup_{n\in\mathbb{N}} F_n^i = \Omega_i$. Posons pour chaque entier n: $F_n = \bigcup_{i\in I} F_n^i$ (on notera que c'est une union d'ensembles disjoints), de sorte que $(F_n, n \in \mathbb{N})$ est une suite croissante de parties finies de Ω telle que $\bigcup_{n\in\mathbb{N}} F_n = \Omega$. On a

$$\sum_I \left(\sum_{F_n^i} b(\omega)\right) = \sum_{F_n} b(\omega),$$

puis en faisant tendre n vers l'infini, on trouve bien

$$\sum_I \left(\sum_{\Omega_i} b(\omega)\right) = \sum_\Omega b(\omega).$$

Supposons finalement que I est dénombrable, et considérons une sous-partie finie $J \subseteq I$. Posons $\Omega_J = \bigcup_{j\in J} \Omega_j$, de sorte qu'on a d'après ce qui précède

$$\sum_J |\sum_{\Omega_j} b(\omega)| \leq \sum_J \left(\sum_{\Omega_j} |b(\omega)|\right) = \sum_{\Omega_J} |b(\omega)| \leq \sum_\Omega |b(\omega)|.$$

Ceci montre que la famille $(\sigma_i, i \in I)$ est bien sommable.

Considérons ensuite une suite croissante $(J(k), k \in \mathbb{N})$ de parties finies de I avec $\bigcup J(k) = I$. On a alors $\Omega_{J(k)} = \bigcup_{j \in J(k)} \Omega_j$. On sait donc que pour tout k

$$\sum_{J(k)} \sigma(j) \; = \; \sum_{J(k)} \left(\sum_{\Omega_j} b(\omega) \right) \; = \; \sum_{\Omega_{J(k)}} b(\omega) \, .$$

Comme $(b(\omega), \omega \in \Omega)$ est sommable, pour tout $\varepsilon > 0$, on peut trouver une partie finie $F \subseteq \Omega$ telle que

$$\sum_{\Omega'} |b(\omega)| \; < \; \varepsilon \qquad \text{pour tout } \Omega' \subseteq \Omega - F.$$

Or $\bigcup \Omega_{J(k)} = \Omega$, de sorte qu'il existe un entier k_0 tel que $F \subseteq \Omega_{J(k)}$ pour tout $k \geq k_0$. En particulier, on a alors pour $k \geq k_0$

$$|\sum_{\Omega} b(\omega) \; - \; \sum_{J(k)} \sigma(j)| \; = |\sum_{\Omega - \Omega_{J(k)}} b(\omega)| \; \leq \; \sum_{\Omega - F} |b(\omega)| \; \leq \; \varepsilon \, .$$

En faisant tendre ε vers 0, on voit que la formule de sommation par paquets est correcte. ■

2.2 Mesure de probabilité

On utilisera désormais le langage probabiliste au lieu du langage ensembliste, à savoir qu'on appellera Ω l'univers, les parties de Ω des événements, et un élément $\omega \in \Omega$ un événement élémentaire.

On appelle mesure de probabilité sur Ω une application $\mathbb{P} : \mathcal{P}(\Omega) \to [0, 1]$ qui vérifie les axiomes suivants.

A-1: $\mathbb{P}(\Omega) = 1$

A-2: Si Λ et Λ' sont deux parties disjointes de Ω, alors $\mathbb{P}(\Lambda \cup \Lambda') = \mathbb{P}(\Lambda) + \mathbb{P}(\Lambda')$.

A-3: Si $(\Lambda_n, n \in \mathbb{N})$ est une suite croissante d'événements, alors

$$\mathbb{P}(\bigcup \Lambda_n) \; = \; \lim_{n \to \infty} \mathbb{P}(\Lambda_n) \, .$$

En utilisant A-2 et en passant aux événements complémentaires, on voit qu'on peut aussi écrire A-3 sous la forme équivalente:

A-3(bis) Si $(\Lambda_n, n \in \mathbb{N})$ est une suite décroissante d'événements, alors

$$\mathbb{P}(\bigcap \Lambda_n) \; = \; \lim_{n \to \infty} \mathbb{P}(\Lambda_n) \, .$$

L'axiome A-3 n'est utile que lorsque l'univers Ω est infini. En effet, dans un univers fini, toute suite croissante d'événements est stationnaire (c'est-à-dire constante à partir d'un certain rang), et A-3 est trivialement satisfait.

La mesure de probabilité $\mathbb{P}$ induit en particulier une application $p : \Omega \to [0, 1]$, donnée par $p(\omega) = \mathbb{P}(\{\omega\})$.

Proposition. *Pour tout événement Λ, on a*

$$\mathbb{P}(\Lambda) \;=\; \sum_{\Lambda} p(\omega).$$

Preuve: C'est évident si $\Lambda = \{\omega\}$ est un singleton. En appliquant le second axiome, la proposition est encore vérifiée si Λ est un doubleton, et par récurrence, on voit que la proposition est vraie si Λ est un événement fini.

Passons au cas où Λ est infini. On peut numéroter ses éléments, $\Lambda = \{\omega_0, \cdots, \omega_n, \cdot\}$, et pour tout entier n, on pose $\Lambda_n = \{\omega_0, \cdots, \omega_n\}$. On sait que $\mathbb{P}(\Lambda_n) \;=\; \sum_0^n p(\omega_i)$. La suite des événements Λ_n est croissante, sa limite est Λ, et on déduit de l'axiome A-3 que

$$\mathbb{P}(\Lambda) \;=\; \sum_{0}^{\infty} p(\omega_i)\,.$$

D'après un résultat vu dans la première section, on sait que le terme de droite est égal à $\sum_\Lambda p(\omega)$. ■

En particulier, $(p(\omega), \omega \in \Omega)$ est une famille sommable à valeurs positives et telle que $\sum_\Omega p(\omega) = 1$ (d'après l'axiome A-1). On a une réciproque:

Proposition *Soit $(p(\omega), \omega \in \Omega)$ est une famille sommable à valeurs positives et telle que $\sum_\Omega p(\omega) = 1$. Pour tout $\Lambda \in \mathcal{P}(\Omega)$, on pose*

$$\mathbb{P}(\Lambda) \;=\; \sum_{\Lambda} p(\omega).$$

Alors $\mathbb{P}$ est une mesure de probabilité sur Ω.

Preuve: L'axiome A-1 est immédiatement vérifié. Pour établir A-2, considérons deux événements disjoints, Λ et Λ'. Si Λ et Λ' sont tous les deux finis, on a bien

$$\sum_{\Lambda\cup\Lambda'} p(\omega) \;=\; \left(\sum_{\Lambda} p(\omega)\right) + \left(\sum_{\Lambda'} p(\omega)\right)$$

(par commutativité de l'addition). Passons au cas général. Pour tout $\varepsilon > 0$, on peut trouver une partie finie $F \subseteq \Lambda$ et une partie finie $F' \subseteq \Lambda'$ telles que

$$\sum_{\Lambda} p(\omega) - \sum_{F} p(\omega) \;< \varepsilon \qquad \text{et} \qquad \sum_{\Lambda'} p(\omega) - \sum_{F'} p(\omega) \;< \varepsilon\,.$$

Comme Λ et Λ' sont disjoints, il en est de même pour F et F', et on déduit que

$$\sum_{\Lambda} p(\omega) + \sum_{\Lambda'} p(\omega) - \sum_{F\cup F'} p(\omega) \leq 2\varepsilon\,.$$

Or $F \bigcup F'$ est une partie finie de $\Lambda \bigcup \Lambda'$, et comme ε est arbitrairement petit, on a

$$\sum_{\Lambda} p(\omega) + \sum_{\Lambda'} p(\omega) \;\leq\; \sum_{\Lambda\cup\Lambda'} p(\omega)\,.$$

Pour établir l'inégalité inverse, on se donne une partie finie $G \subseteq \Lambda \bigcup \Lambda'$, qu'on peut écrire sous la forme $G = F \bigcup F'$, où $F \subseteq \Lambda$ et $F' \subseteq \Lambda'$ sont toutes deux finies. On a donc

$$\sum_G p(\omega) \;=\; \sum_F p(\omega) + \sum_{F'} p(\omega) \;\leq\; \sum_\Lambda p(\omega) + \sum_{\Lambda'} p(\omega)\,.$$

En prenant le supremum sur les parties finies G, on a bien

$$\sum_\Lambda p(\omega) + \sum_{\Lambda'} p(\omega) \;\geq\; \sum_{\Lambda\cup\Lambda'} p(\omega)\,.$$

Vérifions maintenant l'axiome A-3. Soit $(\Lambda_n, n \in \mathbb{N})$ une suite croissante d'évènements; posons $\Lambda = \bigcup \Lambda_n$. La suite réelle $\sum_{\Lambda_n} p(\omega)$, $n \in \mathbb{N}$ est croissante, elle admet donc une limite que nous notons ℓ. Si $F \subseteq \Lambda$ est un évènement fini, alors il existe un entier n tel que $F \subseteq \Lambda_n$ et on déduit que $\ell \geq \sum_\Lambda p(\omega)$. Réciproquement, on a $\Lambda_n \subseteq \Lambda$ pour tout entier n, et donc

$$\sum_{\Lambda_n} p(\omega) \leq \sum_\Lambda p(\omega)\,;$$

en faisant tendre n vers ∞, ceci entraine $\ell \leq \sum_\Lambda p(\omega)$. ■

2.3 Variables aléatoires discrètes et leurs lois

Une variable aléatoire est une application X sur Ω, à valeurs dans un certain ensemble E. Le plus souvent, $E = \mathbb{N}, \mathbb{Z}, \mathbb{R}$ ou $\mathbb{R}^d$. De façon informelle, une variable aléatoire represente l'effet de l'aléas sur une expérience; par exemple X peut representer le gain d'un joueur à la loterie...

Une variable aléatoire permet de transporter la mesure de probabilité $\mathbb{P}$ sur Ω en une mesure de probabilité sur l'ensemble des valeurs prises par X.

Proposition et Définition *Pour tout sous-ensemble $E' \subseteq E$, on note*

$$\{X \in E'\} \;=\; \{\omega \in \Omega : X(\omega) \in E'\}\,.$$

L'application $P_X : \mathcal{P}(E) \to [0,1]$ définie par

$$P_X(E') \;=\; \mathbb{P}(\{X \in E'\})\,, \qquad E' \in \mathcal{P}(E)$$

est une mesure de probabilité sur $\mathcal{P}(E)$. On l'appelle la loi de la variable aléatoire X, ou encore sa distribution.

Preuve: On a $P_X(E) = \mathbb{P}(\{X \in E\}) = 1$, et donc l'axiome A-1 est satisfait.

Si E_1 et E_2 sont deux parties disjointes de E, alors les événements $\{X \in E_1\}$ et $\{X \in E_2\}$ sont contradictoires et donc

$$\begin{aligned} P_X(E_1 \cup E_2) &= \mathbb{P}(\{X \in E_1 \cup E_2\}) = \mathbb{P}(\{X \in E_1\} \cup \{X \in E_2\}) \\ &= \mathbb{P}(\{X \in E_1\} + \mathbb{P}(\{X \in E_2\}) \;=\; P_X(E_1) + P_X(E_2). \end{aligned}$$

Ceci établit A-2.

Enfin, soit $(E_n, n \in \mathbb{N})$ une suite croissante de parties de E; on note $E_\infty = \bigcup E_n$. La suite d'événements $\{X \in E_n\}$ est elle aussi croissante, et on vérifie immédiatement que

$$\bigcup \{X \in E_n\} \,=\, \{X \in E_\infty\}\,.$$

On a donc

$$\lim_{n\to\infty} P_X(E_n) \,=\, \lim_{n\to\infty} \mathbb{P}\left(\{X \in E_n\}\right) \,=\, \mathbb{P}\left(\{X \in E_\infty\}\right) \,=\, P_X(E_\infty)\,,$$

et A-3 est démontré. ■

Si on note $\{x_n, n \in I\}$ l'ensemble des valeurs que peut prendre X (l'ensemble I est nécessairement au plus dénombrable, puisque c'est le cas pour Ω), on voit donc que la loi de X est caractérisée par la donnée de

$$p_X(x_i) \,=\, P_X(\{x_i\}) \,=\, \mathbb{P}(\{X = x_i\})\,, \qquad i \in I.$$

Plus précisément, on a pour tout $E' \subseteq E$

$$P_X(E') \,=\, \sum_{x\in E'} p_X(x)\,.$$

Nous allons maintenant supposer que X est une variable aléatoire réelle, c'est-à-dire que $E = \mathbb{R}$.

Définition *On appelle fonction de répartition de X la fonction $F_X : \mathbb{R} \to [0,1]$ donnée par:*

$$F_X(x) \,=\, P_X(]-\infty, x]) \,=\, \mathbb{P}(\{X \le x\})\,, \qquad x \in \mathbb{R}\,.$$

Nous allons montrer maintenant qu'une fonction de répartition vérifie les propriétés élémentaires suivantes.

Proposition *(i) La fonction F_X est croissante, avec*

$$\lim_{x\to-\infty} F_X(x) = 0 \quad \textit{et} \quad \lim_{x\to\infty} F_X(x) = 1.$$

(ii) La fonction F_X est continue à droite, sa limite à gauche au point x vaut

$$F_X(x-) \,=\, \lim_{y\to x-} F_X(y) \,=\, P_X(]-\infty, x[)\,.$$

(iii) Pour tout $x \in \mathbb{R}$, on a

$$p_X(x) = \mathbb{P}(\{X = x\}) = F_X(x) - F_X(x-)\,;$$

en particulier, la fonction de répartition F_X caractérise la loi de X.

preuve: (i) Si $x \le x'$, alors $]-\infty, x] \subseteq]-\infty, x']$ et on a bien $F_X(x) \le F_X(x')$. Comme $]-\infty, \infty[= \bigcup_{n\in\mathbb{N}}]-\infty, n]$, on a

$$\lim_{n\to\infty} F_X(n) \,=\, P_X(]-\infty, \infty[) \,=\, 1.$$

De même $]-\infty,\infty[=\bigcup_{n\in\mathbb{N}}]-n,\infty[$ et comme $F_X(-n)=1-P_X(]-n,\infty[)$, on a donc

$$1-\lim_{n\to\infty} F_X(-n) \;=\; P_X(]-\infty,\infty[) \;=\; 1\,.$$

(ii) Soit $(x_n, n\in\mathbb{N})$ une suite réelle qui décroît vers x. La suite des évènements $\{X\le x_n\}$ est décroissante et $\{X\le x\} = \bigcap\{X\le x_n\}$. On a donc $F_X(x) = \lim F_X(x_n)$. Supposons maintenant que $(x'_n, n\in\mathbb{N})$ est une suite strictement croissante qui converge vers x. Alors la suite des événements $\{X\le x'_n\}$ est croissante et $\{X<x\} = \bigcup\{X\le x'_n\}$. On a donc $F_X(x-) = \lim F_X(x'_n)$.

(iii) Il suffit d'écrire $\{X\le x\} = \{X<x\}\cup\{X=x\}$ et d'appliquer A-2. ■

2.4 Moments d'une variable aléatoire réelle

On suppose à nouveau que X est une variable aléatoire réelle, et on introduit la notion suivante:

Définition *Si la famille $\{|X(\omega)|p(\omega), \omega\in\Omega\}$ est sommable, on dit que X admet un moment d'ordre 1 et on pose*

$$\mathbb{E}(X) \;=\; \sum_{\Omega} X(\omega)p(\omega).$$

On appelle cette quantité l'espérance, ou la moyenne, de X.

Plus généralement, pour tout réel $k>0$, on dit que X admet un moment d'ordre k si la variable aléatoire $|X|^k$ admet un moment d'ordre 1.

Il est immédiat de voir que si la variable X est bornée, c'est-à-dire s'il existe un réel $M>0$ tel que $|X(\omega)|\le M$ pour tout $\omega\in\Omega$, alors X admet des moments de tous ordres. C'est le cas en particuler lorsque l'univers Ω est fini. L'espérance satisfait les propriétés suivantes:

Proposition *(i) Si la variable aléatoire X admet un moment d'ordre 1, et si $\{x_i, i\in I\}$ désigne l'ensemble des valeurs prises par X, alors la famille $(x_i\mathbb{P}(\{X=x_i\}), i\in I)$ est sommable et*

$$\mathbb{E}\,(X) \;=\; \sum_{i\in I} x_i\mathbb{P}(\{X=x_i\})\,.$$

(ii) Si X et Y sont deux variables aléatoires qui admettent toutes les deux un moment d'ordre 1, alors pour tout $\alpha,\beta\in\mathbb{R}$, c'est aussi le cas pour la variable $\alpha X+\beta Y$ et on a

$$\mathbb{E}\,(\alpha X+\beta Y) \;=\; \alpha\mathbb{E}(X)+\beta\mathbb{E}(Y).$$

Preuve: (i) Posons $\Omega_i=\{\omega\in\Omega : X(\omega)=x_i\}$. La famille $\{\Omega_i, i\in I\}$ est une partition de Ω, et il ne reste qu'à appliquer la formule de sommation par paquets.

(ii) C'est une conséquence de la linéarité de la somme pour les familles sommables. ■

Proposition et définition *Si X admet un moment d'ordre 2, alors il admet également un moment d'ordre 1. On pose alors*

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

et on appelle cette quantité la variance de X. On a l'identité

$$\mathrm{Var}(X) \;=\; \mathbb{E}\left((X - \mathbb{E}(X))^2\right)\,.$$

En conséquence, $\mathrm{Var}(X) \geq 0$*, on appelle* $\sigma = \sqrt{\mathrm{Var}(X)}$ *l'écart-type de X.*

Preuve: En considérant la partition de l'univers en $\Omega = \{|X| \leq 1\} \cup \{|X| > 1\}$, on déduit que

$$\sum_{\Omega} |X(\omega)|p(\omega) \leq \sum_{\{|X|\leq 1\}} p(\omega) + \sum_{\{|X|>1\}} X(\omega)^2 p(\omega) \leq \mathbb{P}(\{|X| \leq 1\}) + \mathbb{E}(X^2)\,,$$

ce qui montre que X a un moment d'ordre 1. Ensuite, il suffit de poser $m = \mathbb{E}(X)$ et d'écrire

$$\mathbb{E}((X-m)^2) \;=\; \mathbb{E}(X^2 - 2mX + m^2) \;=\; \mathbb{E}(X^2) - 2m\mathbb{E}(X) + m^2 \;=\; \mathbb{E}(X^2) - m^2. \quad \diamond$$

Les notions d'espérance et de variance sont très utile pour estimer la queue de la distribution d'une variable aléatoire.

Inégalité de Markov *Soit X une variable aléatoire qui admet un moment d'ordre 1. Pour tout réel $a > 0$, on a*

$$\mathbb{P}\left(\{|X| \geq a\}\right) \;\leq\; a^{-1}\mathbb{E}(|X|).$$

Preuve: Il suffit d'écrire

$$\begin{aligned}\mathbb{E}(|X|) \;=\; \sum_{\Omega} |X(\omega)|p(\omega) \quad &\geq \quad \sum_{|X(\omega)|\geq a} |X(\omega)|p(\omega) \\ &\geq \quad \sum_{|X(\omega)|\geq a} ap(\omega) = a\mathbb{P}\left(\{|X| \geq a\}\right)\,.\end{aligned}$$

■

Inégalité de Bienaymé-Chebitchev *Soit X une variable aléatoire qui admet un moment d'ordre 2. Pour tout réel $a > 0$, on a*

$$\mathbb{P}\left(\{|X - \mathbb{E}(X)| \geq a\}\right) \;\leq\; a^{-2}\mathrm{Var}(X).$$

Preuve: Pour simplifier, posons $m = \mathbb{E}(X)$. On a

$$\mathbb{P}\left(|X - m| \geq a\right) \;=\; \mathbb{P}\left(|X-m|^2 \geq a^2\right)\,,$$

et il ne reste qu'à appliquer l'inégalité de Markov. ■

Corollaire *Une variable aléatoire est constante avec probabilité 1 si et seulement si sa variance est nulle. Elle est alors égale à sa valeur moyenne.*

Preuve: Si X a une variance nulle, alors l'inégalité de Bienaymé-Chebitchev entraine que pour tout $\varepsilon > 0$, $\mathbb{P}(\{|X - \mathbb{E}(X)| > \varepsilon\}) = 0$, c'est-à-dire $\mathbb{P}(\{X = \mathbb{E}(X)\}) = 1$. La réciproque est évidente. ■

2.5 Fonction génératrice d'une v.a. à valeurs entières

On suppose dans cette partie que X est une v.a. qui prend toutes ses valeurs dans $\mathbb{N}$.

Définition *On appelle fonction génératrice de X la fonction $G_X : [0,1] \to [0,1]$ donnée par*

$$G_X(s) \;=\; \mathbb{E}(s^X) \;=\; \sum_{n=0}^{\infty} s^n \mathbb{P}(X = n)\,.$$

(On utilise la convention $0^0 = 1$ dans le cas $s = n = 0$.)

Proposition *La fonction génératrice est une fonction entière sur $[0,1]$, de rayon de convergence supérieur ou égal à 1. Elle détermine la loi de X, plus précisément on a pour tout entier n*

$$\mathbb{P}(X = n) \;=\; \frac{G_X^{(n)}(0)}{n!}\,,$$

où $G_X^{(n)}$ désigne la dérivée n-ième de G_X.

Preuve: Le fait que G_X soit une fonction entière est clair sur sa définition. Comme la série correspondant à $G_X(1)$ a tous ses coéfficients positifs et que $G_X(1) = \mathbb{E}(1^X) = 1$, le rayon est au moins égal à 1. La formule qui donne les coefficients d'une série entière en fonction de ses dérivées à l'origine termine la preuve. ■

Exemples fondamentaux • Prenons pour X une variable de Bernoulli de paramètre $p \in [0,1]$, c'est-à-dire $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$. On a $G_X(s) = 1 - p + sp$.

• Prenons pour X une variable binomiale de paramètres (n, p), où $n \in \mathbb{N}$ et $p \in [0,1]$, c'est-à-dire

$$\mathbb{P}(X = k) \;=\; \tbinom{n}{k}\, p^k (1-p)^{n-k}\,, \qquad k = 0, 1, \cdots, n\,.$$

(Pour vérifier que $\sum_0^n \mathbb{P}(X = k)$ vaut bien 1, il suffit d'appliquer la formule de Newton.) On a d'après la formule de Newton:

$$G_X(s) \;=\; \sum_{k=0}^{n} \tbinom{n}{k}\, p^k (1-p)^{n-k} s^k \;=\; (ps + 1 - p)^n\,.$$

• Prenons pour X une variable géométrique de paramètre $p \in]0,1[$, c'est-à-dire

$$\mathbb{P}(X = n) \;=\; (1-p)p^n \qquad n \in \mathbb{N}.$$

(Pour vérifier que $\sum_0^\infty \mathbb{P}(X=n)$ vaut bien 1, il suffit d'appliquer la formule pour la somme d'une série géométrique.) On a alors

$$G_X(s) \;=\; \sum_{n=0}^{\infty} s^n(1-p)p^n \;=\; \frac{1-p}{1-sp}\,.$$

• Prenons pour X une variable de Poisson de paramètre $c>0$, c'est-à-dire

$$\mathbb{P}(X=n) \;=\; \mathrm{e}^{-c}\frac{c^n}{n!}\,, \qquad n\in\mathbb{N}.$$

On a alors

$$G_X(s) \;=\; \mathrm{e}^{-c}\sum_{n=0}^{\infty} s^n\frac{c^n}{n!} \;=\; \mathrm{e}^{-c}\mathrm{e}^{cs} \;=\; \mathrm{e}^{c(s-1)}\,.$$

2.6 Variables aléatoires indépendantes

Soient $X_1,\cdots,X_n$ une famille de n variables aléatoires discrètes, à valeurs dans $E_1,\cdots,E_n$.

Définition. *On dit que les v.a. $X_1,\cdots,X_n$ sont indépendantes si*

$$\mathbb{P}\left(X_1=x_1,\cdots,X_n=x_n\right) \;=\; \prod_{i=1}^{n}\mathbb{P}(X_i=x_i)$$

pour tout $x_1\in E_1,\cdots,x_n\in E_n$.

Il convient de prendre garde au point suivant: il existe des variables X_1,X_2,X_3 qui sont deux à deux indépendantes sans que le triplet le soit. Par exemple, si B_1 et B_2 sont deux variables de Bernoulli indépendantes, on peut prendre $X_1=B_1$, $X_2=B_2$ et $X_3=B_1B_2$.

Il est immédiat de renforcer cette propriété.

Proposition. *Les v.a. $X_1,\cdots,X_n$ sont indépendantes si et seulement si*

$$\mathbb{E}\left(f_1(X_1)\cdots f_n(X_n)\right) \;=\; \prod_{i=1}^{n}\mathbb{E}(f_i(X_i))\,,$$

pour toutes les fonctions bornées $f_i : E_i\to\mathbb{R}$.

Preuve: Pour simplifier, nous supposerons $n=2$; le cas général est analogue mais avec des notations plus lourdes

$$\begin{aligned}
&\mathbb{E}(f_1(X_1)f_2(X_2))\\
=\;&\sum_{\omega\in\Omega} f_1(X_1(\omega))f_2(X_2(\omega))p(\omega)\\
=\;&\sum_{E_1\times E_2} f_1(x_1)f_2(x_2)\mathbb{P}(X_1=x_1,X_2=x_2) \quad \text{(sommation par paquets)}\\
=\;&\sum_{E_1\times E_2} f_1(x_1)f_2(x_2)\mathbb{P}(X_1=x_1)\mathbb{P}(X_2=x_2) \quad \text{(indépendance)}\\
=\;&\left(\sum_{E_1} f(x_1)\mathbb{P}(X_1=x_1)\right)\left(\sum_{x_2\in E_2} f(x_2)\mathbb{P}(X_2=x_2)\right) \quad \text{(sommation par paquets)}\\
=\;&\mathbb{E}(f_1(X_1))\mathbb{E}(f_2(X_2))\,.
\end{aligned}$$

La réciproque est évidente. ■

Voici une application simple de ce résultat.

Corollaire *Si X et Y sont deux variables aléatoires réelles indépendantes qui admettent toutes les deux un moment d'ordre deux, alors* $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Preuve: Grâce à la linéarité et à l'indépendance, on a

$$\begin{aligned} \mathbb{E}((X+Y)^2) &= \mathbb{E}\left(X^2 + 2XY + Y^2\right) = \mathbb{E}(X^2) + 2\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(Y^2) \\ \mathbb{E}(X+Y)^2 &= \mathbb{E}(X)^2 + 2\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(Y)^2\,, \end{aligned}$$

ce qui conduit au résultat en prenant la différence. ■

Les fonctions génératrices sont très utiles pour étudier les sommes de variables indépendantes comme le montre le résultat suivant.

Corollaire *Si $X_1, \cdots, X_n$ sont n variables aléatoires à valeurs entières indépendantes, alors la fonction génératrice de la variable $X_1 + \cdots + X_n$ est le produit des fonctions génératrices, $G_{X_1+\cdots+X_n} = \prod_{i=1}^n G_{X_i}$.*

Preuve: On a pour tout $s \in [0,1]$:

$$G_{X_1+\cdots+X_n}(s) = \mathbb{E}\left(s^{X_1+\cdots+X_n}\right) = \mathbb{E}\left(s^{X_1}\cdots s^{X_n}\right) = \prod_{i=1}^n \mathbb{E}\left(s^{X_i}\right) = \prod_{i=1}^n G_{X_i}(s)\,. \diamond$$

Corollaire *Soit $X_1, \cdots, X_n$, n variables indépendantes qui suivent toutes la loi de Bernoulli de paramètre p, $0 < p < 1$. Alors $S = X_1 + \cdots + X_n$ suit la loi binomiale de paramètres (n,p).*

Preuve: D'après la proposition précédente, la fonction génératrice de S est $G_S = G_X^n$, où X désigne une variable de Bernoulli de paramètre p. Comme $G_X(s) = 1 - p + sp$, on a donc $G_S(s) = (1-p+sp)^n$, qui est la fonction génératrice de la loi binomiale de paramètres (n,p). Comme la fonction génératrice caractérise la loi, le corollaire est démontré. ■

On peut démontrer de la même manière que si X et X' sont deux variables indépendantes qui suivent des lois de Poisson de paramètres c et c', respectivement, alors $X + X'$ suit encore un loi de Poisson de paramètre $c + c'$ (exercice).

Chapitre 3

Espaces de Probabilités Généraux et Variables Aléatoires

3.1 Axiomatique de Kolmogorov

On se donne un espace abstrait Ω, qu'on suppose muni d'une tribu $\mathcal{F}$. C'est-à-dire que $\mathcal{F}$ est une partie de $\mathcal{P}(\Omega)$ telle que:

(a) $\Omega \in \mathcal{F}$.

(b) Si $\Lambda \in \mathcal{F}$, alors $\Lambda^c \in \mathcal{F}$.

(c) Si $(\Lambda_n, n \in \mathbb{N})$ est une suite d'éléments dans $\mathcal{F}$, alors $\bigcup \Lambda_n \in \mathcal{F}$.

Par passage au complémentaire, on a aussi sous les hypothèses de (c) que $\bigcap \Lambda_n \in \mathcal{F}$.

On appelle mesure de probabilité sur $(\Omega, \mathcal{F})$ une application $\mathbb{P} : \mathcal{F} \to [0,1]$ qui vérifie les axiomes suivants:

A.1: $\mathbb{P}(\Omega) = 1$

A.2: Si $\Lambda, \Lambda' \in \mathcal{F}$ sont disjoints, alors $\mathbb{P}(\Lambda \cup \Lambda') = \mathbb{P}(\Lambda) + \mathbb{P}(\Lambda')$.

A.3: Si $(\Lambda_n, n \in \mathbb{N})$ est une suite croissante d'éléments dans $\mathcal{F}$, alors $\mathbb{P}(\bigcup \Lambda_n) = \lim_{n\to\infty} \mathbb{P}(\Lambda_n)$.

On fera référence à A.3 comme la propriété de σ-additivité de $\mathbb{P}$. On notera que, par passage au complémentaire dans A.3, si $(\Lambda_n, n \in \mathbb{N})$ est une suite décroissante d'éléments dans $\mathcal{F}$, alors $\mathbb{P}(\bigcap \Lambda_n) = \lim_{n\to\infty} \mathbb{P}(\Lambda_n)$.

On dira que la tribu $\mathcal{F}$ est complète pour $\mathbb{P}$ si pour toute sous-partie A d'une partie $N \in \mathcal{F}$ de probabilité nulle, $\mathbb{P}(N) = 0$, on a $A \in \mathcal{F}$. Il est de plus clair que dans ce cas, $\mathbb{P}(A) = 0$. Il facile de construire une tribu $\mathcal{F}'$ qui contient $\mathcal{F}$ et d'étendre $\mathbb{P}$ à $\mathcal{F}'$ de telle sorte que $\mathcal{F}'$ soit complète pour l'extension de $\mathbb{P}$. Il sera plus commode pour nous de toujours supposer que $\mathcal{F}$ est complète pour $\mathbb{P}$. On dira qu'un évènement Λ est vérifié presque sûrement si son complémentaire $\Lambda^c = \Omega \backslash \Lambda$ a une probabilité nulle. On utilisera l'abréviation p.s. pour presque-sûr.

Si E est un espace muni d'une tribu $\mathcal{E}$ (on prendra souvent pour E la droite réelle ou un espace euclidien et pour $\mathcal{E}$ la tribu borélienne), on appelle variable aléatoire à valeurs dans E toute application mesurable $X : \Omega \to E$.

Lorsque X est une v.a. réelle, on dit que X admet un moment d'ordre 1 si $\int_\Omega |X(\omega)| d\mathbb{P}(\omega) < \infty$. Dans ce cas, on pose

$$\mathbb{E}(X) \;=\; \int_\Omega X(\omega) d\mathbb{P}(\omega)$$

et on lit le terme de gauche "espérance de X", ou "moyenne de X". L'espace des variables aléatoires réelles (ou complexes) qui admettent un moment d'ordre 1 est noté $L^1(\Omega, \mathbb{P})$; c'est un espace vectoriel et l'application $X \to \mathbb{E}(X)$ est linéaire et positive, i.e.

$$0 \le X \le Y \quad \Longrightarrow 0 \le \mathbb{E}(X) \le \mathbb{E}(Y)\,.$$

Dans ce cours on fera souvent appel aux résultats suivants de la théorie de la mesure:

Théorème de convergence monotone *Si $(X_n, n \in \mathbb{N})$ est une suite croissante de variables aléatoires positives, c'est-à-dire que p.s. $0 \le X_n \le X_{n+1}$ pour tout $n \in \mathbb{N}$, et si on note $X_\infty = \lim_{n\to\infty} X_n$, alors*

$$\mathbb{E}(X_\infty) \;=\; \lim_{n\to\infty} \mathbb{E}(X_n)\,.$$

Lemme de Fatou *Si $(X_n, n \in \mathbb{N})$ est une suite de variables aléatoires positives, alors*

$$\mathbb{E}\left(\liminf_{n\to\infty} X_n\right) \;\le\; \liminf_{n\to\infty} \mathbb{E}(X_n)\,.$$

Théorème de convergence dominée *Si $(X_n, n \in \mathbb{N})$ est une suite de variables aléatoires réelles qui converge p.s. vers une variable X_∞, et s'il existe une variable Y intégrable avec $|X_n| \le Y$ p.s. pour tout $n \in \mathbb{N}$ alors*

$$\mathbb{E}(X_\infty) \;=\; \lim_{n\to\infty} \mathbb{E}(X_n)\,.$$

Plus généralement, pour tout $p > 0$, on note $L^p(\Omega, \mathbb{P})$ l'espace des v.a. X telles que $\mathbb{E}(|X|^p) < \infty$. On rappelle que pour $p \ge 1$, $L^p(\Omega, \mathbb{P})$ est un espace de Banach lorsqu'il est muni de la norme $||X||_p = \mathbb{E}(|X|^p)^{1/p}$.

Les résultats que nous avons établis dans le chapitre précédent sur les v.a. discrètes (définition et formules pour la variance, inégalité de Bienaymé-Chebitchev,...) sont encore valables dans le cadre général; nous ne les ré-écrivons pas et laissons la preuve comme exercice.

3.2 Loi d'une variable aléatoire

Soit $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ une variable aléatoire. On appelle loi de X la mesure image de $\mathbb{P}$ par X, c'est-à-dire la mesure de probabilité P_X sur $(E, \mathcal{E})$ donnée par

$$P_X(A) \;=\; \mathbb{P}(X \in A) \;=\; \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in A\}\right)\,, \qquad A \in \mathcal{E}\,.$$

(Il est très facile de vérifier que P_X est bien une mesure de probabilités sur $(E, \mathcal{E})$.) On peut aussi formuler cette définition sous la forme équivalent:

Proposition. *P_X est l'unique mesure sur $(E, \mathcal{E})$ telle que pour toute fonction mesurable et bornée $f : E \to \mathbb{R}$, on ait*

$$\mathbb{E}(f(X)) \;=\; \int_E f(x) dP_X(x)\,.$$

Preuve L'unicité est évidente en prenant pour f l'indicatrice d'un ensemble mesurable générique $A \in \mathcal{F}$, puisque dans ce cas

$$\mathbb{E}(f(X)) \;=\; \mathbb{E}(\mathbf{1}_{X\in A}) \;=\; \mathbb{P}(X \in A) \;=\; P_X(A)\,.$$

Par linéarité, l'identité de l'énoncé est vérifiée pour toute fonction mesurable f étagée. Par passage à la limite, on voit que l'identitée reste vraie pour f mesurable bornée générale. ■

La loi d'une variable aléatoire réelle est donc une mesure sur $\mathbb{R}$ (muni de la tribu borélienne), de masse 1. Voyons quelques exemples fondamentaux. Tout d'abord, commençons par quatre lois discrètes que nous avons déjà vues.

• La masse de Dirac en $x \in \mathbb{R}$, δ_x, est la loi de la variable aléatoire qui vaut x presque partout. Une telle variable est notée x par un (leger) abus de notation.

• La loi de Bernoulli de paramètre $p \in [0,1]$ est $p\delta_1 + (1-p)\delta_0$, où δ_x est le symbole de la masse de Dirac en x. De même, la loi de Poisson de paramètre $c > 0$ est $\sum_{n=0}^{\infty} \mathrm{e}^{-c} c^n (n!)^{-1} \delta_n$; et la loi géométrique de paramètre $p \in]0,1[$ est $\sum_{n=0}^{\infty}(1-p)p^n\delta_n$.

Passons ensuite à des lois sur $\mathbb{R}$ absolument continues par rapport à la mesure de Lebesgue, qu'on notera dx.

• La mesure de Lebesgue sur $[0,1]$ est une mesure de probabilité, qu'on appelle la loi uniforme sur $[0,1]$. Plus généralement, si $a < b$, on appelle loi uniforme sur $[a,b]$ la mesure de probabilité $(b-a)^{-1}\mathbf{1}_{[a,b]}(x)dx$.

• La mesure sur $\mathbb{R}$ de densité $\frac{1}{\pi}(1+x^2)^{-1}dx$ a bien pour masse 1, c'est une mesure de probabilité sur $\mathbb{R}$ qu'on appelle la loi de Cauchy standard.

• Pour tout $q > 0$, la mesure $q\mathrm{e}^{-qx}\mathbf{1}_{\{x\geq 0\}}dx$ a pour masse 1, c'est donc une mesure de probabilité qu'on appelle la loi exponentielle de paramètre q.

• En écrivant

$$\begin{aligned}\left(\int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}dx\right)^2 &= \left(\int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}dx\right)\left(\int_{-\infty}^{\infty} \mathrm{e}^{-y^2/2}dy\right)\\ &= \int_{-\infty}^{\infty}dx\int_{-\infty}^{\infty}dy\mathrm{e}^{-(x^2+y^2)/2}\\ &= 2\pi\int_0^{\infty}\mathrm{e}^{-r^2/2}rdr \;=\; 2\pi\,,\end{aligned}$$

on obtient l'identité $\int_{-\infty}^{\infty}\mathrm{e}^{-x^2/2}dx = \sqrt{2\pi}$. Ainsi, la mesure $\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}dx$ a pour masse 1, on l'appelle la loi normale standard. Plus généralement, pour tout $m \in \mathbb{R}$ et $\sigma^2 > 0$,

la mesure sur $\mathbb{R}$ de densité

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}$$

a également pour masse 1 (il s'agit d'un changement de variables simple), on l'appelle la loi Gauss de variance σ^2 centrée en m, et on la note $\mathcal{N}(m, \sigma^2)$.

Exercice: -a- Calculer la moyenne et la variance de la loi de Gauss (on trouve m et σ^2, respectivement).

-b- Pour tout entier k, calculer le moment d'ordre k de la loi exponentielle de paramètre q (on trouve $k!q^{-k}$) .

-c- Montrer que la loi de Cauchy n'admet pas de moment d'ordre 1.

Tout comme dans le cas discret, il est souvent commode de décrire une loi de probabilité sur $\mathbb{R}$ à l'aide de sa fonction de répartition:

Définition. *Si P_X est la loi d'une variable aléatoire réelle X, on appelle la fonction $F_X : \mathbb{R} \to [0,1]$ donnée par*

$$F_X(x) \;=\; P_X(]-\infty, x]) \;=\; \mathbb{P}(X \leq x)\,,$$

la fonction de répartition de X (ou de P_X). C'est une fonction croissante, continue à droite. On a $F_X(x-) = \mathbb{P}(X < x)$.

Par exemple, la fonction de répartition de la loi uniforme U sur $[0,1]$ est $F_U(x) = 0$ si $x \leq 0$, $F_U(x) = x$ si $0 \leq x \leq 1$, et $F_X(x) = 1$ si $x \geq 1$. La fonction de répartion de la loi de Cauchy standard C est $F_C(x) = \frac{1}{\pi}\arctan(x)$. La fonction de répartion de la loi de Gauss n'admet pas d'expression explicite simple.

La fonction de répartition F_X caractérise la loi P_X, puisque pour tout intervalle de type $]a, b]$, on a $F_X(a) - F_X(b) = P_X(]a, b])$, et qu'une mesure borélienne sur $\mathbb{R}$ est déterminée par la donnée des masses qu'elle attribue aux intervalles de ce type. Plus précisément, en tant que fonction croissante, F_X induit une mesure de Stieltjes sur $\mathbb{R}$ notée dF_X, et on a $P_X(dx) = dF_X(x)$.

On dit qu'une loi réelle est continue (ou aussi qu'elle n'a pas d'atome) si sa fonction de répartition F est continue. Une loi absolument continue (i.e. ayant une densité) par rapport à la mesure de Lebesgue est a fortiori continue. La fonction réciproque F^{-1} d'une distribution continue definie par

$$F^{-1}(y) \;=\; \inf\{x \in \mathbb{R} : F(x) = y\}\;, \qquad y \in]0,1[$$

s'appelle le *quantile* de la distribution F. Cette notion joue un rôle important en statistique.

Voyons une application à la simulation de variables aléatoires sur un ordinateur, qui repose sur l'idée suivante. Il est clair que si $f : \mathbb{R} \to \mathbb{R}$ est une fonction mesurable, alors l'image par f d'une v.a. X est une v.a. $Y = f(X)$. La loi de Y peut être déduite de celle de X par changement de variable.

Exemples: • Si U est une variable de loi uniforme sur $[0,1]$, alors $X = \log 1/U$ est une variable de loi exponentielle de paramètre 1. En effet, pour tout $x \geq 0$, on a

$\mathbb{P}(X \leq x) = \mathbb{P}(U \geq \mathrm{e}^{-x}) = 1 - \mathrm{e}^{-x}$, et on reconnait la fonction de répartition de la loi exponentielle de paramètre 1.

• Soit X une v.a. de loi exponentielle de paramètre q. On note $Y = [X]$ sa partie entière. On a donc

$$\mathbb{P}(Y = n) \;=\; \mathbb{P}(n \leq X < n+1) \;=\; \int_n^{n+1} q\mathrm{e}^{-qx} dx = e^{-qn}(1 - \mathrm{e}^{-q})$$

et on voit donc que Y a pour loi une loi géométrique de paramètre e^{-q}.

• Soit X une v.a. normale standard, i.e. de loi $\mathcal{N}(0,1)$. Vérifier que $\sigma X + m = Y$ suit une loi de Gauss centrée en m et de variance σ^2, i.e. $\mathcal{N}(m, \sigma^2)$.

La plupart des ordinateurs permettent de générer des variables aléatoires de loi uniforme sur $[0,1]$, c'est-à-dire que l'ordinateur fournit à l'utilisateur un nombre réel aléatoire X tel que $X \in [0,x]$ avec probabilité x pour $x \in [0,1]$. La transformation précédente permet de générer des variables aléatoires de loi arbitraires à partir d'une variable X suivant une loi uniforme sur $[0,1]$.

Proposition. *Soit $F : \mathbb{R} \to [0,1]$ la fonction de répartition associée à une loi de probabilité μ sur $\mathbb{R}$, i.e. $F(x) = \mu(]-\infty, x])$. Soit $F^{-1} : [0,1] \to \mathbb{R}$ l'inverse continu à droite de F, c'est-à-dire*

$$F^{-1}(x) \;=\; \inf\{y : F(y) > x\}\,, \qquad x \in [0,1]\,.$$

Si X suit la loi uniforme sur $[0,1]$, alors la variable aléatoire $Y = F^{-1}(X)$ suit la loi μ.

Preuve: Il s'agit de vérifier que Y a pour fonction de répartition F. On a pour tout $x \in \mathbb{R}$

$$Y \leq x \iff F^{-1}(X) \leq x \iff \inf\{y : F(y) > X\} \leq x\,. \qquad (\dagger)$$

D'une part, la dernière assertion de $(\dagger)$ est satisfaite dès que $F(x) > X$, autrement dit on a $\{X < F(x)\} \subseteq \{Y \leq x\}$. Comme X suit la loi uniforme, il en découle que

$$F(x) \;=\; \mathbb{P}(X < F(x)) \;\leq \mathbb{P}\,(Y \leq x)\;.$$

D'autre part, si la dernière assertion de $(\dagger)$ est vérifiée, alors $F(y) > X$ pour tout $y > x$, c'est-à-dire $\{Y \leq x\} \subseteq \{X < F(y)\}$. On a donc $\mathbb{P}\,(Y \leq x) \leq \mathbb{P}\,(X \leq F(y)) = F(y)$ (puisque X suit la loi uniforme). On fait tendre y vers x, et comme F est continue à droite, on obtient

$$\mathbb{P}\,(Y \leq x) \;\leq\; F(x)\,.$$

En conclusion, on a donc vérifié que F est la fonction de répartition de Y. ■

Cette méthode permet de simuler un grand nombre de variables aléatoires, mais pas toutes. Par exemple, on ne peut pas simuler ainsi une variable gaussienne, car on ne connait pas explicitement la fonction de répartition de la loi de Gauss, et a fortiori pas son inverse. On verra dans la section suivante comment contourner cette difficulté.

3.3 Indépendance

Définition. *Soient X et X' deux variables aléatoires à valeurs dans E et E', respectivement. On dit que X et X' sont indépendants si pour tout $A \in \mathcal{E}$ et tout $A' \in \mathcal{E}'$:*

$$\mathbb{P}\left(X \in A\,,\, X' \in A'\right) \,=\, \mathbb{P}(X \in A)\mathbb{P}(X' \in A')\,.$$

Quand on exprime l'indépendance en termes des distributions, on obtient que deux v.a. X et X' sont indépendantes si et seulement si la loi du couple $(X, X',\ P_{X,X'}$ vérifie

$$P_{X,X'}(A \times A') \,=\, P_X(A)P_{X'}(A')\,, \qquad \forall A \in \mathcal{E}, A' \in \mathcal{E}'\,.$$

Autrement dit, la loi $P_{X,X'}$ sur $E \times E'$ est la loi produit $P_X \otimes P_{X'}$. En particulier, étant données deux lois de probabilites μ et μ' sur E et E', il est facile de construire un espace de probabilité et deux v.a. X et X' indépendantes et de lois respectives μ et μ'. Plus précisément, on prend l'espace produit $E \times E'$, qu'on muni de la tribu produit $\mathcal{E} \otimes \mathcal{E}'$, et de la mesure produit $\mu \otimes \mu'$. Ensuite on prend pour X la première projection $X : E \times E' \to E$, et pour $X' : E \times E' \to E'$ la seconde projection.

On a immédiatement

Proposition *Pour que deux v.a. X et X' soient indépendantes, il faut et il suffit que*

$$\mathbb{E}\left(f(X)g(X')\right) \,=\, \mathbb{E}(f(X))\mathbb{E}(g(X'))$$

pour toute fonction mesurable bornée $f : E \to \mathbb{R}$ et toute fonction mesurable bornée $g : E' \to \mathbb{R}$.

Preuve: Si X et X' sont indépendants, alors on a bien

$$\mathbb{E}\left(f(X)g(X')\right) \,=\, \mathbb{E}(f(X))\mathbb{E}(g(X'))$$

lorsque $f = \mathbf{1}_A$ et $g = \mathbf{1}_B$. Par linéarité, la formule reste valable pour f et g étagées. Le cas général en découle par approximation. ■

Exercice: Vérifier que si X et X' sont indépendantes, alors il en est de même pour $f(X)$ et $g(X')$ pour toutes fonctions mesurables f et g.

Plus généralement, on dira que n+1 variables aléatoires $X_1, \cdots, X_{n+1}$ sont indépendantes si les n premières variables $X_1, \cdots, X_n$ sont indépendantes et si les variables $(X_1, \cdot, X_n)$ et X_{n+1} le sont également. Il est très facile de vérifier la propriété suivante:

Proposition *Pour que n v.a. $X_1, \cdots, X_n$ soient indépendantes, il faut et il suffit que*

$$\mathbb{E}\left(f_1(X_1) \cdots f_n(X_n)\right) \,=\, \mathbb{E}(f_1(X_1)) \cdots \mathbb{E}(f_n(X_n))$$

pour toutes les fonctions mesurables bornées $f_k : E_k \to \mathbb{R}$.

Voyons maintenant comment la notion d'indépendance permet la simulation de, non pas une, mais deux variables gaussiennes indépendantes. Pour cela, on commence par

simuler une variable U de loi uniforme sur $[0,1]$, et une variable S de loi exponentielle de paramètre $1/2$, indépendante de U. On pose ensuite

$$X \;=\; \sqrt{S}\cos(2\pi U) \quad , \quad Y \;=\; \sqrt{S}\sin(2\pi U)\,.$$

On peut vérifier alors par un changement de variable en coordonnées polaires que X et Y sont deux variables indépendantes, toutes les deux de loi $\mathcal{N}(0,1)$. En effet, soit $f : \mathbb{R}^2 \to \mathbb{R}$ une application mesurable bornée. On a

$$\begin{aligned}
\mathbb{E}(f(X,Y)) &= \frac{1}{2}\int_0^\infty ds \int_0^1 du\, \mathrm{e}^{-s/2} f\left(\sqrt{s}\cos(2\pi u), \sqrt{s}\sin(2\pi u)\right) \\
&= \int_0^\infty dr \int_0^1 du\, \mathrm{e}^{-r^2/2} r f\left(r\cos(2\pi u), r\sin(2\pi u)\right) \\
&= \frac{1}{2\pi}\int_{\mathbb{R}^2} \mathrm{e}^{-(x^2+y^2)/2} f\left(x,y\right)\, dxdy\,,
\end{aligned}$$

où la première égalité découle du changement de variable $s = r^2$ et la seconde du passage en coordonnées polaires. On voit maintenant que la loi du couple (X,Y) est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^2$, avec pour densité

$$\frac{1}{\sqrt{2\pi}}\exp\left\{-x^2/2\right\} \;\times\; \frac{1}{\sqrt{2\pi}}\exp\left\{-y^2/2\right\}\,, \qquad (x,y)\in\mathbb{R}^2\,.$$

Chapitre 4

Chaines de Markov sur un espace fini et matrices de transitions

4.1 Préliminaires

Un processus stochastique est une suite de variables aléatoires $X_0, X_1, \cdots, X_n, \cdots$ qui décrit l'évolution d'un phénomène aléatoire. On travaille en temps discret (indexé par les entiers) et on supposera de plus que l'espace d'état E dans lequel le processus prend ses valeurs est fini. Les processus stochastiques interviennent de façon très naturelle dans les applications.

Exemples. • Météorologie: X_n = température en un lieu donné au jour n, ou encore hauteur des précipitations pour l'année n.

• Bourse: X_n = cours d'une action le jour n, ou chiffre d'affaire d'une société l'année n.

• Assurance: X_n est le montant total des indemnités versées par une compagnie d'assurance pour des sinistres survenus le mois n.

• Epidémiologie: X_n = nombre d'individus infectés par une maladie contagieuse au bout de n jours.

• Jeux: X_n = ordre des cartes d'un jeu qui a été battu n fois.

Afin de décrire l'évolution du processus, on a besoin de la notion de probabilités conditionnelles.

Définition. *Soit A un événement de probabilité non nulle. Pour tout événement B, on note*

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

et on appelle cette quantité la probabilité de B sachant A, ou conditionnellement à A.

Il est immédiat de vérifier que l'application $\mathcal{F} \ni B \to \mathbb{P}(B \mid A)$ est une mesure de probabilité sur Ω. Nous nous intéressons au cas où $A = \{X_0 = x_0, \cdots, X_n = x_n\}$ et $B = \{X_{n+1} = x_{n+1}\}$.

De façon informelle, lorsque l'évolution d'un processus stochastique après une date n, ne dépend du passé $X_0, \cdots, X_n$ qu'à travers sa position au temps n (et non pas

du trajet qu'il a suivi pour atteindre cet état), on dit qu'il vérifie la propriété de Markov. Ce phénomène apparait souvent. Par exemple, si on bat un jeu de cartes suivant toujours la même technique, la distribution aléatoire des cartes au rang $n+1$ sachant les ordres aux rangs $1, \cdots, n$, ne dépend que l'ordre au rang n et pas des précédents. On se gardera de croire que la propriété de Markov est le lot de tous processus stochastiques.

Voici la définition précise.

Définition. *On dit que $X_0, \cdots, X_n, \cdots$ est une chaine de Markov (homogène) s'il existe des nombres réels positifs $(P_{ij})_{i,j\in E}$ avec $\sum_{j\in E} P_{ij} = 1$ pour tout $i \in E$, tels que pour tout entier n, tous $i, j, x_0, \cdots, x_n \in E$:*

$$\mathbb{P}\left(X_{n+1} = j \mid X_0 = x_0, \cdots, X_{n-1} = x_{n-1}, X_n = i\right) \,=\, P_{ij}\,,$$

pourvu bien sûr que le terme de droite soit bien défini. On appelle $(P_{ij})_{j\in E}$ la probabilité de transition pour l'état i. Lorsque l'espace d'état E est fini, $(P_{ij})_{i,j\in E}$ est une matrice qu'on appelle la matrice de transition de la chaine.

Quand on se donne des probabilités de transitions $(P_{ij})_{i,j\in E}$, il est facile de simuler sur ordinateur une chaine ayant ces probabilités de transitions. Pour tout entier n et tout $i \in E$, on simule par $\mathrm{Random}(n, i)$ des v.a. indépendantes de loi $(P_{ij})_{j\in E}$. On prend $X_1 = \mathrm{Random}(1, i)$. Connaissant $X_1 = x_1$, on prend ensuite $X_2 = \mathrm{Random}(2, x_1)$, puis on recommence, etc...

Voyons maintenant des exemples concrets.

Exemples. • Chaine à deux états: Considérons l'état d'une ligne de téléphone $X_n = 0$ si la ligne est libre à l'instant n, et $X_n = 1$ si la ligne est occupée. Supposons que sur chaque intervalle de temps, il y a une probabilité p qu'un appel arrive (un appel au plus). Si la ligne est déjà occupée, l'appel est perdu. Supposons également que si la ligne est occupée au temps n, il y a une probabilité q qu'elle se libère au temps $n+1$. On peut modèliser ainsi une chaine de Markov à valeurs dans $E = \{0, 1\}$, avec matrice de transition

$$P \,=\, \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}\,.$$

• File d'attente simple: On modifie l'exemple précédent en supposant qu'on peut mettre un appel en attente. Les appels arrivent et la ligne se libère comme avant. Si un appel arrive pendant que la ligne est occupée et si le système n'est pas saturé, l'appel est mis en attente. Si un appel arrive alors qu'il y a déjà un en attente, il est perdu. Cette fois l'espace d'état est $E = \{0, 1, 2\}$. On a

$$p(0,0) = 1-p\,, \quad p(0,1) \,=\, p\,, \quad p(0,2) = 0\,.$$

De même

$$p(2,0) = 0\,, \quad p(2,1) = q\,, \quad p(2,2) \,=\, 1-q\,.$$

Le cas où il y a exactement un appel retenu au temps n est un peu plus délicat. On a $p(1,0) = q(1-p)$ (l'appel se termine et pas d'appel nouveau arrive) et $p(1,2) = p(1-q)$ (un appel nouveau arrive et celui en cours continue). Comme la somme

$p(1,0)+p(1,1)+p(1,2)$ doit valoir 1, on a donc $p(1,1) = 1-q(1-p)-p(1-q)$. Finalement, la matrice de transition de la chaine est

$$\begin{pmatrix} 1-p & p & 0 \\ q(1-p) & 1-q(1-p)-p(1-q) & p(1-q) \\ 0 & q & 1-q \end{pmatrix}.$$

On supposera désormais que $X_0, X_1, \cdots$ est une chaine de Markov avec pour probabilités de transition $(P_{ij})_{i,j\in E}$. On décrit tout d'abord l'état de la chaine.

Proposition. *Pour $x_0, \cdots, x_n$ dans E, on a*

$$\mathbb{P}\left(X_1 = x_1, \cdots, X_n = x_n \mid X_0 = x_0\right) \;=\; P_{x_0,x_1} \cdots P_{x_{n-1},x_n}\,.$$

En particulier, si l'espace d'état est fini, disons $E = \{1, \cdots, N\}$,

$$\mathbb{P}\left(X_n = j \mid X_0 \,=\, i\right) \;=\; P_{ij}^n\,,$$

où $P^n = P \times \cdots \times P$ au sens des produits de matrices et $P^n = \left(P_{ij}^n\right)_{i,j\in E}$.

Preuve: La première formule s'obtient par récurrence. Plus précisément, on a

$$\begin{aligned} & \mathbb{P}\left(X_0 = x_0, \cdots, X_n = x_n \mid X_0 = x_0\right) \\ = \;& \mathbb{P}\left(X_1 = x_1, \cdots, X_{n-1} = x_{n-1} \mid X_0 = x_0\right) \mathbb{P}\left(X_n = x_n \mid X_0 = x_0, \cdots, X_{n-1} = x_{n-1}\right) \\ = \;& \mathbb{P}\left(X_1 = x_1, \cdots, X_{n-1} = x_{n-1} \mid X_0 = x_0\right) P_{x_{n-1},x_n} \\ = \;& P_{x_0,x_1} \cdots P_{x_{n-1},x_n} \end{aligned}$$

La seconde formule en découle également part un calcul simple de récurrence. ■

Une application importante de la proposition est la suivante: fixons des entiers n et k, et soient $x_0, \cdots, x_n, \cdots, x_{n+k}$ des points de E. On a alors

$$\begin{aligned} & \mathbb{P}\left(X_{n+1} = x_{n+1}, \cdots, X_{n+k} = x_{n+k} \mid X_0 = x_0, \cdots, X_n = x_n\right) \\ = \;& \frac{\mathbb{P}\left(X_0 = x_0, \cdots, X_n = x_n, X_{n+1} = x_{n+1}, \cdots, X_{n+k} = x_{n+k}\right)}{\mathbb{P}\left(X_0 = x_0, \cdots, X_n = x_n\right)} \\ = \;& \frac{P_{x_0,x_1} \cdots P_{x_{n-1},x_n} P_{x_n,x_{n+1}} \cdots P_{x_{n+k-1},x_{n+k}}}{P_{x_0,x_1} \cdots P_{x_{n-1},x_n}} \\ = \;& P_{x_n,x_{n+1}} \cdots P_{x_{n+k-1},x_{n+k}} \\ = \;& \mathbb{P}\left(X_1 = x_{n+1}, \cdots, X_k = x_{n+k} \mid X_0 = x_n\right)\,. \end{aligned}$$

En mots, si on sait qu'à l'instant n la chaine est en x_n, alors la chaine $X_0' = X_n, \cdots, X_k' = X_{n+k}, \cdots$, obtenue par translation temporelle, ne dépend pas de la trajectoire suivie pour arriver en x_n au temps n, et a même loi que la chaine initiale quand elle est issue de x_n. C'est une propriété conforme à l'intuition qu'on peut avoir.

4.2 Comportement asymptotique des chaines de Markov

On rappelle que P désigne la matrice de transition d'une chaîne de Markov. On s'intéresse à la distribution de X_n quand $n \to \infty$, ce qui revient à étudier la suite de matrices P^n quand $n \to \infty$.

Considèrons d'abord l'exemple simple de la chaine à deux états

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix},$$

avec $0 < p, q < 1$ Lorsque les réels p et q sont connus, on peut élever P à la puissance n en utilisant un ordinateur. Si on ne dispose pas d'ordinateur, ou si on veut traiter le cas général, on peut diagonaliser P. Les valeurs propres sont 1 et $1-p-q$, et on diagonalise $D = Q^{-1}PQ$ avec

$$Q = \begin{pmatrix} 1 & -p \\ 1 & q \end{pmatrix} \quad , \quad Q^{-1} = \begin{pmatrix} q/(p+q) & p/(p+q) \\ -1/(p+q) & 1/(p+q) \end{pmatrix} \quad ,$$

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 1-p-q \end{pmatrix}.$$

La diagonale de D est constituée des valeurs propres. Les colonnes de Q sont les vecteurs propres à droite de P, et les lignes de Q^{-1} les vecteurs propres à gauches. Les vecteurs propres sont uniques à une constante multiplicative près; on a choisi la constante pour la valeur propre 1 de sorte que la première ligne de Q^{-1} soit un vecteur probabilité (on verra plus tard pourquoi). On a alors

$$P^n = \left(QDQ^{-1}\right)^n = QD^nQ^{-1}$$

et un calcul simple donne

$$P^n = \begin{pmatrix} (q+p(1-p-q)^n)/(p+q) & (p-p(1-p-q)^n)/(p+q) \\ (q-q(1-p-q)^n)/(p+q) & (p+q(1-p-q)^n)/(p+q) \end{pmatrix}.$$

Comme $|1-p-q| < 1$, on voit que

$$\lim_{n\to\infty} P^n = \begin{pmatrix} q/(p+q) & p/(p+q) \\ q/(p+q) & p/(p+q) \end{pmatrix}.$$

Le fait que la seconde valeur propre, $1-p-q$ soit en module strictement inférieure à 1 est primordial dans le calcul; la matrice limite est constituée de deux vecteurs lignes identiques, qui est le vecteur propre à gauche de P, normalisé de sorte à être une probabilité.

Supposons maintenant que l'état initial de la chaine est donné par une probabilité $\mathbb{P}(X_0 = 0) = a$, $\mathbb{P}(X_0 = 1) = 1-a$. Nous avons vu que l'état de la chaine au temps n suit la loi caractérisée par

$$\mathbb{P}(X_n = j \mid X_0 = i) = P_{ij}^n \qquad i, j = 0 \text{ ou } 1$$

et on déduit que quand $n \to \infty$, quelque soit l'état initial i de la chaine,

$$\lim_{n\to\infty} \mathbb{P}(X_n = 0 \mid X_0 = i) = \frac{q}{p+q} \quad , \quad \lim_{n\to\infty} \mathbb{P}(X_n = 1 \mid X_0 = i) = \frac{p}{p+q}.$$

On notera que la limite en loi qu'on a obtenue est donnée par le vecteur propre associé à la valeur propre 1, normalisé pour être une probabilité.

Définition. *Un vecteur probabilité $v = (v_1, \cdots, v_N)$ sur E (i.e. les coefficients de v sont positifs et de somme 1) est dit probabilité invariante, ou probabilité d'équilibre, pour la chaine de matrice de transition P si v est un vecteur propre à gauche de valeur propre 1, i.e. $v = vP$.*

Pour justifier la terminologie, on voit que si v est une probabilité invariante et si l'état intial de la chaine X_0 suit la loi v, alors X_n suit également la loi v pour tout n (puisque la loi de X_n est $vP^n = v$).

Il existe toujours une probabilité invariante (quand on travaille avec des chaines de Markov sur un espace d'état fini). En effet; la condition $\sum_j P_{ij} = 1$ montre que le vecteur $1 = (1, \cdots, 1)$ est vecteur propre à droite associé à la valeur propre 1. Donc 1 admet un vecteur propre à gauche, et on peut montrer qu'il est toujours possible de choisir le vecteur propre à gauche avec des coefficients positifs, et donc de le normaliser pour en faire un vecteur probabilité. Les questions naturelles qui se posent alors sont de savoir si cette probabilité invariante est unique, puis si quelque soit la distribution initiale de la chaine, l'état converge quand le temps tend vers l'infini, vers cette probabilité invariante.

On peut observer que si, pour une distribution initiale donnée $m = (m_1, \cdots, m_N)$, X_n converge en loi vers un vecteur probabilité v, autrement dit $\lim_{n\to\infty} mP^n = v$, alors $vP = v$, et v est nécessairement une probabilité invariante.

Le théorème de Perron-Frobenius permet de répondre à cette question pour une très grande famille de matrices. Nous allons énoncer le résultat sans le démontrer.

Théorème. (Perron-Frobenius) *Soit (P_{ij}) une matrice $N \times N$ dont les coefficients sont strictement positifs. Alors la plus grande valeur propre (en module) λ, est > 0 et simple, et tout vecteur propre correspondant a tous ses coefficients de même signe.*

Sous les hypothèses du théorème de Perron Frobenius, on peut écrire P sous la forme $P = QDQ^{-1}$, où

$$D = \begin{pmatrix} \lambda & 0 \\ 0 & M \end{pmatrix},$$

avec M matrice $(N-1) \times (N-1)$ dont la plus grande valeur propre est strictement inférieure à λ. Cette observation entraine le résultat suivant pour les chaines de Markov.

Corollaire. *Soit $X_0, \cdots, X_n, \cdots$ une chaine de Markov sur un espace à d éléments, avec pour matrice de transition P. On suppose que les coefficients de P sont tous strictement positifs. Alors il existe une unique probabilité invariante, $v = (v_1, \ldots, v_d)$, et quelque que soit la distribution initiale de la chaine, X_n converge en loi vers v quand n tend vers ∞, c'est-à-dire qu'on a pour tous les états i, j*

$$\lim_{n\to\infty} \mathbb{P}(X_n = j \mid X_0 = i) = v_j.$$

Preuve: On applique le théorème de Perron-Frobenius. Il est facile de voir que la plus grande valeur propre de P (dont on sait qu'elle est réelle positive) vaut nécessairement 1. Le vecteur propre à droite a tous ses coefficients égaux à 1, le vecteur propre à gauche $v = (v_1, \cdots, v_N)$ est normalisé de sorte à être un vecteur probabilité.

Comme la plus grande valeur propre λ' de la matrice $(N-1)\times(N-1)$ M vérifie $|\lambda'| < 1$, on a en particulier $\lim_{n\to\infty} M^n = 0$ (plus précisément, la vitesse de convergence est en λ'^n).

En conséquence, nous avons

$$\lim_{n\to\infty} P^n = Q \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} = \begin{pmatrix} v \\ \vdots \\ v \end{pmatrix} .$$

La convergence en loi vers v en découle, ainsi que l'unicité de la probabilité invariante. ■

D'une façon un peu plus générale, on a l'extension suivante.

Corollaire. *Soit $X_0, \cdots, X_n, \cdots$ une chaine de Markov sur un espace fini, avec pour matrice de transition P. On suppose qu'il existe un entier k pour lequel les coefficients de P^k sont tous strictement positifs. Alors il existe une unique probabilité invariante, v, et quelque que soit la distribution initiale de la chaine, X_n converge en loi vers v quand n tend vers ∞.*

Preuve: Considérons une loi initiale m, fixons $r \in \{0, \cdots, k\}$ et notons $\mu = mP^r$. On a $mP^{kd+r} = \mu P^{kd}$, et d'après le premier corollaire, mP^{kd+r} converge vers v quand d tend vers ∞. La limite ne dépend pas de r, donc quelque soit la distribution initiale, X_n converge en loi vers le vecteur propre probabilité (à gauche) associé à la valeur propre 1 de P^k. Ceci assure l'unicité de la probabilité invariante. ■

On se gardera de croire que le résultat précédent est valable en toute généralité. Il est facile de construire des chaines de Markov possèdant plusieurs probabilités invariantes.

Chapitre 5

Suites et Séries de Variables Aléatoires

On considère une suite $X_1, \cdots, X_n, \cdots$ de v.a. à valeurs réelles et on s'intéresse au comportement asymptotique de cette suite aléatoire. L'étude repose en partie sur un résultat élémentaire très utile.

5.1 Le lemme de Borel-Cantelli

On considère une suite d'événements $\{\Lambda_n, n \in \mathbb{N}\}$ dans Ω. La suite des événements $\bigcup_{k\geq n} \Lambda_k$, $n \in \mathbb{N}$ est décroissante, son intersection

$$\Lambda = \limsup_{n\to\infty} \Lambda_n = \bigcap_{n\geq 0} \bigcup_{k\geq n} \Lambda_k$$

est encore un événement dans $\mathcal{F}$. On remarquera que Λ représente l'ensemble des aléas ω qui appartiennent à une infinité d'événements Λ_n, autrement dit $\mathbf{1}_\Lambda = \limsup_{n\to\infty} \mathbf{1}_{\Lambda_n}$; ce qui justifie la notation.

Lemme de Borel-Cantelli. (partie directe) *Si la série* $\sum_{n=0}^{\infty} \mathbb{P}(\Lambda_n)$ *converge, alors* $\mathbb{P}(\Lambda) = 0$.

Preuve: Fixons $\varepsilon > 0$. Il existe un entier N tel que $\sum_{n=N}^{\infty} \mathbb{P}(\Lambda_n) < \varepsilon$, et en conséquence $\mathbb{P}\left(\bigcup_{n\geq N} \Lambda_n\right) < \varepsilon$. De la définition de Λ, on tire *a fortiori* $\mathbb{P}(\Lambda) < \varepsilon$. ■

Il y a une réciproque partielle à la partie directe sous une hypothèse d'indépendance entre les événements.

Définition. *On dit que les événements* $\Lambda_n, n \in \mathbb{N}$ *sont indépendants si pour tout entier* N, *les v.a.* $\mathbf{1}_{\Lambda_1}, \cdots, \mathbf{1}_{\Lambda_N}$ *sont indépendantes.*

En particulier, si les événements $\Lambda_n, n \in \mathbb{N}$ sont indépendants, alors pour toute famille

finie d'entiers $\{n(i), i \in I\}$, on a

$$\mathbb{P}\left(\bigcap_I \Lambda_{n(i)}\right) \;=\; \prod_I \mathbb{P}\left(\Lambda_{n(i)}\right) .$$

Lemme de Borel-Cantelli. (partie réciproque) *Si les événements $\Lambda_n, n \in \mathbb{N}$ sont indépendants et si la série $\sum_{n=0}^{\infty} \mathbb{P}(\Lambda_n)$ diverge, alors $\mathbb{P}(\Lambda) = 1$.*

Preuve: Comme les Λ_n sont indépendants, il en est de même pour leurs complémentaires, et on a

$$\mathbb{P}\left(\bigcup_{j=k}^{m} \Lambda_j\right) \;=\; 1 - \mathbb{P}\left(\bigcap_{j=k}^{m} \Lambda_j^c\right) \;=\; 1 - \prod_{j=k}^{m} \mathbb{P}(\Lambda_j^c) \;\geq\; 1 - \exp\left\{-\sum_{j=k}^{m} \mathbb{P}(\Lambda_j)\right\}.$$

On fait tendre m vers ∞, et on tire

$$\mathbb{P}\left(\bigcup_{j=k}^{\infty} \Lambda_j\right) \;\geq\; 1 - \exp\left\{-\sum_{j=k}^{\infty} \mathbb{P}(\Lambda_j)\right\} \;=\; 1\,.$$

Comme les événements $\bigcup_{j=k}^{\infty} \Lambda_j$ décroissent vers Λ quand $k \to \infty$, on a bien $\mathbb{P}(\Lambda) \geq 1$ (et donc $= 1$). ■

5.2 Divers modes de convergence

Définition. (convergence p.s.) *On dit que la suite $(X_n, n \in \mathbb{N})$ converge presque sûrement vers une v.a. X s'il existe un événement Λ avec $\mathbb{P}(\Lambda) = 1$, tel que*

$$\lim_{n\to\infty} X_n(\omega) \;=\; X(\omega) \qquad \textit{pour tout } \omega \in \Lambda\,.$$

Définition. (convergence en probabilité) *On dit que la suite $(X_n, n \in \mathbb{N})$ converge en probabilité vers une v.a. X si pour tout $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| > \varepsilon\right) \;=\; 0\,.$$

Il est facile de voir que les résultats usuels sur les limites (unicité de la limite, linéarité ...) sont valables dans les deux cas.

Proposition. *Si $X_n \to X$ p.s., alors la convergence a lieu également en probabilité.*

Preuve: Soit Λ l'événement de probabilité 1 qui apparait dans la définition. On fixe $\varepsilon > 0$, et pour chaque entier n, on considère l'événement

$$\Gamma_n \;=\; \{\omega \in \Lambda : \sup_{k\geq n} |X_k(\omega) - X(\omega)| > \varepsilon\}\,.$$

La suite des Γ_n est décroissante et $\bigcap \Gamma_n = \emptyset$ (puisque X_n converge vers X sur Λ). On a donc $\lim_{n\to\infty} \mathbb{P}(\Gamma_n) = 0$. Comme

$$\{|X_n - X| > \varepsilon\} \subseteq \Gamma_n \cup \Lambda^c,$$

on a *a fortiori* $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$. ■

La réciproque est fausse; voici un contre exemple: On prend $\Omega = [1, 2]$ qu'on munit de la mesure de Lebesgue. Pour chaque n, on note k l'unique entier tel que $2^k \leq n < 2^{k+1}$, et on prend $X_n(t) = \mathbf{1}_{[n2^{-k},(n+1)2^{-k}]}$. Il est clair que pour tout $\varepsilon \in]0, 1[$, $\mathbb{P}(|X_n| > \varepsilon) = 2^{-k}$, de sorte que $X_n \to 0$ en probabilité. Néanmoins, pour tout $t \in [0, 1]$, il existe une infinité d'entiers n pour lesquels $X_n(t) = 1$, et $X_n(t)$ ne converge pas vers 0.

En revanche on a une réciproque partielle.

Proposition. *Si $X_n \to X$ en probabilité, alors il existe une suite extraite $X_{N(n)}$ qui converge vers X p.s.*

Preuve: Comme $\mathbb{P}(|X - X_n| > \varepsilon) \to 0$ pour tout $\varepsilon > 0$, on peut trouver pour tout $n \geq 1$ un entier $N(n)$ tel que $\mathbb{P}\left(|X - X_{N(n)}| > 1/n\right) \leq 2^{-n}$. La série $\sum \mathbb{P}\left(|X - X_{N(n)}| > 1/n\right)$ converge donc. D'après le lemme de Borel-Cantelli, la probabilité de l'événement

$$\bigcap_{n\geq 0} \bigcup_{k\geq n} \{|X - X_{N(n)}| > 1/n\}$$

est nulle. Si Λ désigne l'événement complémentaire, on a donc $\mathbb{P}(\Lambda) = 1$, et (par définition de Λ) pour tout $\omega \in \Lambda$, il existe un entier $k(\omega)$ tel que

$$|X(\omega) - X_{N(n)}(\omega)| < 1/n \qquad \text{pour tout } n \geq k(\omega).$$

Ceci montre que $X_{N(n)}(\omega) \to X(\omega)$. ■

En théorie de l'intégration, on a vu d'autres modes de convergence qu'on rappelle.

Définition. (convergence dans L^p) *Pour tout $p \geq 1$, on dit qu'une suite $(X_n, n \in \mathbb{N})$ de v.a. converge dans $L^p(\Omega, \mathbb{P})$ vers une v.a. $X \in L^p(\Omega, \mathbb{P})$ si*

$$\lim_{n\to\infty} \mathbb{E}(|X - X_n|^p) = 0.$$

Comparons cette notion avec les précédentes.

Proposition. *Si $X_n \to X$ dans L^p, alors la convergence a encore lieu en probabilité. Réciproquement, si $X_n \to X$ en probabilité et si il existe une v.a. réelle $Y \in L^p$ telle que $|X_n| \leq Y$ pour tout n, alors $X_n \to X$ dans L^p.*

Preuve: Fixons $\varepsilon > 0$ et $\eta > 0$. On sait qu'on peut trouver un entier n_0 tel que $\mathbb{E}(|X - X_n|^p) \leq \eta\varepsilon^p$ dès que $n \geq n_0$. En appliquant l'inégalité de Markov, on obtient dans ce cas

$$\mathbb{P}(|X - X_n| > \varepsilon) \leq \varepsilon^{-p}\mathbb{E}(|X - X_n|^p) \leq \eta.$$

Réciproquement, si la suite $(X_n, n \in \mathbb{N})$ converge vers X en probabilité, c'est encore le cas pour toute suite extraite, disons $\left(X_{M(n)}, n \in \mathbb{N}\right)$. On sait qu'on peut extraire

de cette dernière une sous-suite, disons $\left(X_{N(n)}, n \in \mathbb{N}\right)$ qui converge p.s. vers X. Par hypothèse, on peut appliquer le théorème de convergence dominée, i.e. $X_{N(n)} \to X$ dans $L^p(\Omega, \mathbb{P})$. Ainsi, de toute suite extraite de $(X_n, n \in \mathbb{N})$, on a su extraire une sous-sous-suite qui converge vers X dans L^p. Donc $X_n \to X$ dans L^p. ■

Exercice: On se donne une suite de v.a. $(X_n, n \in \mathbb{N})$, suivant toutes la même loi.

(1) Montrer que X_n/n converge toujours vers 0 en probabilité.

(2) Montrer que si $X_1 \in L^1(\Omega, \mathbb{P})$, alors la suite X_n/n converge vers 0 p.s. et dans $L^1(\mathbb{P})$.

(3) On suppose maintenant que ces v.a. sont indépendantes, i.e. toute sous-suite finie est constituée de v.a. indépendantes. Montrer que si de plus $\mathbb{E}(|X_1|) = \infty$, alors $\limsup_{n\to\infty} n^{-1} X_n = \infty$ p.s.

Solution (à détailler):

(1) $\mathbb{P}(|X_n/n| > \varepsilon) = \mathbb{P}(|X_1| > n\varepsilon) \to 0$.

(2) La convergence dans L^1 est évidente. Fixons un réel $k > 0$. On a en intégrant par parties

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n/n| \geq 1/k) = \sum_{n=1}^{\infty} \mathbb{P}(k|X_1| \geq n) \; = \; \mathbb{E}([k|X_1|]) < \infty \, .$$

On applique la partie directe de Borel-Cantelli. L'événement

$$\Lambda_k = \{\limsup_{n\to\infty} n^{-1}|X_n| < 1/k\}$$

a pour probabilité 1, il en est de même pour l'intersection de ces événements, i.e. $\limsup_{n\to\infty} n^{-1}|X_n| = 0$ p.s.

(3) Cette fois, $\sum_{n=1}^{\infty} \mathbb{P}(|X_n/n| > k) = \infty$ et on applique la partie réciproque de Borel-Cantelli.

5.3 Séries de variables aléatoires indépendantes $(*)$

On se donne une suite de v.a. réelles $(X_n, n \in \mathbb{N}^*)$ que l'on suppose indépendantes, c'est-à-dire que pour tout n, $X_1, \cdots, X_n$ sont des v.a. indépendantes. On note $S_n = X_1 + \cdots + X_n$ la somme partielle jusqu'à l'ordre n, et on s'intéresse à la convergence de la suite des v.a. S_n. Commençons par un exemple très simple.

Exemple: On suppose que $(X_n, n \in \mathbb{N})$ est une suite de v.a. Gaussiennes indépendantes, X_n de loi $\mathcal{N}(0, \sigma_n^2)$. Pour $m < n$, on sait que $S_n - S_m$ a pour loi $\mathcal{N}(0, \sum_{m+1}^{n} \sigma_k^2)$, et on vérifie aisément que si la série $\sum_1^{\infty} \sigma_n^2$ diverge, alors S_n ne converge pas probabilité.

Voyons la réciproque en supposant maintenant que la série $\sum_1^{\infty} \sigma_n^2$ converge. La suite $(X_n, n \in \mathbb{N})$ est une suite orthogonale dans $L^2(\Omega, \mathbb{P})$, ce qui montre que la suite de sommes partielles converge dans L^2, et donc en probabilité. Comme la somme partielle S_n a pour loi $\mathcal{N}(0, \sum_1^n \sigma_k^2)$, la v.a. limite suit une loi Gaussienne de loi $\mathcal{N}(0, \sum_1^{\infty} \sigma_k^2)$. En fait, les résultats qui vont suivrent montrent qu'on a également convergence p.s.

On commence par étudier le cas où les variables X_j sont *symétriques*, c'est-à-dire que

X_j et $-X_j$ ont la même loi. L'étude repose sur une inégalité très utile dûe à Paul Lévy.

Lemme. (Inégalité de Lévy) *Si les variables X_j sont symétriques, alors pour tout entier n et tout réel $x > 0$:*

$$\mathbb{P}\left(\max_{1\le j\le n} |S_j| > x\right) \le 2\mathbb{P}(|S_n| > x)\,.$$

Preuve: La démonstration repose sur un argument de réflexion. Pour chaque entier $j \ge 0$, on note $S^{(j)}$ la suite des sommes partielles associée à la suite de v.a. indépendantes $X_1, \cdots, X_j, -X_{j+1} \cdots$. L'hypothèse de symétrie garantie que les suites $(S_n, n \in \mathbb{N})$ et $(S_n^{(j)}, n \in \mathbb{N})$ ont la même loi.

Soit $N = \inf\{k : |S_k| > x\}$, avec la convention $\inf \emptyset = \infty$; ainsi $\{\max_{1\le k\le n} |S_k| > x\} = \{N \le n\}$. Il est évident que pour tout entier $j \le n$, les événements $\{N = j\}$ et $\{N^{(j)} = j\}$ coincident, où on a noté $N^{(j)} = \inf\{k : |S_k^{(j)}| > x\}$. En conséquence,

$$\mathbb{P}(|S_n| > x) \;=\; \sum_{j=1}^{n} \mathbb{P}(|S_n| > x, N = j) \;=\; \sum_{j=1}^{n} \mathbb{P}(|S_n^{(j)}| > x, N = j)\,.$$

D'autre part, on a $2S_j = S_n + S_n^{(j)}$ dès que $j \le n$, et d'après l'inégalité triangulaire, $2|S_j| \le |S_n| + |S_n^{(j)}|$. Il en découle que $\{|S_j| > x\} \subseteq \{|S_n| > x\} \cup \{|S_n^{(j)}| > x\}$. On a donc

$$\begin{aligned}
\mathbb{P}(N \le n) &= \sum_{j=1}^{n} \mathbb{P}(|S_j| > x, N = j) \\
&\le \sum_{j=1}^{n} \mathbb{P}(|S_n| > x, N = j) \;+\; \sum_{j=1}^{n} \mathbb{P}(|S_n^{(j)}| > x, N = j) \;=\; 2\mathbb{P}(|S_n| > x)\,.
\end{aligned}$$

L'inégalité de Lévy est donc établie. ■

Corollaire. *Si $(S_n, n \in \mathbb{N})$ est la suite des sommes partielles d'une suite de v.a. symétriques indépendantes, alors S_n converge p.s. si et seulement si S_n converge en probabilité.*

Preuve: Supposons que S_n converge en probabilité; on note S_∞ la limite. Pour tout entier $n > 0$, il existe un entier $N(n)$ tel que $\mathbb{P}(|S_\infty - S_{N(n)}| > 1/n) \le n^{-2}$. En appliquant l'inégalité de Lévy, on tire

$$\mathbb{P}\left(\sup_{k\ge N(n)} |S_\infty - S_{N(n)}| > 1/n\right) \le 2n^{-2}\,.$$

Ceci montre que

$$\sum_{1}^{\infty} \mathbb{P}\left(\sup_{k\ge N(n)} |S_\infty - S_{N(n)}| > 1/n\right) \;<\; \infty\,,$$

et d'après le lemme de Borel-Cantelli, $\sup_{k\ge N(n)} |S_\infty - S_{N(n)}| \le 1/n$ sauf pour au plus un nombre fini d'entiers n, p.s. En réfléchissant un peu, on voit que ceci prouve

que $\lim_{k\to\infty} |S_\infty - S_k| = 0$ p.s., autrement dit la convergence p.s. est établie. La réciproque est évidente. ■

Une technique classique pour passer d'une suite de v.a. indépendantes non-symétrique $(X_n, n \in \mathbb{N})$ à une suite de v.a. symétrique indépendantes consiste à munir chaque X_n d'un signe aléatoire ε_n, où $(\varepsilon_n, n \in \mathbb{N})$ est une suite dite de Rademacher, i.e. suite de v.a. indépendantes et équidistribuées sur $\{-1, +1\}$. Il est facile de vérifier que si les suites $(\varepsilon_n, n \in \mathbb{N})$ et $(X_n, n \in \mathbb{N})$ sont indépendantes, alors la suite $(\varepsilon_n X_n, n \in \mathbb{N})$ est une suite de v.a. symétriques indépendantes.

En fait, le résultat sur la somme de variables aléatoires symétriques indépendantes reste vrai sans l'hypothèse de symétrie, ce qui est a priori assez surprenant.

Théorème. *Si S_n converge en probabilité, alors S_n converge également p.s.*

La preuve de ce théorème difficile repose sur un lemme technique, qui va remplacer l'inégalité de Lévy dans le cas symétrique.

Lemme. *Soient $(U_n, n = 1, \cdots, N)$ et $(V_n, n = 1, \cdots, N)$, deux suites finies de v.a. positives telles que pour chaque $n \leq N$, V_n est indépendant de $(U_n, U_{n+1}, \cdots, U_N)$. On a alors pour tout $x > 0$*

$$\mathbb{P}\left(\max_{n=1,\cdots,N}(U_n - V_n) > x\right) \geq \left(\min_{n=1,\cdots,N}\mathbb{P}(V_n < x)\right) \times \mathbb{P}\left(\max_{n=1,\cdots,N} U_n > 2x\right).$$

Preuve: On observe tout d'abord que

$$\begin{aligned}
& \mathbb{P}\left(\max_{n=1,\cdots,N}(U_n - V_n) > x\right) \\
\geq\ & \mathbb{P}\left(\bigcup_{n=1,\cdots,N}\{V_n < x, U_n > 2x\}\right) \\
\geq\ & \mathbb{P}\left(\bigcup_{n=1,\cdots,N}\{V_n < x, U_n > 2x, U_{n+1} \leq 2x, \cdots, U_N \leq 2x\}\right) \\
=\ & \sum_{n=1}^{N} \mathbb{P}\left(V_n < x, U_n > 2x, U_{n+1} \leq 2x, \cdots, U_N \leq 2x\right).
\end{aligned}$$

Grâce à l'indépendance, on tire

$$\begin{aligned}
=\ & \sum_{n=1}^{N} \mathbb{P}\left(V_n < x\right) \mathbb{P}\left(U_n > 2x, U_{n+1} \leq 2x, \cdots, U_N \leq 2x\right) \\
\geq\ & \left(\min_{n=1,\cdots,N}\mathbb{P}(V_n < x)\right) \times \sum_{n=1}^{N} \mathbb{P}\left(U_n > 2x, U_{n+1} \leq 2x, \cdots, U_N \leq 2x\right) \\
=\ & \left(\min_{n=1,\cdots,N}\mathbb{P}(V_n < x)\right) \times \mathbb{P}\left(\bigcup_{n=1,\cdots,N}\{U_n > 2x, U_{n+1} \leq 2x, \cdots, U_N \leq 2x\}\right) \\
=\ & \left(\min_{n=1,\cdots,N}\mathbb{P}(V_n < x)\right) \times \mathbb{P}\left(\max_{n=1,\cdots,N} U_n > 2x\right).
\end{aligned}$$

■

Preuve du théorème: Soit S la limite des S_n (en probabilité). On sait qu'il existe une suite extraite $S_{A(n)}$ qui converge vers S p.s. (ici, $A : \mathbb{N} \to \mathbb{N}$ est une certaine application strictement croissante). Pour chaque entier k, il existe un unique entier q tel que $A(q-1) \leq k < A(q)$, et on pose $\alpha(k) = A(q-1)$. L'application α est croissante sur $\mathbb{N}$. On introduit

$$U_k = |S - S_k| \quad , \quad V_k = |S_{\alpha(k)} - S_k| \quad , \quad W_k = |S - S_{\alpha(k)}| \, .$$

Par construction, on sait que $W_k \to 0$ p.s. et $V_k \to 0$ en probabilité (critère de Cauchy). De plus, pour chaque k fixé, la v.a. V_k est indépendante de $(U_k, U_{k+1}, \cdots)$.

Fixons $x > 0$ arbitrairement petit. Puisque $W_k \geq U_k - V_k$, on a d'après le lemme précédent que pour tous les entiers M, N avec $M \leq N$

$$\begin{aligned} \mathbb{P}\left(\max_{n=M,\cdots,N} U_n > 2x \right) &\leq \mathbb{P}\left(\max_{n=M,\cdots,N} (U_n - V_n) > x \right) \left(\min_{n=M,\cdots,N} \mathbb{P}(V_n < x) \right)^{-1} \\ &\leq \mathbb{P}\left(\max_{n=M,\cdots,N} W_n > x \right) \left(\min_{n=M,\cdots,N} \mathbb{P}(V_n < x) \right)^{-1} . \end{aligned}$$

Fixons $\varepsilon > 0$. Puisque $V_k \to 0$ en probabilité, on peut trouver un entier M_0 tel que $\mathbb{P}(V_n < x) > 1 - \varepsilon$ dès que $n > M_0$.

Puisque $W_k \to 0$ p.s., on peut trouver un entier M_1 tel que $\mathbb{P}\left(\max_{n=M,\cdots,N} W_n > x\right) < \varepsilon$ pour tout $M_1 \leq M \leq N$.

En conséquence, pour $M \geq M_0 + M_1$

$$\mathbb{P}\left(\max_{n=M,\cdots,N} U_n > 2x \right) \leq \varepsilon(1-\varepsilon)^{-1} , \qquad \forall N \geq M \, ,$$

et donc également, en faisant tendre d'abord N vers ∞ puis M vers ∞

$$\mathbb{P}\left(\limsup_{n\to\infty} U_n > 2x \right) \leq \varepsilon(1-\varepsilon)^{-1} .$$

Comme ε et x sont arbitraires, on a bien $\limsup_{n\to\infty} U_n = 0$ p.s., ce qui termine la preuve du théorème. ■

Lemme. *Supposons que les v.a. X_n sont centrées et admettent toutes un moment d'ordre 2, i.e. $\mathbb{E}(X_n) = 0$ et $\mathbb{E}(X_n^2) < \infty$. Alors S_n converge dans $L^2(\Omega, \mathbb{P})$ si et seulement si la série des moyennes quadratiques converge, $\sum_{n=1}^{\infty} \mathbb{E}(X_n^2) < \infty$. Dans ce cas, la convergence a lieu p.s. également, et la limite, S_∞, est une v.a. centrée dont la variance vaut $\sum_{n=1}^{\infty} \mathbb{E}(X_n^2)$.*

Preuve: La suite $(X_n, n \in \mathbb{N})$ est une suite orthogonale dans L^2. La série converge dans L^2 ssi $\sum ||X_n||_2^2 < \infty$. Si c'est le cas, la convergence a encore lieu en probabilité; il ne reste qu'à appliquer le théorème précédent. ■

■

Exemples: • On prend $X_n = \varepsilon_n n^{-1}$, avec $(\varepsilon_n, n \in \mathbb{N})$ une suite de v.a. de Rademacher (suite de v.a. à valeurs dans $\{-1, 1\}$ équidistribuées et indépendantes). La série $\sum_{n=1}^{\infty} |X_n|$ diverge p.s. alors que la série $\sum_{n=1}^{\infty} X_n$ converge p.s. (semi-convergence).

• Plus généralement, on prend une suite $(Y_n, n \in \mathbb{N})$ de v.a. indépendantes, et une suite de Rademacher $(\varepsilon_n, n \in \mathbb{N})$ comme ci-dessus, indépendante de la suite $(Y_n, n \in \mathbb{N})$. Si $\sum_{n=1}^{\infty} \mathbb{E}(Y_n^2) < \infty$, alors la série $\sum_{n=1}^{\infty} \varepsilon_n Y_n$ converge p.s. et dans $L^2(\mathbb{P})$ (en effet, la suite des $X_n = \varepsilon_n Y_n$ est une suite de v.a. centrées indépendantes, et $\mathbb{E}(X_n^2) = \mathbb{E}(Y_n^2)$. Exercice: Montrer un résultat analogue quand on remplace la suite de Rademacher par une suite de v.a. normales indépendantes.

Théorème des trois séries (Kolmogorov). *Posons $Y_n = X_n$ si $|X_n| \leq 1$ et $Y_n = 0$ sinon. Pour que la suite S_n converge p.s., il faut et il suffit que les trois séries suivantes converges:*

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > 1) \quad , \quad \sum_{n=1}^{\infty} \mathbb{E}(Y_n) \quad , \quad \sum_{n=1}^{\infty} \mathrm{Var}(Y_n)\,.$$

Preuve: On ne montrera que la partie la plus facile (la plus intéressante en pratique, aussi).

Comme $\sum \mathbb{P}(Y_n \neq X_n) = \sum \mathbb{P}(|X_n| > 1) < \infty$, on sait d'après le lemme de Borel-Cantelli que p.s., $X_n(\omega) = Y_n(\omega)$ sauf pour au plus un nombre fini d'entiers. Les séries $\sum X_n(\omega)$ et $\sum Y_n(\omega)$ sont donc p.s. de même nature. La suite des v.a. $(Y_n - \mathbb{E}(Y_n), n \in \mathbb{N})$ est une suite de v.a. centrées indépendantes. On applique le lemme précédent, de sorte que la série $\sum (Y_n - \mathbb{E}(Y_n))$ converge p.s. Comme la série des moyennes $\sum \mathbb{E}(Y_n)$ converge également, la série $\sum Y_n$ converge p.s. ■

5.4 La loi des grands nombres

Pour ne pas passer trop de temps sur des questions techniques un peu délicates, on admettra que si P_X est une loi de probabilité sur $\mathbb{R}$, on peut construire un espace de probabilités adéquat et une suite de v.a. $X_1, \cdots, X_n, \cdots$ indépendantes et toutes de loi P_X, et on s'intéresse à la convergence de la suite au sens de Césaro. Il est un cas simple dans lequel une telle construction est explicite.

Construction de Lebesgue d'une suite de v.a. de Bernoulli independantes: On prend $(\Omega, \mathcal{F}) = ([0, 1), \mathcal{B})$ et $\mathbb{P}$ = mesure de Lebesgue. On écrit $\omega \in \Omega$ sous forme dyadique, $\omega = \sum_{n=1}^{\infty} X_n(\omega) 2^{-n}$ (suivant la convention usuelle, l'écriture où tous les $X_n(\omega)$ valent 1 à partir d'un certain rang est exclue).. Alors la suite $(X_n, n = 1, \cdots)$ est une suite de v.a. indépendantes, chacune suit une loi de Bernoulli de paramètre 1/2. En effet, si on se donne une suite finie $(x_1, \cdots, x_n)$ de 0 et 1 et si on pose $y = \sum_{j=1}^{n} x_j 2^{-j}$, alors $\{X_1 = x_1, \cdots, X_n = x_n\} = [y, y + 2^{-n}[$ et donc

$$\mathbb{P}\left(X_1 = x_1, \cdots, X_n = x_n\right) \,=\, 2^{-n} \,=\, \prod_{j=1}^{n} \mathbb{P}(X_j = x_j)\,.$$

La loi des grands nombres a été découverte par Jacques Bernoulli, précisément pour

une suite de v.a. de Bernoulli indépendantes. C'est l'un des résultats les plus importants de la théorie des probabilités.

Théorème (Loi des Grands Nombres) *(i) Si* $\mathbb{E}(|X|) < \infty$, $\mathbb{E}(X) = \mu$, *alors*

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} \;=\; \mu$$

presque sûrement.

(ii) Si $\mathbb{E}(|X|) = \infty$, *alors la suite* $(n^{-1}(X_1 + \cdots + X_n), n = 1, \cdots,)$ *diverge p.s.*

Preuve: (i) Nous ne démontrerons ici la loi que sous l'hypothèse plus restrictive que X admet un moment d'ordre 2. Nous reviendrons sur le cas général plus tard.

Tout d'abord, en considérant séparément la partie positive de X_n et sa partie négative, on remarque qu'il suffit de démontrer le résultat pour des v.a. positives, ce qu'on suppose désormais.

On note $S(n) = (X_1 + \cdots + X_n)$. On a $\mathbb{E}(S(n)) = n\mu$ et $\mathrm{Var}(S(n)) = n\mathrm{Var}(X)$. En particulier,

$$\mathrm{Var}(n^{-4}(S(n^4) - n^4\mu) = n^{-4}\mathrm{Var}(X)\,.$$

D'après l'inégalité de Bienaymé-Tchebychev,

$$\mathbb{P}(|n^{-4}S(n^4) - \mu| > 1/n) \;\le\; n^{-2}\mathrm{Var}(X)\,.$$

On déduit que la série $\sum_{n=1}^{\infty} \mathbb{P}(|n^{-4}S(n^4) - \mu| > 1/n)$ converge. D'après le lemme de Borel-Cantelli, on a donc presque-sûrement

$$|n^{-4}S(n^4) - \mu| \;\le\; 1/n \qquad \text{sauf pour un nombre fini d'entiers } n.$$

Fixons un aléas ω pour lequel l'événement ci-dessus est réalisé et prenons n suffisament grand. Pour tout entier k tel que $n^4 \le k \le (n+1)^4$, on a donc (puisque la suite $(S_j, j = 1, \cdots,)$ est croissante)

$$k^{-1}S(k) \;\le\; n^{-4}S((n+1)^4) \;=\; \left(\frac{n+1}{n}\right)^4 (n+1)^{-4}S((n+1)^4) \;\le \left(\frac{n+1}{n}\right)^4 (\mu + 1/n)\,.$$

Comme $\left(\frac{n+1}{n}\right)^4$ tend vers 1 quand $n \to \infty$, on a donc établi que pour cet aléas ω

$$\limsup_{k\to\infty} k^{-1}S(k) \;\le\; \mu\,.$$

De même, on a

$$k^{-1}S(k) \;\ge\; (n+1)^{-4}S(n^4) \;=\; \left(\frac{n}{n+1}\right)^4 (n)^{-4}S(n^4) \;\ge \left(\frac{n}{n+1}\right)^4 (\mu - 1/n)\,.$$

Comme $\left(\frac{n}{n+1}\right)^4$ tend vers 1 quand $n \to \infty$, on a donc établi que pour cet aléas ω

$$\limsup_{k\to\infty} k^{-1}S(k) \;\ge\; \mu\,.$$

En conclusion, on a bien pour presque tout ω: $\lim_{k\to\infty} k^{-1}S(k) \;=\; \mu\,.$

(ii) On a déjà vu dans le chapitre précédent que si les v.a. $X_1, \cdots, X_n, \cdots$ sont indépendantes et de même loi et que $\mathbb{E}(|X|) = \infty$, alors p.s., $\limsup_{n\to\infty} n^{-1}|X_n| = \infty$. Dans ces conditions, la suite des moyennes de Césaro ne peut pas converger. ■

On fait souvent référence au théorème comme la loi *forte* des grands nombres. Une conséquence immédiate du théorème est que sous les hypothèses (i), la suite $n^{-1}S(n)$ converge *en probabilités* vers la moyenne μ. On parle ce dernier cas de loi *faible* des grands nombres, par opposition à la loi forte.

La loi des grands nombres justifie l'interprétation fréquentielle de la notion de probabilité. Typiquement, on réalise un grand nombre d'expériences dans des conditions identiques, dont le résultat dépend de l'aléas (exemple, durée de vie d'une ampoule électrique > 1000 heures). On note $X_n = 1$ si la n-ième expérience réussit, $X_n = 0$ sinon. La suite $X_1, \cdots, X_n, \cdots$ est une suite de v.a. de Bernoulli indépendantes, de moyenne $\mathbb{E}(X) = \mathbb{P}(X = 1)$. La fréquence de succès de l'expérience converge donc p.s. vers la probabilité de succès.

Exemple des nombres normaux. On dit qu'un nombre reél est 2-normal si, quand on écrit sa décomposition dyadique, la proportion de zéros dans son écriture jusqu'à l'ordre n converge vers $1/2$ quand $n \to \infty$. D'après la loi des grands nombres, quand on tire un réel $\omega \in [0, 1[$ au hasard suivant la mesure de Lebesgue, on obtient p.s. un nombre normal. Idem avec des nombres b-normaux quand on utilise la base b ($b \geq 2$ entier). Presque tous les réels sont b-normaux pour tout b; mais concrètement, on ne connait pas d'exemple de tels nombres.

5.5 Méthodes de Monte-Carlo

a) Méthode simple

La loi des grands nombres permet le calcul approché de certaines intégrales par une approche probabiliste, la méthode de Monte-Carlo. Typiquement, supposons que nous souhaitions estimer une integrale du type

$$I = \int_{\mathbb{R}^d} \varphi(x) f(x) dx \,,$$

où $f : \mathbb{R}^d \to [0, \infty)$ est la densité d'une mesure de probabilité sur $\mathbb{R}^d$, et $\varphi : \mathbb{R}^d \to \mathbb{R}$ une fonction dans $L^1(\mathbb{R}^d, f(x))dx$.

On remarque tout d'abord que $I = \mathbb{E}(\varphi(X))$, où X désigne une v.a. de loi $P_X(dx) = f(x)dx$. Si on sait simuler une suite $X_1, \cdots, X_n, \cdots$ de v.a. indépendantes toutes de même loi que X, on peut appliquer la loi des grands nombres pour obtenir l'estimation suivante:

$$I \sim I_n = \frac{\varphi(X_1) + \cdots + \varphi(X_n)}{n} \,.$$

Voici l'énoncé précis.

Corollaire. *Sous les hypothèses précédentes, I_n converge p.s. vers I.*

Preuve: La suite des v.a. $\varphi(X_1), \cdots, \varphi(X_n), \cdots$ est i.i.d. et $\mathbb{E}(\varphi(X)) = \int \varphi(x) f(x) dx$. ■

Par rapport aux méthodes déterministes (e.g. méthode des trapèzes), la méthode de Monte Carlo présente plusieurs avantages: Tout d'abord, quand on a obtenu une approximation I_n au bout de n étapes, le calcul pour passer à l'approximation à l'ordre $n+1$ est très simple (et utilise le résultat à l'ordre n), puisque

$$I_{n+1} \;=\; \frac{nI_n + \varphi(X_{n+1})}{n+1}\,.$$

En revanche, dans beaucoup de méthodes deterministes, on choisit d'abord l'ordre de précision souhaité pour déterminer le nombre de pas d'approximation. Si on change l'ordre de précision, il faut tout recommencer. Ensuite, travailler en grandes dimensions ne fait pas croitre la complexité des calculs qui sont souvent en puissance de la dimension pour les méthodes déterministes. Enfin, on peut travailler sans hypothèse de régularité sur les fonctions f et φ.

Pour ce qui est des inconvénients, le premier vient de la difficulté qu'il y a à générer des suite de v.a. indépendentes (à partir d'un rang très grand, les algorithmes ont tendance à fournir des v.a. trop corrélées). En second lieu, il faudrait connaitre une estimation de l'erreur commise au rang n, information que ne fournit pas la loi des grands nombres (cf. suite du cours et théorème central limite); en règle générale, la vitesse de convergence est de l'ordre de $1/\sqrt{n}$, ce qui n'est pas très rapide en dimension 1.

b) Méthode avec rejet

La méthode du rejet est une procédure simple qui permet de générer une suite de v.a. indépendantes de même loi, à partir d'une suite de v.a. i.i.d. donnée, $(X_n, n \in \mathbb{N})$, et d'une suite indépendente de v.a. uniformes. Voici des hypothèses précises:

Soit $(X_n, n \in \mathbb{N})$ une suite de v.a. réelles indépendantes, toutes de loi $P_X(dx) = f(x)dx$. Soit $(U_n, n \in \mathbb{N})$ une suite de v.a. indépendantes, toutes de loi uniforme sur $[0,1]$, indépendante de la suite $(X_n, n \in \mathbb{N})$. Soit enfin $g : \mathbb{R} \to [0, \infty[$ une densité de probabilité sur $\mathbb{R}$ telle que pour tout $x \in \mathbb{R}$, $g(x)/f(x) \leq c$ pour une certaine constante $c \geq 1$.

Pour chaque entier k, on pose

$$B_k \;=\; \begin{cases} 1 & \text{si } cU_k < g(X_k)/f(X_k) \\ 0 & \text{sinon.} \end{cases}$$

Il est immédiat que la suite $B_0, \cdots, B_n, \cdots$ est une suite de Bernoulli avec

$$\mathbb{P}(B=1) \;=\; \int_{-\infty}^{\infty} \frac{g(x)}{cf(x)} f(x)dx \;=\; \frac{1}{c}\,.$$

Puis on pose

$$N(n) \;=\; \min\{k \in \mathbb{N} : B_0 + \cdots + B_k = n\}\,;$$

il découle facilement de la loi des grands nombres que $N(n) \sim cn$. Enfin, on considère les v.a. $Y_1 = X_{N(1)}, \cdots, Y_n = X_{N(n)}, \cdots$. Il est également facile de vérifier que la loi de Y_1 est $P_{Y_1}(dx) = g(x)dx$. En effet, $\mathbb{E}(\varphi(Y_1))$ vaut

$$\mathbb{E}\left(\sum_{k=0}^{\infty} \varphi(X_k)\mathbf{1}_{B_0=0,\cdots,B_{k-1}=0,B_k=1}\right)$$

$$= \sum_{k=0}^{\infty} \mathbb{E}\left(\varphi(X_k)\mathbf{1}_{\{g(X_k)/f(X_k)>cU_k\}}\right)(1-1/c)^k$$
$$= c\int_{-\infty}^{\infty}\left(\int_0^{g(x)/cf(x)} du\right)\varphi(x)f(x)dx \;=\; \int_{-\infty}^{\infty} dx\varphi(x)g(x)\,.$$

Enfin, il est aisé de voir que la suite $Y_1,\cdots,Y_n,\cdots$ est une suite de v.a. i.i.d. On déduit donc de la méthode de Monte-Carlo usuelle le résultat d'approximation suivant; dont on donnera une preuve directe

Corollaire. *Sous les hypothèses précédentes, on a pour toute $\varphi \in L^1(\mathbb{R}, g(x)dx)$:*

$$\lim_{n\to\infty}\frac{c}{n}\left(\sum_{k=1}^{n}\varphi(X_k)\mathbf{1}_{g(X_k)<cU_kf(X_k)}\right) \;=\; \int_{-\infty}^{\infty}\varphi(x)g(x)dx\,.$$

Preuve: Les v.a. $\left(\varphi(X_k)\mathbf{1}_{g(X_k)<cU_kf(X_k)}, k=1,\cdots\right)$ sont i.i.d., et pour $x\neq 0$

$$\mathbb{E}(\varphi(X)\mathbf{1}_{g(X)<cUf(X)}) \;=\; \int_{-\infty}^{\infty}\varphi(x)\frac{g(x)}{cf(x)}f(x)dx \;=\; c^{-1}\int_{-\infty}^{\infty}\varphi(x)g(x)dx\,.$$

Il ne reste qu'à appliquer la loi des grands nombres. ■

On notera que l'erreur croît avec le nombre c, et qu'on aura donc intéret à le choisir le plus petit possible.

5.6 Grandes déviations pour un jeu de pile ou face $(*)$

On se donne $X_1,\cdots,X_n,\cdots$ une suite de v.a. indépendantes de même loi μ, dont la somme partielle est notée $S_n = X_1+\cdots+X_n$. On supposera toujours que $\mathbb{E}|X|) < \infty$. La loi (faible) des grands nombres assure que si $m > \mathbb{E}(X)$, alors $\mathbb{P}(S_n \geq mn) \to 0$ (de même, pour $m < \mathbb{E}(X)$, la limite vaut alors 1). On s'intérèsse à la vitesse à laquelle a lieu la convergence; c'est ce qu'on appelle un problème de grandes déviations (de la moyenne empirique par rapport à la moyenne mathématique). Nous allons présenter un résultat permettant de minorer la vitesse de convergence dans la loi des grands nombres pour les sommes de v.a. indépendantes de même loi, qui admettent des moments exponentiels. Donnons tout d'abord quelques définitions.

Soit μ une loi de probabilité sur $\mathbb{R}$. On appelle cumulant de μ la fonction $\Lambda_\mu : \mathbb{R} \to (-\infty,\infty]$ donnée par

$$\Lambda_\mu(\lambda) \;=\; \log\left(\mathbb{E}(\mathrm{e}^{\lambda X})\right)\,,$$

avec la convention $\log\infty = \infty$.

Lemme. Λ_μ *est une fonction convexe.*

Preuve: La convexité découle de l'inégalité de Hölder. ■

On supposera dorénavant que Λ_μ est finie sur $\mathbb{R}$.

Proposition. *Supposons que le cumulant Λ_μ est fini sur $\mathbb{R}_+$. Soient $X_1, \cdots, X_n, \cdots$ une suite de v.a. indépendantes, toutes de loi μ. Posons enfin $S_n = X_1 + \cdots + X_n$. Pour tout $m \in \mathbb{R}$, on a*

$$-\frac{1}{n} \log \left(\mathbb{P} \left(S_n \geq mn \right) \right) \geq \sup\{\lambda m - \Lambda_\mu(\lambda) : \lambda \in \mathbb{R}_+\} .$$

Preuve: Pour tout $\lambda \in \mathbb{R}_+$, on a

$$\mathbb{E}(\exp\{\lambda S_n\}) = \mathbb{E}\left(\mathrm{e}^{\lambda X_1} \cdots \mathrm{e}^{\lambda X_n}\right) = \mathbb{E}\left(\mathrm{e}^{\lambda X_1}\right) \cdots \mathbb{E}\left(\mathrm{e}^{\lambda X_n}\right) = \exp\{n\Lambda_\mu(\lambda)\} .$$

En appliquant l'inégalité de Markov, il vient

$$\mathbb{P}\left(S_n \geq mn\right) \leq \mathrm{e}^{-\lambda mn} \mathbb{E}\left(\mathrm{e}^{\lambda S_n}\right) = \mathrm{e}^{-\lambda mn} \exp\{n\Lambda_\mu(\lambda)\} .$$

En passant au logarithme, on tire

$$\frac{-1}{n} \log \mathbb{P}\left(S_n \geq mn\right) \geq m\lambda - \Lambda_\mu(\lambda) ;$$

ce qui entraine le lemme. ■

La fonction $m \to \sup\{\lambda m - \Lambda_\mu(\lambda) : \lambda \in \mathbb{R}\}$ s'appelle la transformée de Legendre de μ, on la note souvent Λ_μ^*. On remarquera que c'est une fonction à valeurs positive (prendre $\lambda = 0$), convexe (enveloppe supérieure de fonctions linéaires) et semi-continue inférieurement (idem). De plus, si $m > \mathbb{E}(X_1)$, alors $\Lambda_\mu^*(m) > 0$ (car $\mathbb{E}(X_1)$ est la dérivée de Λ_μ en 0). Autrement dit, quand $m > \mathbb{E}(X_1)$, le terme de gauche dans la proposition précédente décroît au moins exponentiellement vite.

Une question bien plus délicate est de savoir si la vitesse de décroissance est la bonne, c'est-à-sire savoir si on a une minoration du même type. Nous allons maintenant étudier cette question dans le cas simple où μ est une loi de Bernoulli, i.e. S_n correspond à un jeu de pile ou face. Le problème général a été résolu par Cramèr, sous même hypothèse que dans la proposition précédente, i.e. finitude des moments exponentiels. Nous énoncerons donc une forme faible du théorème de Cramèr; la démonstration contient cependant les idées principales qui sont nécéssaires pour traiter le cas général.

On désignera par $G_X(s) = 1 - (1 - s)p$ $(s \geq 0)$, la fonction génératrice de la loi de Bernoulli de paramètre $p \in]0, 1[$. On remarquera que $\Lambda_\mu(\lambda) = \log G_X(\mathrm{e}^\lambda)$. En appliquant la proposition précédente, on tire immédiatement le lemme suivant:

Lemme. *Pour tout $m \in [p, 1[$, on a*

$$\frac{-1}{n} \log \mathbb{P}\left(S_n \geq mn\right) \geq \sup_{s \geq 0} \left(m \log s - \log G_X(s)\right) .$$

Pour obtenir une minoration, on a recours à l'idée clef de la théorie des grandes déviations, qui consiste à travailler sous une loi de probabilité équivalente à $\mathbb{P}$ sous

laquelle les v.a. X_j seront toujours indépendantes, mais cette fois de moyenne m (au lieu de p), puis à appliquer l'inégalité de Bienaymé-Tchebychev. La densité de la nouvelle mesure de probabilité a toujours une expression exponentielle.

Proposition. *Soient $s > 0$ et $n \geq 1$ un entier.*

(i) La v.a. $s^{S_n} \exp\{-n \log G_X(s)\}$ est positive et d'intégrale 1.

(ii) Soit $\mathbb{P}^{(s)}$ la mesure de probabilité sur $(\Omega, \mathcal{F})$ de densité

$$\frac{d\mathbb{P}^{(s)}}{d\mathbb{P}} \;=\; s^{S_n} \exp\{-n \log G_X(s)\}\,.$$

Sous $\mathbb{P}^{(s)}$, les v.a. $X_1, \cdots, X_n$ sont des v.a. de Bernoulli indépendantes, toutes de parametre $p^{(s)} = ps \exp\{-\log G_X(s)\}$.

Preuve: (i) est évident.

(ii) Soient $\varepsilon_1, \cdots, \varepsilon_n$ une suite de n termes dans $\{0, 1\}$. On a

$$\begin{aligned}
& \mathbb{P}^{(s)}\left(X_1 = \varepsilon_1, \cdots, X_n = \varepsilon_n\right) \\
= \;& \exp\{-n \log G_X(s)\} \mathbb{E}\Big(s^{X_1 + \cdots + X_n}, X_1 = \varepsilon_1, \cdots, X_n = \varepsilon_n\Big) \\
= \;& \exp\{-n \log G_X(s)\} s^{\varepsilon_1 + \cdots + \varepsilon_n} \mathbb{P}(X_1 = \varepsilon_1) \cdots \mathbb{P}(X_n = \varepsilon_n) \\
= \;& \left(\exp\{-\log G_X(s)\} s^{\varepsilon_1} \mathbb{P}(X_1 = \varepsilon_1)\right) \cdots \left(\exp\{-\log G_X(s)\} s^{\varepsilon_n} \mathbb{P}(X_n = \varepsilon_n)\right)\,.
\end{aligned}$$

Comme

$$\exp\{-\log G_X(s)\} s^0 \mathbb{P}(X = 0) + \exp\{-\log G_X(s)\} s^1 \mathbb{P}(X = 1) \;=\; 1\,,$$

notre assertion en découle. ■

Lemme. *Si $m \in [p, 1[$, la fonction $s \to m \log s - \log G_X(s)$ atteint son supremum pour $s = s_m = \frac{m - mp}{p - mp} \geq 1$. On a encore $p \exp\{-\log G_X(s_m)\} s_m = m$.*

Preuve: Pour la première assertion, il suffit de dériver. Le supremum est atteint lorsque

$$\frac{m}{s} \;=\; \frac{p}{1 - (1 - s)p}\,,$$

ce qui donne les expressions souhaitées pour s_m. ■

En particulier, on sait désormais que sous $\mathbb{P}^{(s_m)}$, la suite $X_1, \cdots, X_n$ est une suite de v.a. de Bernoulli indépendantes, toutes de moyenne m. On énonce maintenant la majoration.

Lemme. *Pour tout $m \in [p, 1[$, on a*

$$\limsup_{n \to \infty} \frac{-1}{n} \log \mathbb{P}\left(S_n \geq mn\right) \;\leq m \log s_m - \log G_X(s_m)\,.$$

Preuve: Prenons $\varepsilon > 0$ assez petit pour que $m + 2\varepsilon < 1$. Soit $\mathbb{Q}^{(m+\varepsilon)}$ la probabilité $\mathbb{P}^{(s)}$ pour $s = s_{m+\varepsilon}$ (voir le lemme précédent). Nous avons

$$\begin{aligned}\mathbb{P}(S_n > mn) &= \mathbb{Q}^{(m+\varepsilon)}\left(s_{m+\varepsilon}^{-S_n}\exp\{n\log G_X(s_{m+\varepsilon})\}, S_n > mn\right)\\ &\geq s_{m+\varepsilon}^{-(m+\varepsilon)n}\exp\{n\log G_X(s_{m+\varepsilon})\}\mathbb{Q}^{(m+\varepsilon)}(mn < S_n \leq (m+2\varepsilon)n)\,.\end{aligned}$$

Nous savons que sous $\mathbb{Q}^{(m+\varepsilon)}$, les v.a. $X_1, \cdots, X_n$ sont indépendantes et suivent toutes une loi de Bernoulli de parametre $m + \varepsilon$. En conséquence, sous $\mathbb{Q}^{(m+\varepsilon)}$, S_n suit une loi binomiale de paramètres $(n, m + \varepsilon)$, en particulier sa moyenne est $(m + \varepsilon)n$ et sa variance $n((m + \varepsilon) - (m + \varepsilon)^2)$. D'après l'inégalité de Bienaymé-Tchebychev

$$\begin{aligned}\mathbb{Q}^{(m+\varepsilon)}(mn < S_n \leq (m+2\varepsilon)n) &= 1 - \mathbb{Q}^{(m+\varepsilon)}\left(|S_n - n(m+\varepsilon)| \geq \varepsilon n\right)\\ &\geq 1 - \frac{n((m+\varepsilon) - (m+\varepsilon)^2)}{\varepsilon^2 n^2}\,.\end{aligned}$$

Le terme de droite tend vers 1 quand n tend vers ∞. On a donc

$$\limsup_{n\to\infty}\frac{-1}{n}\log\mathbb{P}(S_n \geq mn) \leq (m+\varepsilon)\log s_{m+\varepsilon} - \log G_X(s_m + \varepsilon)\,.$$

Quand on fait tendre ε vers 0, $s_{m+\varepsilon}$ converge vers s_m, et le lemme est démontré. ■

En conclusion, nous avons obtenu le résultat de grandes déviations suivant:

Proposition *Pour tout $m \in [p, 1[$, on a*

$$\lim_{n\to\infty}\frac{-1}{n}\log\mathbb{P}(S_n \geq mn) = \sup_{s\geq 0}\left(m\log s - \log G_X(s)\right)\,.$$

Plus précisément, le point en lequel la fonction $s \to m\log s - \log G_X(s)$ atteint son supremum est $s = s_m = \frac{m-mp}{p-mp}$.

On remarquera que la restriction sur l'ensemble sur lequel m varie n'en est pas une en pratique.

Chapitre 6

Convergence en loi

On introduit un nouveau concept de convergence pour une suite de v.a., qui ne dépend cette fois que de la loi de chacune de ces v.a.

6.1 Convergence d'une suite de mesures

Soient $\mu_1, \cdots, \mu_n, \cdots$ une suite de mesures de Radon sur $\mathbb{R}^d$ (ou, plus généralement, sur un espace topologique localement compact). On a deux notions naturelles de convergence:

Définitions. *Soit μ une mesure de Radon sur $\mathbb{R}^d$. On s'intéresse aux fonctions mesurables $f : \mathbb{R}^d \to [0, \infty[$ pour lesquelles on a*

$$(\dagger) \qquad \int_{\mathbb{R}^d} f(x)\mu_n(dx) \to \int_{\mathbb{R}^d} f(x)\mu(dx)\,.$$

(1) On dit que μ_n converge **étroitement** *vers μ si $(\dagger)$ est vérifié pour toute fonction f continue bornée.*

(2) On dit que μ_n converge **vaguement** *vers μ si $(\dagger)$ est vérifié pour toute fonction f continue à support compact.*

Dans les deux cas, la limite est bien sûr unique.

Exemple: Soit $(x_n : n \in \mathbb{N})$ une suite de réels, et prenons pour $\mu_n = \delta_{x_n}$. Si $x_n \to x$, alors μ_n converge étroitement vers la masse de Dirac en x, δ_x. Réciproquement, supposons que μ_n converge étroitement vers une mesure μ. Alors nécéssairement, μ est une mesure de probabilité. Posons $f(t) = \arctan(t) + \pi/2$. On sait que $f(x_n)$ va converger vers $\mu(f) \in [0, 2\pi]$. On ne pourrait avoir $\mu(f) = 0$ que si x_n tendait vers $-\infty$. Mais dans ce cas on aurait $\mu(g) = 0$ pour tout fonction g continue à support compact, ce qui est impossible. Donc $\mu(f) \neq 0$, et on montre de même que $\mu(f) \neq \pi$. Il existe un unique réel x tel que $\mu(f) = f(x)$, et on voit alors en composant avec la fonction continue tan que x_n converge vers x.

On a évidemment (cv. étroite) $\Rightarrow$ (cv. vague), et la réciproque peut être fausse. Par exemple, si on prend $\mu_n = \delta_n$ (masse de Dirac en n, alors μ_n tend vers 0 vaguement,

mais pas étroitement. Néanmoins, les deux notions coincident en fait si l'on rajoute une hypothèse supplémentaire.

Proposition. *Soit $(\mu_n : n \in \mathbb{N})$ une suite de mesure finies sur $\mathbb{R}^d$, et μ une mesure finie sur $\mathbb{R}^d$. Pour que μ_n converge étroitement vers μ, il faut et il suffit que μ_n converge vaguement vers μ et que $\mu_n(\mathbb{R}^d) \to \mu(\mathbb{R}^d)$ (convergence des masses).*

Preuve: Si on a convergence étroite, alors $\mu_n(\mathbb{R}^d) = \int 1\mu_n(dx) \to \int 1\mu(dx) = \mu(\mathbb{R}^d)$, et donc la convergence des masses est assurée.

Pour la réciproque, soit $f \geq 0$ une fonction continue bornée et $\varepsilon > 0$ fixé. On peut supposer que $|f| \leq 1$. Comme μ est une mesure finie, il existe une fonction φ continue à support compact avec $0 \leq \varphi \leq 1$ et $\int \varphi(x)\mu(dx) > \mu(\mathbb{R}^d) - \varepsilon/5$. On déduit qu'il existe donc un entier N tel que $\int \varphi(x)\mu_n(dx) > \mu_n(\mathbb{R}^d) - \varepsilon/4$ dès que $n \geq N$. D'autre part, la fonction $f\varphi$ est continue et à support compact; on sait donc trouver un entier M tel que

$$\left| \int \varphi(x)f(x)\mu_n(dx) - \int \varphi(x)f(x)\mu(dx) \right| < \varepsilon/2 \qquad \text{dès que } n \geq M\,.$$

On a alors pour tout $n \geq M + N$

$$\begin{aligned}
& \left| \int f(x)\mu_n(dx) - \int f(x)\mu(dx) \right| \\
\leq\ & \left| \int \varphi(x)f(x)\mu(dx) - \int \varphi(x)f(x)\mu_n(dx) \right| \\
& \quad + \left| \int (1-\varphi(x))f(x)\mu(dx) \right| + \left| \int (1-\varphi(x))f(x)\mu_n(dx) \right| \\
\leq\ & \frac{\varepsilon}{2} + \int (1-\varphi(x))\mu_n(dx) + \int (1-\varphi(x))\mu(dx) \\
\leq\ & \frac{\varepsilon}{2} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \epsilon\,.
\end{aligned}$$

Ceci montre qu'on a convergence étroite. ■

6.2 Convergence en loi d'une suite de v.a.

On va utiliser les notions précédentes dans le cas où les mesures μ_n sont des lois de probabilité, i.e. représentent des lois de variables aléatoires.

Définition. *On dit qu'une suite de v.a. $X_1, \cdots, X_n, \cdot$ converge en loi vers une v.a. X si les mesures de probabilites P_{X_n} convergent étroitement vers P_X. Autrement dit, pour toute fonction continue bornée f:*

$$\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))\,.$$

En appliquant le résultat du paragraphe précédent, on voit qu'il suffit d'avoir la convergence vague (la convergence des masses est automatique).

Tout d'abord, on compare la convergence en loi avec les autres types de convergence pour des suites de v.a.

Proposition. *Si $X_1, \cdots, X_n, \cdots$ est une suite de v.a; qui converge en probabilité vers une v.a. X (en particulier, si la convergence a lieu dans $L^1(\Omega, \mathbb{P})$, ou p.s.), alors X_n converge également vers X en loi.*

Preuve: On sait que de toute sous-suite, on peut extraire une sous-sous-suite qui converge p.s. vers X. En appliquant le théorème de convergence dominée ($|f(X_n)|$ reste dominé par $\sup_{\mathbb{R}} |f(x)| < \infty$), on voit que $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ le long de cette sous-sous-suite. Comme ce résultat est valable pour toute suite extraite, on a bien $\lim_{n\to\infty} \mathbb{E}(f(X_n)) = \mathbb{E}(f(X))$. ■

La réciproque est bien sûr fausse (si $X_1, \cdots, X_n, \cdots$ est une suite de v.a. i.i.d., par exemple de Bernoulli, il est facile de vérifier que X_n ne converge pas en probabilités, puisque le critère de Cauchy ne peut pas être vérifié). Cependant, on a une réciproque très partielle qu'on laisse en exercice:

Exercice: Montrer que si une suite de v.a. définies sur un même espace probabilisé converge en loi vers une constante p.s., alors la convergence a également lieu en probabilités.

6.3 Cas des v.a. à valeurs entières

Etudions la convergence en loi pour des v.a. entières, où les résultats sont très simples.

Proposition. *Soient $X, X_1, \cdots, X_n, \cdots$ des v.a. à valeurs dans $\mathbb{N}$. Les assertions suivantes sont équivalentes:*

(i) X_n converge vers X en loi.

(ii) Pour chaque $k \in \mathbb{N}$, on a $\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k)$.

(iii) G_{X_n} converge simplement vers G_X sur $[0, 1]$ (comme d'habitude, G_Y est la fonctin génératrice de Y).

Preuve: (i) $\Rightarrow$ (ii) Prendre pour f une fonction continue qui vaut 1 sur en k et 0 pour les autres entiers.

(ii) $\Rightarrow$ (i) Soit f une fonction continue à support compact, disons dans $[0, N]$. On a

$$\mathbb{E}(f(X_n)) = \sum_{k=0}^{N} f(k)\mathbb{P}(X_n = k) \to \sum_{k=0}^{N} f(k)\mathbb{P}(X = k) = \mathbb{E}(f(X)) .$$

(ii) $\Rightarrow$ (iii) Même argument que ci-dessus en ajoutant la convergence dominée.

(iii) $\Rightarrow$ (ii) On rappelle un résultat classique sur la convergence des fonctions holomorphes. Soit $(h_n : n \in \mathbb{N})$ une suite de fonctions holomorphes sur le disque unité, telle que pour chaque n, les coefficients de h_n dans son expression comme série entière sont tous positifs. Si h_n converge simplement sur $[0, 1]$, alors h_n converge également simplement sur tout le disque unité, et la limite h est une fonction holomorphe sur le disque. De plus, pour tout entier k, la dérivée k-ième de h_n converge simplement vers la dérivée k-ième de h (sur le disque unité ouvert).

Appliquons ceci à $h_n = G_{X_n}$. La dérivée k-ième de h_n au point 0 est $k! \times \mathbb{P}(X_n = k)$, et donc on a (ii). ■

6.4 Convergence en loi et fonctions de répartition

On va ensuite étudier la convergence en loi des v.a. réelles à l'aide de leurs fonctions de répartition F_{X_n}. On dira qu'un réel x est un point de continuité d'une fonction de répartition F_X si $F_X(x) = F_X(x-)$, et que c'est un point de discontinuité sinon. Autrement dit, x est une point de discontinuité pour F_X si et seulement si la loi de X, P_X, a un atome en x.

Proposition *Soient $X_1, \cdots, X_n, \cdots$ une suite de v.a. réelles, de fonctions de répartition F_{X_n}. Soit F_X la fonction de répartition d'une v.a. réelle X. Les assertions suivantes sont équivalentes:*

(1) X_n converge en loi vers X.

(2) Si x est un point de continuité de F_X, alors $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$.

Preuve: (1) $\Rightarrow$ (2) Pour tout $\varepsilon > 0$, il existe une fonction $\varphi_\varepsilon : \mathbb{R} \to [0,1]$ continue telle que $\varphi_\varepsilon(t) = 1$ si $t \leq x$ et $\varphi_\varepsilon(t) = 0$ si $t \geq x + \varepsilon$. On a alors d'une part

$$\lim_{n\to\infty} \mathbb{E}(\varphi_\varepsilon(X_n)) \;=\; \mathbb{E}(\varphi_\varepsilon(X)) \;\leq \mathbb{P}(X \leq x + \varepsilon) \;=\; F_X(x+\varepsilon)\,.$$

Or $\mathbb{E}(\varphi_\varepsilon(X_n)) \geq \mathbb{P}(X_n \leq x) = F_{X_n}(x)$, et donc $\limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x+\varepsilon)$. Comme $\varepsilon > 0$ peut être arbitrairement petit et que F_X est continue à droite, on déduit que

$$\limsup_{n\to\infty} F_{X_n}(x) \;\leq\; F_X(x)\,.$$

Pour établir l'inégalité réciproque, on prend une fonction continue $\psi_\varepsilon : \mathbb{R} \to [0,1]$ telle que $\psi(t) = 1$ si $t \leq x - \varepsilon$ et $\psi(t) = 0$ si $t \geq x$. Comme précédemment, on tire cette fois dans un premier temps

$$\lim_{n\to\infty} \mathbb{E}(\psi_\varepsilon(X_n)) \;=\; \mathbb{E}(\psi_\varepsilon(X)) \;\geq \mathbb{P}(X \leq x - \varepsilon) \;=\; F_X(x-\varepsilon)\,,$$

puis $\liminf_{n\to\infty} F_{X_n}(x) \geq F_X(x-\varepsilon)$. Comme, par hypothèse, $F_X(x-) = F_X(x)$ et $\varepsilon > 0$ peut être pris arbitrairement petit, on a finalement

$$\liminf_{n\to\infty} F_{X_n}(x) \;\geq\; F_X(x)\,.$$

(2) $\Rightarrow$ (1) Supposons d'abord que $f : \mathbb{R} \to \mathbb{R}$ est une fonction de classe $\mathcal{C}^1$ à support compact. En particulier, sa dérivée f' est bornée et a un support compact. A l'aide du Théorème de Fubini (c'est-à-dire dans ce cas d'une intégration par parties), on a

$$\mathbb{E}(f(X_n)) \;=\; \int_{\mathbb{R}} P_{X_n}(dt) \left(\int_{-\infty}^{t} f'(s) ds \right) \;=\; \int_{-\infty}^{\infty} (1 - F_{X_n}(s))\, f'(s) ds\,.$$

L'ensemble des points de discontinuité d'une fonction croissante est au plus dénombrable, donc de mesure de Lebesgue nulle. On sait que hors de ces points, $1 - F_{X_n}$ converge

vers $1 - F_X$, en restant bien évidemment borné par 1. D'autre part, la mesure (positive) $|f'(s)|ds$ a clairement une masse totale finie. On déduit par convergence dominée que

$$\lim_{n\to\infty} \mathbb{E}(f(X_n)) \;=\; \int_{-\infty}^{\infty} (1 - F_X(s))\, f'(s) ds \;=\; \mathbb{E}(f(X))\,.$$

Voyons comment traiter maintenant le cas où f est une fonction continue à support compact, mais plus nécessairement dérivable. Pour tout $\varepsilon > 0$, on sait trouver une fonction g de classe $\mathcal{C}^1$ à support compact, telle que $|f(t) - g(t)| < \varepsilon/4$ pour tout $t \in \mathbb{R}$ (pour cela, on peut par exemple prendre la convolée de f avec une approximation de l'unité de classe $\mathcal{C}^1$ à support compact). D'autre part, on sait trouver un entier N tel que $|\mathbb{E}(g(X_n)) - \mathbb{E}(g(X))| \leq \varepsilon/2$ dès que $n \geq N$. En conséquence, pourvu que $n \geq N$, on a:

$$\begin{array}{rl} & |\mathbb{E}\,(f(X_n)) - \mathbb{E}\,(f(X))\,| \\ \leq & \mathbb{E}\,(|f(X_n) - g(X_n)|) + |\mathbb{E}\,(g(X_n) - g(X))\,| + \mathbb{E}\,(|f(X) - g(X)|) \\ \leq & \dfrac{\varepsilon}{4} + \dfrac{\varepsilon}{2} + \dfrac{\varepsilon}{4} = \varepsilon\,. \end{array}$$

Ceci établit la convergence vague de X_n vers X, donc en fait la convergence en loi. ■

Pour conclure cette section, nous allons présenter (une forme simple d') un résultat remarquable de représentation dû à Skorohod.

Corollaire *Soient $X_1, \ldots$ une suite de variables aléatoires réelles qui converge en loi vers une variable aléatoire X_∞. Il existe alors des variables $Y_1, \ldots$ et Y_∞ telles que pour chaque indice n, X_n et Y_n ont la même loi, avec $\lim_{n\to\infty} Y_n = Y_\infty$ presque-sûrement.*

Preuve: Pour simplifier, nous allons supposer que les fonctions de répartition F_n de X_n et F_∞ de X_∞ sont des bijections de $\mathbb{R}$ dans $]0,1[$ (i.e. ce sont des fonctions continues et strictement croissantes). La démonstration peut être adaptée au cas général au prix de quelques difficultés techniques.

Soit F_n^{-1} la fonction réciproque de F_n, et F_∞^{-1} celle de F_∞. Introduisons une variable U de loi uniforme sur $[0,1]$. On a vu (cf simulation de variables aléatoires) que $F_n^{-1}(U) := Y_n$ est une variable aléatoire de même loi que X_n. L'hypothèse que X_n converge en loi vers X_∞ assure que F_n converge simplement vers F_∞, et il en découle que F_n^{-1} converge simplement vers F_∞^{-1}. On voit maintenant que Y_n converge presque sûrement (en fait sauf quand U prend la valeur 0 ou 1) vers $Y_\infty = F_\infty^{-1}(U)$. ■

6.5 Fonctions Caractéristiques

L'objet de cette longue section est d'introduire la notion de fonction caractéristique d'une variable aléatoire (c'est la version probabiliste de la transformée de Fourier d'une mesure), afin d'en donner une application à la convergence en loi.

6.5.1 Définition et exemples

Dans tout cette section, X désigne une v.a. à valeurs dans $\mathbb{R}^d$ et P_X sa loi, c'est-à-dire que P_X est une mesure de probabilité sur $\mathbb{R}^d$.

Définition. *L'application* $\Phi_X : \mathbb{R}^d \to \mathbb{C}$ *donnée par*

$$\Phi_X(\lambda) \;=\; \mathbb{E}\left(\mathrm{e}^{\mathrm{i}\langle\lambda,X\rangle}\right) \;=\; \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\langle\lambda,x\rangle} P_X(dx)$$

s'appelle la fonction caractéristique de X.

Autrement dit, la fonction caractéristique d'une variable aléatoire est la transformée de Fourier de sa loi.

Exemples fondamentaux: On se concentre sur la dimension $d = 1$.

• Si P_X est la loi uniforme U sur $[a, b]$,

$$\Phi_U(\lambda) = \frac{1}{\mathrm{i}(b-a)\lambda}\left(\mathrm{e}^{\mathrm{i}b\lambda} - \mathrm{e}^{\mathrm{i}a\lambda}\right).$$

• Si P_X est la loi exponentielle de paramètre $q > 0$, $P_{\mathbf{e}}(dx) = q\mathrm{e}^{-qx}\mathbf{1}_{\{x\geq 0\}}dx$, alors

$$\Phi_{\mathbf{e}}(\lambda) \;=\; \int_0^\infty q\mathrm{e}^{-qx}\mathrm{e}^{\mathrm{i}\lambda x}dx \;=\; \frac{q}{q-\mathrm{i}\lambda}\,.$$

• Lorsque P_X est la loi de Cauchy standard, $P_C(dx) = \frac{1}{\pi}(1+x^2)^{-1}dx$, on applique le théorème des résidus à la fonction méromorphe $z \to (1+z^2)^{-1}\mathrm{e}^{\mathrm{i}\lambda z}$, dont l'unique pôle dans le demi-plan supérieur est ı, avec résidu $(2\imath)^{-1}\mathrm{e}^{-\lambda}$. On déduit facilement en intégrant sur un grand demi-cercle centré en 0 et en appliquant le théorème des résidus que

$$\Phi_C(\lambda) \;=\; \frac{1}{\pi}\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\lambda x}(1+x^2)^{-1}dx \;=\; \mathrm{e}^{-\lambda}\,, \qquad \lambda \geq 0.$$

Par symétrie, on voit que $\Phi_C(\lambda) = \Phi_C(-\lambda)$, et donc, pour $\lambda \in \mathbb{R}$ de signe quelconque, on a $\Phi_C(\lambda) = \mathrm{e}^{-|\lambda|}$. A titre d'exercice, on pourra vérifier cette formule en utilisant la transformée de Fourier inverse.

• Si P_X est la loi normale standard $\mathcal{N}(0,1)$, alors

$$\Phi_{\mathcal{N}(0,1)}(\lambda) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left\{-\frac{x^2}{2}\right\}\mathrm{e}^{\mathrm{i}\lambda x}dx$$

est à valeurs réelles (par symétrie). Une dérivation sous le signe somme donne

$$\Phi'_{\mathcal{N}(0,1)}(\lambda) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{i}x\exp\left\{-\frac{x^2}{2}\right\}\mathrm{e}^{\mathrm{i}\lambda x}dx\,.$$

A l'aide d'une intégration par parties (dériver $x \to \mathrm{e}^{\mathrm{i}\lambda x}$ et intégrer $x \to x\mathrm{e}^{-x^2/2}$), on obtient

$$\Phi'_{\mathcal{N}(0,1)}(\lambda) \;=\; -\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \lambda\exp\left\{-\frac{x^2}{2}\right\}\mathrm{e}^{\mathrm{i}\lambda x}dx \;=\; -\lambda\Phi_{\mathcal{N}(0,1)}(\lambda)\,.$$

La résolution de cette équation différentielle donne $\log|\Phi_{\mathcal{N}(0,1)}(\lambda)| = -\lambda^2/2 + c$. Comme $\Phi_{\mathcal{N}(0,1)}(0) = 1$, on conclut que

$$\Phi_{\mathcal{N}(0,1)}(\lambda) = \exp\left\{-\lambda^2/2\right\}.$$

• Plus généralement, on peut utiliser la representation $Y = \sigma X + m$ pour Y de loi $\mathcal{N}(m, \sigma^2)$ en fonction d'une variable normale standard X afin d'obtenir

$$\Phi_{\mathcal{N}(m,\sigma^2)}(\lambda) = \exp\left\{\mathrm{i}\lambda m - \sigma^2\lambda^2/2\right\}.$$

6.5.2 Principales propriétés des fonctions caractéristiques

Voyons d'abord quelque propriétés élémentaires:

• On a toujours $\Phi_X(0) = 1$.

• La fonction caractéristique prend ses valeurs dans le disque unitié, i.e. $|\Phi_X(\lambda)| \leq 1$ pour tout $\lambda \in \mathbb{R}^d$. En effet, $|\mathbb{E}\left(\mathrm{e}^{\mathrm{i}\langle\lambda,X\rangle}\right)| \leq \mathbb{E}\left(|\mathrm{e}^{\mathrm{i}\langle\lambda,X\rangle}|\right) = 1$.

• La fonction caractéristique est continue sur $\mathbb{R}^d$. C'est une conséquence immédiate de la continuité de l'application $\lambda \to \mathrm{e}^{\mathrm{i}\langle\lambda,x\rangle}$ pour chaque $x \in \mathbb{R}^d$, et du théorème de convergence dominée.

• Lorsque la loi P_X est absolument continue par rapport à la mesure de Lebesgue, i.e. $P_X(dx) = g(x)dx$, où $g : \mathbb{R}^d \to [0, \infty]$ est une fonction mesurable telle que $\int_{\mathbb{R}^d} g(x)dx = 1$, alors Φ_X est la transformée de Fourier de la fonction g. En particulier, d'après le théorème de Riemann Lebesgue, on sait que Φ_X tend vers 0 à l'infini.

Comme son nom l'indique, la fonction caractéristique caractérise la loi de X. Voici une façon d'établir ce résultat (pour simplifier les arguments, nous supposerons dans la suite de la variable X est à valeurs entières.

Lemme. *Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction continue à support compact. On a*

$$\mathbb{E}(f(X)) = (2\pi)^{-1} \lim_{a\to 0} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \mathrm{e}^{-i\lambda x}\mathrm{e}^{-a|\lambda|}\Phi_X(\lambda)d\lambda\right) f(x)dx\,.$$

En conséquence, deux variables aléatoires réelles, X et Y, ont même loi si et seulement si $\Phi_X = \Phi_Y$.

Preuve: Si $P_X = P_Y$, on a en particulier pour tout $\lambda \in \mathbb{R}^d$

$$\Phi_X(\lambda) = \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\langle\lambda,x\rangle} P_X(dx) = \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\langle\lambda,x\rangle} P_Y(dx) = \Phi_Y(\lambda)\,.$$

Pour la réciproque, on va pour supposer pour simplifier que la dimension est $d = 1$. Le cas des dimensions supérieures est similaire, mais avec des notations plus lourdes. On remarque d'abord que

$$\int_{-\infty}^{\infty} \mathrm{e}^{i\lambda x}\mathrm{e}^{-a|\lambda|}d\lambda = \frac{1}{a - ix} + \frac{1}{a + ix} = \frac{2a}{a^2 + x^2}\,.$$

On a donc

$$\int_{\mathbb{R}} \frac{2a}{a^2 + (y - x)^2} P_X(dy) = \int_{\mathbb{R}} P_X(dy) \int_{-\infty}^{\infty} d\lambda \mathrm{e}^{i\lambda(y-x)}\mathrm{e}^{-a|\lambda|}\,;$$

puis en appliquant le théorème de Fubini

$$\int_{\mathbb{R}} \frac{2a}{a^2+(y-x)^2} P_X(dy) \;=\; \int_{-\infty}^{\infty} \mathrm{e}^{-i\lambda x}\mathrm{e}^{-a|\lambda|}\Phi_X(\lambda)d\lambda\,.$$

Considérons ensuite une fonction continue f à support compact. En appliquant le théorème de Fubini et un changement de variable, on tire

$$\int_{-\infty}^{\infty} f(x)\left(\int_{\mathbb{R}} \frac{2a}{a^2+(y-x)^2} P_X(dy)\right)dx \;=\; \int_{\mathbb{R}}\left(\int_{-\infty}^{\infty} f(y-at)\frac{2}{1+t^2}dt\right)P_X(dy)\,.$$

Quand on fait tendre a vers 0, le terme entre parenthèses dans le membre de droite converge vers $2\pi f(y)$ en restant dominé par $2\pi||f||_\infty$. En appliquant le théorème de convergence dominée, on voit donc que

$$2\pi\int_{\mathbb{R}} f(y)P_X(dy) \;=\; \lim_{a\to 0}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\mathrm{e}^{-i\lambda x}\mathrm{e}^{-a|\lambda|}\Phi_X(\lambda)d\lambda\right)f(x)dx\,.$$

Le terme de droite est une fonction de f et de Φ_X, et donc Φ_X caractérise bien la loi P_X. ■

Voici un corollaire utile de la formule du Lemme précédent.

Corollaire. *Si Φ_X est une fonction intégrable pour la mesure de Lebesgue, alors la loi de X est absolument continue par rapport à la mesure de Lebesgue et admet une densité continue qui est donnée par la formule d'inversion de Fourier*

$$P_X(dx)/dx \;=\; \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{e}^{-\mathrm{i}\lambda x}\Phi_X(\lambda)d\lambda\,, \qquad x\in\mathbb{R}.$$

Preuve: Il suffit d'appliquer le théorème de convergence dominée dans la formule du lemme précédent. ■

Lemme. *Soient $X_1,\cdots,X_d$ sont d variables aléatoires à valeurs réelles, posons $X=(X_1,\cdots,X_d)$. Les v.a. $X_1,\cdots,X_d$ sont indépendantes si et seulement si pour tout $\lambda=(\lambda_1,\cdots,\lambda_d)\in\mathbb{R}^d$, on a*

$$\Phi_X(\lambda) \;=\; \prod_{n=1}^{d}\Phi_{X_n}(\lambda_n)\,.$$

Preuve: Si les v.a. $X_1,\cdots,X_d$ sont indépendantes, alors

$$\mathbb{E}\left(\mathrm{e}^{\mathrm{i}\langle\lambda,X\rangle}\right) \;=\; \mathbb{E}\left(\exp\left\{\mathrm{i}\sum_{n=1}^{d}\lambda_n X_n\right\}\right) \;=\; \prod_{n=1}^{d}\mathbb{E}\left(\mathrm{e}^{\mathrm{i}\lambda_n X_n}\right)\,.$$

Réciproquement, on suppose que l'identité précédente est vérifiée, et on considère $Y_1,\cdots,Y_d$, d v.a. indépendantes, avec X_n de même loi que Y_n pour chaque n (les Y_n

peuvent être construites sur un espace produit). Posons $Y = (Y_1, \cdots, Y_d)$. D'après la première partie, X et Y ont même fonction caractéristique, donc même loi. En conséquence, pour tout pavé borélien $B = B_1 \times \cdots \times B_d$, on a

$$\begin{aligned} \mathbb{P}(X_1 \in B_1, \cdots, X_d \in B_d) &= \mathbb{P}(X \in B) \\ &= \mathbb{P}(Y \in B) = \mathbb{P}(Y_1 \in B_1, \cdots, Y_d \in B_d) \\ &= \prod_{n=1}^{d} \mathbb{P}(Y_n \in B_n) = \prod_{n=1}^{d} \mathbb{P}(X_n \in B_n)\,. \end{aligned}$$

■

Corollaire. *Si $X_1, \cdots, X_n$ sont n v.a. réelles indépendantes, et si $S = X_1 + \cdots + X_n$, alors*

$$\Pi_S(\lambda) = \Phi_{X_1}(\lambda) \times \cdots \times \Phi_{X_n}(\lambda)\,, \qquad \lambda \in \mathbb{R}.$$

Exemples • Si C_1 et C_2 sont deux v.a. de Cauchy standard indépendantes, alors

$$\Phi_{C_1+C_2}(\lambda) = \mathrm{e}^{-|\lambda|} \times \mathrm{e}^{-|\lambda|} = \mathrm{e}^{-2|\lambda|} = \mathrm{e}^{-|2\lambda|} = \Phi_{2C}(\lambda)\,.$$

Comme la fonction caractéristique détermine la loi, on a donc montré que la somme de deux v.a. de Cauchy indépendantes a même loi que deux fois une v.a. de Cauchy standard.

• Si X a pour loi $\mathcal{N}(m, \sigma^2)$ et X a pour loi $\mathcal{N}(m', \sigma'^2)$, avec X et X' indépendantes, alors $X + X'$ a pour loi a pour loi $\mathcal{N}(m + m', \sigma^2 + \sigma'^2)$.

Il est également important de comprendre ce résultat en terme de convolution. Si μ et ν sont deux mesures finies sur $\mathbb{R}$ (voire, sur $\mathbb{R}^d$) on rappelle que la mesure convolée $m * \nu$ est définie par la formule

$$(\dagger) \qquad \int_{\mathbb{R}} f(t)\mu * \nu(dt) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x+y)\mu(dx)\nu(dy)\,,$$

où f désigne une fonction borélienne bornée générique. En prenant $f(t) = \mathrm{e}^{\imath\lambda t}$ dans $(\dagger)$ et en appliquant le théorème de Fubini, on tire que la transformée de Fourier de la mesure finie $\mu * \nu$ est le produit des transformées de Fourier de μ et de ν.

Si on suppose maintenant que μ et ν sont des mesures de probabilités, i.e. deux lois de v.a. réelles. La mesure produit $\mu \otimes \nu$ sur $\mathbb{R} \times \mathbb{R}$ est donc la loi de deux v.a. *indépendantes*, X et Y, de loi respectives μ et ν. La formule $(\dagger)$ montre que $\mu * \nu$ est donc la loi de la somme $X + Y$, et on retrouve ainsi le corollaire. On peut bien sûr itérer le procédé. On retiendra le résultat suivant:

Proposition. *Soient $X_1, \cdots, X_n$, n-v.a. réelles indépendantes de lois respectives $P_{X_1}, \cdots, P_{X_n}$. La loi de $X_1 + \cdots + X_n$ est le produit de convolution $P_{X_1} * \cdots * P_{X_n}$.*

6.5.3 Application au calcul des moments

On suppose dans cette partie que la dimension est $d = 1$. La fonction caractéristique d'une v.a. X permet de calculer très facilement les moments de X, pourvu qu'ils existent.

Proposition. *Supposons que X admet un moment d'ordre $n \in \mathbb{N}$. Alors Φ_X est de classe $\mathcal{C}^n$ et*

$$\imath^n \mathbb{E}(X^n \mathrm{e}^{\mathrm{i}\lambda X}) \;=\; \Phi_X^{(n)}(\lambda)\,, \qquad \lambda \in \mathbb{R}\,.$$

Preuve: Nous allons prouver la formule pour $n = 1$; le cas général en découle par récurrence. On remarque tout d'abord que pour tout $t \in \mathbb{R}$

$$|\sin(t)| \le |t| \quad , \quad 1 - \cos(t) = 2\sin^2(t/2) \le 2 \wedge t^2 \le 2|t|\,.$$

On a pour tout $\lambda \in \mathbb{R}$

$$\frac{\Phi_X(\lambda+\varepsilon) - \Phi_X(\lambda)}{\varepsilon} \;=\; \int_{\mathbb{R}} \left(\frac{\mathrm{e}^{\mathrm{i}(\lambda+\varepsilon)x} - \mathrm{e}^{\mathrm{i}\lambda x}}{\varepsilon} \right) P_X(dx) \;=\; \int_{\mathbb{R}} \left(\frac{\mathrm{e}^{\mathrm{i}\varepsilon x} - 1}{\varepsilon} \right) \mathrm{e}^{\mathrm{i}\lambda x} P_X(dx)\,.$$

On divise par ε. L'intégrand dans le terme de droite est dominé par $3|x|$ et converge vers $\mathrm{i}x$ quand $\varepsilon \to 0$. On applique le théorème de convergence dominée pour obtenir la formule

$$\mathbb{E}(X \mathrm{e}^{\mathrm{i}\lambda X}) \;=\; -\mathrm{i}\Phi_X'(\lambda)\,.$$

L'application $\lambda \to x\mathrm{e}^{\mathrm{i}\lambda x}$ est continue pour chaque x, et est dominée par $|x|$ indépendamment de λ. Par convergence dominée, on a

$$\lim_{\lambda \to \lambda_0} \int_{\mathbb{R}} x \mathrm{e}^{\mathrm{i}\lambda x} P_X(dx) \;=\; \int_{\mathbb{R}} x \mathrm{e}^{\mathrm{i}\lambda_0 x} P_X(dx)\,,$$

ce qui établit la continuité de Φ_X'. ■

N.B. On se gardera de croire que si Φ_X admet une dérivée d'ordre n à l'origine, alors $\mathbb{E}(X^n) = \Phi_X^{(n)}(0)$; il existe des v.a. dont la fonction caractéristique est continument dérivable, et qui n'ont pas de moment d'ordre 1. Par exemple, supposons que

$$P_X(dx) \;=\; c\mathbf{1}_{|x|>1} \frac{dx}{|x|^2 \log|x|}\,,$$

où c est la constante de normalisation. D'une part, on sait que $\int_{-\infty}^{\infty} |x| P_X(dx) = \infty$ (intégrale de Bertrand). D'autre part, par symétrie, on a

$$1 - \Phi_X(\varepsilon) \;=\; c \int_{|x|>1} \left(1 - \mathrm{e}^{\mathrm{i}\varepsilon x}\right) \frac{dx}{|x|^2 \log|x|} \;=\; 2c \int_1^{\infty} (1 - \cos(\varepsilon x)) \frac{dx}{x^2 \log x}\,.$$

En effectuant le changement de variables $\varepsilon x = t$, on trouve

$$\varepsilon^{-1} (1 - \Phi_X(\varepsilon)) \;=\; 2c \int_{\varepsilon}^{\infty} (1 - \cos(t)) \frac{dt}{t^2 (\log t + \log(1/\varepsilon))}\,,$$

et le terme de droite tend vers 0 quand $\varepsilon \to 0$.

Utilisons cette méthode pour calculer les moments d'une variable N de loi $\mathcal{N}(0,1)$. La fonction caractéristique de la loi normale admet le développement en série entière

$$\exp\left\{-\frac{\lambda^2}{2}\right\} \;=\; \sum_{k=0}^{\infty} \left(-\frac{\lambda^2}{2}\right)^k \frac{1}{k!}\,.$$

En identifiant les coefficients de cette série en terme des dérivées successives en zéro, il vient que $\mathbb{E}\left(N^{2k+1}\right) = 0$ (les moments d'ordre impairs sont tous nuls, ce qu'on peut voir directement par symétrie), et les moments d'ordre pairs sont donnés par

$$\mathbb{E}\left(N^{2k}\right) \;=\; \frac{(2k)!}{k!}2^{-k}\,.$$

6.5.4 Convergence en loi et fonctions caractéristiques

On présente enfin un résultat extrèmement utile (qui a motivé en grande partie cette section) pour savoir si une suite de v.a. converge en loi, et déterminer sa limite. On rappelle qu'on note Φ_Y la fonction caractéristique d'une v.a. Y.

Théorème. *On considère comme à l'accoutumée une suite de v.a. $X, X_1, \cdots, X_n, \cdots$ à valeurs dans $\mathbb{R}^d$. Les conditions suivantes sont équivalentes:*

(1) X_n converge en loi vers X.

(2) Φ_{X_n} converge simplement vers Φ_X.

Preuve: (1) $\Rightarrow$ (2) Pour tout $\lambda \in \mathbb{R}^d$, la fonction $x \to \mathrm{e}^{\mathrm{i}\langle\lambda,x\rangle}$ est continue et bornée. On a donc

$$\Phi_{X_n}(\lambda) \;=\; \mathbb{E}\left(\mathrm{e}^{\mathrm{i}\langle\lambda,X_n\rangle}\right) \;\to\; \mathbb{E}\left(\mathrm{e}^{\mathrm{i}\langle\lambda,X\rangle}\right) \;=\; \Phi_X(\lambda)\,.$$

(2) $\Rightarrow$ (1) Pour simplifier, nous allons supposer que la dimension est $d = 1$; le cas général est similair, mais avec des notations plus lourdes. Supposons tout d'abord que f est une fonction continue à support compact, dont la transformée de Fourier $\widehat{f}(\lambda) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\lambda x} f(x)dx$ est dans $L^1(\mathbb{R})$. On sait par Fourier inverse que

$$f(x) \;=\; (2\pi)^{-1}\int_{-\infty}^{\infty} \mathrm{e}^{-i\lambda x}\,\widehat{f}(\lambda)d\lambda\,.$$

On a donc en appliquant le théorème de Fubini (justifié)

$$\begin{aligned}\mathbb{E}\left(f(X_n)\right) \;&=\; (2\pi)^{-1}\mathbb{E}\left(\int_{-\infty}^{\infty} \mathrm{e}^{-i\lambda X_n}\,\widehat{f}(\lambda)d\lambda\right)\\ &=\; (2\pi)^{-1}\int_{-\infty}^{\infty} \mathbb{E}\left(\mathrm{e}^{-i\lambda X_n}\right)\widehat{f}(\lambda)d\lambda \;=\; (2\pi)^{-1}\int_{-\infty}^{\infty} \Phi_{X_n}(-\lambda)\widehat{f}(\lambda)d\lambda\,.\end{aligned}$$

Par hypothèse, on sait que $\Phi_{X_n}(-\lambda)$ converge vers $\Phi_X(-\lambda)$ pour chaque λ. D'autre part, $|\Phi_{X_n}(-\lambda)| \le 1$ pour tout $\lambda \in \mathbb{R}$. Comme $\widehat{f}$ est intégrable, le théorème de convergence dominée s'applique, et on tire

$$(*) \qquad \lim_{n\to\infty} \mathbb{E}\left(f(X_n)\right) == (2\pi)^{-1}\int_{-\infty}^{\infty} \Phi_X(-\lambda)\widehat{f}(\lambda)d\lambda \;=\; \mathbb{E}\left(f(X)\right)\,.$$

Ensuite on remarque que la condition $\widehat{f} \in L^1(\mathbb{R})$ est satisfaite pour toute fonction f de classe $\mathcal{C}^2$ à support compact. En effet, la transformée de Fourier de f'' est $\lambda^2\widehat{f}(\lambda)$ (intégration par parties), et on sait que c'est une fonction bornée. On a donc $\widehat{f}(\lambda) = O(\lambda^{-2})$ à l'infini, et comme $\widehat{f}$ est elle aussi bornée, c'est bien une fonction intégrable. Ainsi, on sait que $(*)$ est vérifié pour toute fonction f de classe $\mathcal{C}^2$ à support compact.

Maintenant, si f est juste une fonction continue à support compact, on sait trouver pour chaque $\varepsilon > 0$ une fonction $g = g_\varepsilon$ de classe $\mathcal{C}^2$ à support compact telle que $|f(t) - g(t)| < \varepsilon/5$ pour tout $t \in \mathbb{R}$ (comme dans la section précedente, il suffit par exemple de prendre pour g la convolée de f avec une approximation de l'unité de classe $\mathcal{C}^2$ à support compact). On peut répéter le même argument que dans la preuve de la proposition sur les fonctions de répartitions pour voir qu'il existe un entier N tel que

$$|\mathbb{E}(f(X_n)) - \mathbb{E}(f(X))| \leq \varepsilon \qquad \text{dès que } n \geq N\,.$$

La preuve du théorème est complète. ■

Une difficulté possible quand on cherche à appliquer le théorème que nous venons de voir, est que l'on doit savoir *a priori* que la limite des fonctions génératrices Φ_{X_n} est elle aussi une fonction caractéristique (ça n'a rien d'automatique). Il y a bien un théorème de Bernstein qui donne une condition nécéssaire et suffisante pour qu'une fonction soit la fonction caractéristique d'une loi de probabilité, mais elle n'est pas toujours facile à vérifier. On admettra le critère suivant très simple:

Critère de Lévy. *Si pour chaque entier n, Φ_n est la fonction caractéristique d'une loi de probabilité sur $\mathbb{R}^d$, et si Φ_n converge simplement vers une fonction Φ continue en 0, alors Φ est la fonction caractéristique d'une loi de probabilité sur $\mathbb{R}^d$.*

Autrement dit, si on sait que la suite des fonctions caractéristiques Φ_{X_n} converge simplement vers une fonction Φ continue en 0, alors X_n converge en loi vers une v.a. X dont la fonction caractéristique est Φ.

6.6 Compactité relative et théorème de Prohorov $(*)$

Le problème qui va nous intéresser maintenant est de savoir si étant donnée une famille de variables aléatoires $(X_i : i \in I)$ à valeurs dans $\mathbb{R}^d$, il est possible d'extraire de n'importe quelle suite issue de cette famille, une sous-suite qui converge en loi. Dans ce cas, on dira que la suite est *relativement compacte* (pour la convergence en loi).

Montrons d'abord sur un exemple simple que ce n'est pas toujours le cas. Prenons $I = \mathbb{N}$ en $X_n = n$. Pour toute fonction $f \in \mathcal{C}_0$ (continue et tendant vers 0 à l'infini), on a $\mathbb{E}(f(X_n)) = f(n) \to 0$, et on voit ainsi qu'il est impossible d'extraire une sous-suite qui converge en loi. La raison intuitive est que 'la masse est partie à l'infini'. Le résultat principal de cette section est que si ce phénomène de perte de masse à l'infini n'est pas possible pour la famille $(X_i : i \in I)$, alors elle est relativement compact, et réciproquement. On introduit d'abord la notion de tension.

Définition. *Soient $(\mu_i : i \in I)$ une famille de lois de probabilités sur $\mathbb{R}^d$. On dit que cette famille est tendue si*

$$\lim_{M \to \infty} \sup_{i \in I} \mu_i \Big(\{x \in \mathbb{R}^d : ||x|| > M\} \Big) = 0\,,$$

c'est-à-dire que pour tout $\varepsilon > 0$, il existe $M > 0$ tel que

$$\mu_i(\{x \in \mathbb{R}^d, ||x|| > M\}) \leq \varepsilon \qquad \textit{pour tout } i \in I.$$

Nous sommes maintenant en mesure d'énoncer le résultat clef.

Théorème de Prohorov. *Une famille de mesures de probabilités sur $\mathbb{R}^d$ est relativement compacte (pour la convergence étroite) si et seulement si elle est tendue.*

Preuve. Partie directe. Supposons que $(\mu_i : i \in I)$ est relativement compacte mais pas tendue. Il existe donc un réel $\varepsilon > 0$ tel que pour tout entier n, on peut trouver un indice $i = i_n$ pour lequel $\mu_{i_n}(||x|| > n) > \varepsilon$. Puisque $(\mu_i : i \in I)$ est relativement compacte, on peut extraire de la suite $(\mu_{i_n} : n \geq 0)$ une sous-suite notée $(\nu_n : n \geq 0)$ qui converge étroitement vers une mesure de probabilité ν sur $\mathbb{R}^d$. On a bien sûr $\nu_n(||x|| > n) \geq \varepsilon$ pour tout n. Il est facile de construire une suite $(f_k : k \in \mathbb{N})$ de fonctions continues bornées par 1, telles que $f_k(x) = 1$ si $||x|| \leq k$ et $f_k(x) = 0$ si $||x|| > k + 1$. On a alors pour tout k

$$\nu(||x|| \leq k) \leq \int_{\mathbb{R}^d} f_k(x)\nu(dx) = \lim_{n\to\infty} \int_{\mathbb{R}^d} f_k(x)\nu_n(dx) \leq 1 - \varepsilon\,.$$

On fait tendre k vers l'infini et on obtient $\nu(\mathbb{R}^d) \leq 1 - \varepsilon$, d'où contradiction.

Partie réciproque. On ne fera la démonstration que pour la dimension $d = 1$. Il n'est pas très difficile d'adapter l'idée en dimension supérieure, mais les notations deviennent plus lourdes. On suppose que $(\mu_i : i \in I)$ est tendue, et on prend une suite notée par abus $(\mu_n : n \in \mathbb{N})$ dans cette famille; elle est bien sûr tendue elle aussi. On désigne par F_n la fonction de répartion de μ_n, de sorte que $\mu_n(dx) = dF_n(x)$.

Pour chaque nombre rationnel q, la suite $(F_n(q) : n \in \mathbb{N})$ est à valeurs dans $[0, 1]$; on peut donc en extraire une sous-suite qui converge. Comme $\mathbb{Q}$ est dénombrable, le procédé d'extraction diagonal fournit une sous-suite notée $(G_n : n \in \mathbb{N})$ qui converge en tout point rationnel vers une limite $G : \mathbb{Q} \to [0, 1]$. Il est clair que G est croissante.

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

$$F(x) = \inf\{G(q) : q \in \mathbb{Q} \text{ et } q > x\}\,.$$

Il est clair que $F : \mathbb{R} \to [0, 1]$ est une fonction croissante, continue à droite. Par un argument de monotonie, on voit que si $x' < x < x''$, alors

$$\limsup_{n\to\infty} G_n(x') \leq F(x) \leq \liminf_{n\to\infty} G_n(x'')\,,$$

de sorte que $\lim_{n\to\infty} G_n(x) = F(x)$ dès que F est continue au point x. Le fait que la suite des mesures $(\mu_n : n \in \mathbb{N})$ soit tendue dit que pour tout $\varepsilon > 0$, on peut trouver x assez grand tel que

$$F_n(-x) \leq \varepsilon \quad \text{et} \quad 1 - F_n(x) \leq \varepsilon \qquad \text{pour tout } n.$$

Il en découle que $\lim_{x\to\infty} F(-x) = 0$ et $\lim_{x\to\infty} 1 - F(x) = 0$; autrement dit F est la fonction de répartition de la mesure de probabilité dF. On a vu que $G_n(x)$

converge vers $F(x)$ en tout point de continuité x de F, et ainsi la suite des mesure de probabilités $(dG_n : n \in \mathbb{N})$ qui est extraite de $(\mu_n : n \in \mathbb{N})$ converge étroitement vers dF. ■

Chapitre 7

Autour du Théorème Central Limite

7.1 Retour sur la loi faible des grands nombres

Nous allons tout d'abord établir la loi faible des grands nombres (i.e. pour la convergence en probabilité, et non pas pour la convergence presque sûre) sous la seule hypothèse de finitude du moment d'ordre 1. Rappelons ce dont il s'agit:

Loi faible des grands nombres. *On se donne $X_1, \cdots, X_n, \cdots$ une suite de v.a. réelles i.i.d., avec $\mathbb{E}(|X|) < \infty$; on note $S_n = X_1 + \cdots + X_n$ la somme partielle au rang n. Alors S_n/n converge en probabilité vers la moyenne $\mathbb{E}(X) = m \in \mathbb{R}$.*

Preuve: Etablissons d'abord la stratégie. Pour cela, on sait qu'il suffit de montrer la convergence en loi (puisque la limite est une constante). A cette fin, on va calculer la fonction caractéristique de S_n/n et vérifier qu'elle converge bien vers la fonction caractéristique de la constante m.

Notons $\Phi_X(\lambda) = \mathbb{E}\left(\mathrm{e}^{\mathrm{i}\lambda X}\right)$ $(\lambda \in \mathbb{R})$ la fonction caractéristique de X. Comme les v.a. $X_1, \cdots, X_n$ sont indépendantes et toutes de même loi que X, la fonction caractéristique de S_n est $\Phi_{S_n} = \Phi_X^n$. On a donc

$$\Phi_{S_n/n}(\lambda) \;=\; \mathbb{E}\left(\mathrm{e}^{\mathrm{i}\lambda S_n/n}\right) \;=\; \Phi_{S_n}(\lambda/n) \;=\; \Phi_X(\lambda/n)^n\,.$$

On veut étudier le comportement du terme de droite quand $n \to \infty$. Pour cela, on le ré-écrit comme $(1-(1-\Phi_X(\lambda/n)))^n$, et on se souvient de ce que

$$\Phi_X(0) \;=\; 1 \quad , \quad \Phi'_X(0) \;=\; \mathrm{i}\mathbb{E}(X)\,.$$

On sait donc que $1-\Phi_X(\lambda/n) \sim -\mathrm{i}\lambda\mathbb{E}(X)/n$ quand $n \to \infty$. Il en découle que

$$\log \Phi_{S_n/n}(\lambda) \;=\; n \log\left(1-(1-\Phi_X(\lambda/n))\right) \;\sim\; \mathrm{i}\lambda\mathbb{E}(X)\,.$$

En prenant l'exponentielle, on tire finalement

$$\Phi_{S_n/n}(\lambda) \;\sim\; \mathrm{e}^{\mathrm{i}\lambda\mathbb{E}(X)}\,.$$

Or le terme de droite est la fonction caractéristique de la v.a. constante $\mathbb{E}(X)$; c'est précisement ce qu'on cherchait à établir. ■

C'est cette même idée qui va servir à établir un résultat du second ordre pour la moyenne de Césaro, le théorème central limite.

7.2 Le théorème central limite unidimensionel

Le théorème central limite est l'un des (si ce n'est le) résultats les plus importants de la théorie des probabilités. De façon informelle, ce théorème donne une estimation très précise de l'erreur qu'on commet en approchant la moyenne mathématique par la moyenne empirique (i.e. la moyenne de Césaro).

Il a d'abord été observé par Gauss, qui l'appelait la loi des erreurs; mais ce dernier n'en a pas donné de démonstration rigoureuse. La preuve en a été apportée par Moivre et Laplace, le théorème porte parfois leurs noms. La dénomination actuelle est apparue vers 1950 (en anglais: central limit theorem, ce qu'on a parfois traduit à tort par "le théorème de la limite centrale").

Théorème Central Limite. *Soient $X_1, \cdots, X_n, \cdots$ une suite de v.a. réelles i.i.d., avec $\mathbb{E}(X^2) < \infty$. On note $S_n = X_1 + \cdots + X_n$ la somme partielle au rang n, $m = \mathbb{E}(X)$ la moyenne et $\sigma^2 = \mathrm{Var}(X) = \mathbb{E}(X^2) - m^2$. Alors*

$$\lim_{n\to\infty} \frac{S_n - n\mathbb{E}(X)}{\sqrt{n}} \;=\; \mathcal{N}(0,\sigma^2) \quad \textit{en loi},$$

où on a noté $\mathcal{N}(0,\sigma^2)$ la loi de Gauss centrée et de variance σ^2, i.e. la loi sur $\mathbb{R}$ de densité

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\}.$$

Preuve: On va adopter la même approche que dans la preuve de la loi faible des grands nombres, mais au second ordre au lieu du premier ordre. Quitte à remplacer X par $X - m$, on peut supposer que $\mathbb{E}(X) = 0$.

Rappelons que la fonction caractéristique de $S_n/\sqrt{n}$ est donnée par

$$\Phi_{S_n/\sqrt{n}}(\lambda) \;=\; \Phi_X(\lambda/\sqrt{n})^n\,.$$

D'autre part, comme X a un moment d'ordre 2, sa fonction caractéristique est de classe $\mathcal{C}^2$, et son developpement de Taylor à l'origine est donné par

$$\Phi_X(\lambda) \;=\; 1 + \lambda\Phi_X'(0) + \frac{\lambda^2}{2}\Phi_X''(0) + o(\lambda^2)\,.$$

Le fait que $\mathbb{E}(X) = 0$ et $\mathbb{E}(X^2) = \sigma^2$ entraine que $\Phi_X'(0) = 0$ et $\Phi_X''(0) = -\sigma^2$. Autrement dit, on a

$$\Phi_X(\lambda) \;=\; 1 - \frac{\sigma^2\lambda^2}{2} + o(\lambda^2)\,, \quad (\lambda \to 0)\,.$$

Il en découle que pour chaque $\lambda \in \mathbb{R}$ fixé

$$n \log \Phi_X(\lambda/\sqrt{n}) \;\sim -n\left(\frac{\sigma^2\lambda^2}{2n}\right) \;=\; -\frac{\sigma^2\lambda^2}{2} \qquad (n \to \infty)\,,$$

et donc, en passant à l'exponentielle

$$\Phi_{S_n/\sqrt{n}}(\lambda) \sim \exp\left\{-\frac{\sigma^2\lambda^2}{2}\right\}.$$

Le terme de droite est la fonction caractéristique de la loi $\mathcal{N}(0,\sigma^2)$, et le théorème est établi. ■

On observera que les conditions d'application du théorème central limite sont plus restrictives que celles pour la loi des grands nombres (finitude du moment d'ordre 2 pour le premier, et seulement du moment d'ordre 1 pour le premier). Bien sûr, les conclusions du théorèmes peuvent être mises en défaut si les hypothèses ne sont pas remplies. Par r exemple, si $X_1, \ldots$ sont des variables indépendantes toutes de loi de Cauchy standard, alors la moyenne empirique $(X_1 + \cdots + X_n)/n$ est elle aussi une variable de Cauchy pour tous les indices n.

Voici un exemple d'application amusant: nous allons démontrer la formule de Stirling à l'aide du théorème central limite. Prenons pour X une loi de Poisson de paramètre 1, de sorte que S_n suit une loi de Poisson de paramètre n (puisque la fonction génératrice de la loi de Poisson de paramètre n est $s \to \mathrm{e}^{-n(1-s)}$). Rappelons que la moyenne de X est 1, ainsi que sa variance. Posons $T_n = (S_n - n)/\sqrt{n}$ et notons N une v.a. normale standard. On sait donc d'après le théorème central limite et la caractérisation de la convergence en loi en terme des fonctions de répartition que pour tout $x > 0$

$$(\dagger) \qquad \lim_{n\to\infty} \mathbb{P}(x \le T_n) = \mathbb{P}(x \le N).$$

D'autre part, on a pour tout entier n d'après l'inégalité de Bienaymé-Chebytchev

$$\mathbb{P}(x \le T_n) = \mathbb{P}\left(S_n - n > x\sqrt{n}\right) \le \frac{1}{nx^2}\mathrm{Var}(S_n) = x^{-2}.$$

Comme $\int_0^\infty (1 \wedge x^{-2})dx < \infty$, nous sommes en droit d'appliquer le théorème de convergence dominée dans (†), ce qui donne

$$\mathbb{E}(T_n^+) = \int_0^\infty \mathbb{P}(x \le T_n)dx \to \int_0^\infty \mathbb{P}(x \le N)dx = \mathbb{E}(N^+),$$

où on a noté x^+ la partie positive de x. Le terme de droite vaut

$$\frac{1}{\sqrt{2\pi}}\int_0^\infty x\exp\{-x^2/2\}dx = \frac{1}{\sqrt{2\pi}}.$$

Quant au terme de gauche, on peut l'écrire comme

$$\begin{aligned}
\mathrm{e}^{-n}\sum_{j=n+1}^{\infty}\frac{n^j}{j!}\left(\frac{j-n}{\sqrt{n}}\right) &= \frac{\mathrm{e}^{-n}}{\sqrt{n}}\lim_{k\to\infty}\sum_{j=n+1}^{k}\left[\frac{n^j}{(j-1)!} - \frac{n^{j+1}}{j!}\right] \\
&= \frac{\mathrm{e}^{-n}}{\sqrt{n}}\lim_{k\to\infty}\left[\frac{n^{n+1}}{n!} - \frac{n^{k+1}}{k!}\right] \\
&= \frac{\mathrm{e}^{-n}}{\sqrt{n}}\frac{n^{n+1}}{n!} = \frac{\mathrm{e}^{-n}n^n\sqrt{n}}{n!}.
\end{aligned}$$

En remettant les pièces en place, on a donc démontré la formule de Stirling

$$\frac{\mathrm{e}^{-n}n^n\sqrt{n}}{n!} \sim \frac{1}{\sqrt{2\pi}}.$$

7.3 Vitesse de convergence dans le théorème central limite $(*)$

Nous allons maintenant nous pencher sur la vitesse de convergence dans le théorème central limite, c'est-à-dire étudier la vitesse à laquelle $\mathbb{E}[f((S_n - nm)/\sqrt{n})]$ approche $\mathbb{E}[f(\mathcal{N}(0,\sigma^2))]$ pour une fonction f assez régulière. Nous allons prendre des hypothèses plus fortes que précédemment, ce qui nous permettra de donner des arguments très simples.

Proposition. *Soit $X_1, \cdots, X_n$ une suite de v.a. réelles i.i.d. On suppose que $\mathbb{E}(|X|^3) < \infty$. On note $\mathbb{E}(X) = m$ et $\sigma^2 = \mathbb{E}(X^2)$, $T_n = (X_1 + \cdots + X_n - mn)/\sqrt{n}$, et N une v.a. de loi $\mathcal{N}(0,\sigma^2)$.*

Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction de classe $\mathcal{C}^3$, dont la dérivée troisième est bornée,i.e. $|f'''(x)| \leq c$ pour tout $x \in \mathbb{R}$. On a alors

$$\mathbb{E}(f(T_n) \; - \; \mathbb{E}(f(N)) \; = \; O(1/\sqrt{n}) \, .$$

Preuve: Sans perdre de généralité, on supposera que $m = 0$. On introduit une suite de $N_1, \cdots, N_n, \cdots$ de v.a. i.i.d. de loi $\mathcal{N}(0,\sigma^2)$ et on pose $\overline{T}_n = (N_1 + \cdots + N_n)/\sqrt{n}$, de sorte que pour chaque n, $\overline{T}_n$ a encore pour loi $\mathcal{N}(0,\sigma^2)$.

Pour chaque $k = 1, \cdots, n$, on considère

$$\begin{aligned} T_n^k &= \frac{X_1 + \cdots + X_k + N_{k+1} + \cdots + N_n}{\sqrt{n}} \, , \\ \widehat{T}_n^k &= \frac{X_1 + \cdots + X_{k-1} + N_{k+1} + \cdots + N_n}{\sqrt{n}} \, . \end{aligned}$$

On cherche donc à estimer

$$\begin{aligned} & \left| \mathbb{E}\left(f(T_n)\right) - \mathbb{E}\left(f(\overline{T}_n)\right) \right| \\ = & \left| \sum_{k=0}^{n-1} \mathbb{E}\left(f(T_n^k)\right) - \mathbb{E}\left(f(T_n^{k+1})\right) \right| \\ \leq & \sum_{k=0}^{n-1} \left| \mathbb{E}\left(f(T_n^k) - f(\widehat{T}_n^k) + f(\widehat{T}_n^k) - f(T_n^{k+1})\right) \right| \, . \end{aligned}$$

Pour évaluer chaque différence, on écrit le développement de Taylor de f au point $\widehat{T}_n^k$, en notant que $T_n^k = \widehat{T}_n^k + X_k/\sqrt{n}$ et $T_n^{k+1} = \widehat{T}_n^k + N_k/\sqrt{n}$:

$$\begin{aligned} f(T_n^k) - f(\widehat{T}_n^k) &= f'\left(\widehat{T}_n^k\right) \frac{X_k}{\sqrt{n}} + \frac{1}{2} f''\left(\widehat{T}_n^k\right) \left(\frac{X_k}{\sqrt{n}}\right)^2 + R_n^k \\ f(T_n^{k+1}) - f(\widehat{T}_n^k) &= f'\left(\widehat{T}_n^k\right) \frac{N_k}{\sqrt{n}} + \frac{1}{2} f''\left(\widehat{T}_n^k\right) \left(\frac{N_k}{\sqrt{n}}\right)^2 + \widehat{R}_n^k \end{aligned}$$

La clef consiste à observer que $\widehat{T}_n^k$ est indépendant de N_k et de X_k, et que, par hypothèse, ces deux dernieres v.a. ont la même moyenne et la même variance. Quand

on prend l'espérance, tout s'annule sauf les restes, $\mathbb{E}\left(R_n^k\right)$ et $\mathbb{E}\left(\widehat{R}_n^k\right)$. Or

$$|R_n^k| \leq ||f'''||_\infty |X_k|^3 n^{-3/2} \quad \text{et} \quad |\widehat{R}_n^k| \leq ||f'''||_\infty |N_k|^3 n^{-3/2}\,.$$

Au bout du compte, on a donc

$$\left|\mathbb{E}\left(f(T_n)\right) - \mathbb{E}\left(f(\overline{T}_n)\right)\right| \;\leq\; c\sum_{k=0}^{n-1} n^{-3/2} \;=\; cn^{-1/2}\,.$$

7.4 Variables gaussiennes multi-dimensionnelles

On rappelle que la loi de Gauss de moyenne $m \in \mathbb{R}$ et de variance $\sigma^2 > 0$ a pour densité

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}, \qquad x \in \mathbb{R}\,.$$

Il est commode de convenir qu'une masse de Dirac δ_m est une loi de Gauss de moyenne m et de variance nulle.

On considère une v.a. $X = (X^{(1)}, \cdots, X^{(d)})$ à valeurs dans $\mathbb{R}^d$. On dit que X est un vecteur gaussien si toute combinaison linéaire de ses coordonnées (c'est-à-dire $\langle \lambda, X\rangle = \lambda^{(1)}X^{(1)} + \cdots + \lambda^{(d)}X^{(d)}$ pour $\lambda = (\lambda^{(1)}, \cdots, \lambda^{(d)}) \in \mathbb{R}^d$) suit une loi de Gauss. Par exemple, un vecteur aléatoire dont les coordonnées sont des variables gaussiennes indépendantes est un vecteur gaussien (car une combinaison linéaire de varibales aléatoires gaussiennes indépendantes suit encore une loi gaussienne.

Bien sûr, si $X = (X^{(1)}, \cdots, X^{(d)})$ est guassien, chaque coordonnée $X^{(i)}$ est une v.a. gaussienne réelle; mais on fera bien attention qu'à l'inverse cette dernière condition n'est pas suffisante pour assurer que X est un vecteur gaussien. Par exemple, si N est une variable aléatoire réelle de loi $\mathcal{N}(0,1)$ et ε une variable indépendante de loi de Bernoulli, $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$, il est facile de vérifier que εN suit encore la loi $\mathcal{N}(0,1)$. En revanche, le couple $(N, \varepsilon N)$ n'est pas gaussien (par exemple, on voit que $\mathbb{P}(N + \varepsilon N = 0) = 1/2$.

Définition. (Vecteur moyenne, matrice de covariance) *Soit X un vecteur gaussien dans $\mathbb{R}^d$. Le vecteur $m = (m^{(1)}, \cdots, m^{(d)})$ donné par $m^{(j)} = \mathbb{E}(X^{(j)})$ s'appelle la moyenne de X.*

La matrice $d \times d$ donnée par

$$D \;=\; \left(D_{i,j} \;=\; \mathbb{E}(X^{(i)}X^{(j)}) - \mathbb{E}(X^{(i)})\mathbb{E}(X^{(j)}) : i,j = 1, \cdots d\right)$$

s'appelle la covariance, ou la matrice de dispersion.

On dira que X est centré si son vecteur moyenne m est nul. Il est immédiat de vérifier que $X - m$ est un vecteur gaussien centré, qui a la même matrice de dispersion que X. On montre sans peine que si $X = \left(X^{(1)}, \ldots, X^{(\mathtt{d})}\right)$ et $Y = \left(Y^{(1)}, \ldots, Y^{(\mathtt{d})}\right)$ sont deux vecteurs gaussiens indépendants, alors leur somme $X + Y$ est encore un vecteur gaussien, de moyenne la somme des vecteurs moyenne, et de matrice de dispersion la somme des matrices de dispersion de X et Y.

On observe d'abord le résultat élémentaire suivant:

Proposition. *Pour tout $\lambda \in \mathbb{R}^d$, la v.a. réelle $\langle \lambda, X\rangle$ a pour moyenne $m_\lambda = \langle \lambda, m\rangle$ et pour variance $\sigma_\lambda^2 = \langle \lambda, D\lambda\rangle$.*

En conséquence, D est une matrice symétrique positive, c'est-à-dire $D = D^t$ et $\langle \lambda, D\lambda\rangle \geq 0$ pour tout $\lambda \in \mathbb{R}^d$.

Preuve: Le fait que $m_\lambda = \langle \lambda, m\rangle$ découle de la linéarité de l'espérance.

On a

$$\mathbb{E}(\langle \lambda, X\rangle^2) \;=\; \sum_{i,j} \lambda^{(i)}\lambda^{(j)} \mathbb{E}\left(X^{(i)}X^{(j)}\right) \quad \text{et} \quad \langle \lambda, m\rangle^2 \;=\; \sum_{i,j} \lambda^{(i)}\lambda^{(j)} \mathbb{E}(X^{(i)})\mathbb{E}(X^{(j)})\,,$$

d'où on tire

$$\mathbb{E}(\langle \lambda, X\rangle^2) - \langle \lambda, m\rangle^2 \;=\; \sum_{i,j} \lambda^{(i)} D_{i,j} \lambda^{(j)} \;=\; \langle \lambda, D\lambda\rangle\,.$$

La seconde assertion est évidente. ■

Corollaire. *La fonction caractéristique d'un vecteur gaussien X est donnée par*

$$\Phi_X(\lambda) \;=\; \exp\left\{\mathrm{i}\langle \lambda, m\rangle - \langle \lambda, D\lambda\rangle/2\right\}\,, \qquad \lambda \in \mathbb{R}^d\,.$$

En conséquence, la loi d'un vecteur gaussien est complètement déterminée par sa moyenne et sa matrice de dispersion; on notera $\mathcal{N}(m, D)$.

Preuve: Pour chaque $\lambda \in \mathbb{R}^d$, on sait que $\langle \lambda, X\rangle$ est une v.a. gaussienne réelle de moyenne m_λ et de variance σ_λ^2. Sa fonction caractéristique est donc $x \to \exp\{\mathrm{i}xm_\lambda - x^2\sigma_\lambda^2/2\}$ $(x \in \mathbb{R})$. En prenant $x = 1$, on a donc d'après la proposition précédente:

$$\Phi_X(\lambda) \;=\; \mathbb{E}\left(\mathrm{e}^{\mathrm{i}\langle \lambda, X\rangle}\right) \;=\; \exp\left\{\mathrm{i}\langle \lambda, m\rangle - \langle \lambda, D\lambda\rangle/2\right\}\,.$$

Nous allons maintenant montrer comment construire un vecteur gaussien de moyenne et de covariance donnée.

Proposition. *Soit $m \in \mathbb{R}^d$ un vecteur et D une matrice $d \times d$ symétrique positive. Alors, il existe un vecteur gaussien d-dimensionnel de moyenne m et de matrice de dispersion D.*

Preuve: On sait construire d variables normales (standard) indépendantes, $N^{(1)}, \cdots, N^{(d)}$. On note $N = (N^{(1)}, \cdots, N^{(d)})$. D'autre part, toute matrice symétrique positive admet une racine carrée, i.e. il existe une matrice symétrique $C = (C_{i,j})$ telle que $C \cdot C = D$. Posons $X = C \cdot N + m$. Pour tout $\lambda \in \mathbb{R}^d$, on a alors grâce à la symétrie de C

$$\langle \lambda, X\rangle \;=\; \langle C\lambda, N\rangle + \langle \lambda, m\rangle\,.$$

Comme toute combinaison de v.a. normales indépendantes est une v.a. gaussienne, on voit que X est un vecteur gaussien. La moyenne de $\langle \lambda, X\rangle$ est clairement $\langle \lambda, m\rangle$. D'autre part, sa variance vaut

$$\mathbb{E}\left(\langle C\lambda, N\rangle^2\right) \;=\; \mathbb{E}\left(\left(\sum_{i,j} C_{i,j}\lambda^{(i)}N^{(j)}\right)\left(\sum_{k,\ell} C_{k,\ell}\lambda^{(k)}N^{(\ell)}\right)\right)$$

$$
\begin{aligned}
&= \mathbb{E}\left(\sum_{i,j,k,\ell} C_{i,j}\lambda^{(i)}N^{(j)}C_{k,\ell}\lambda^{(k)}N^{(\ell)}\right)\\
&= \sum_{i,j,k} C_{i,j}\lambda^{(i)}C_{k,j}\lambda^{(k)} \;=\; \langle C\lambda, C\lambda\rangle \;=\; \langle \lambda, D\lambda\rangle\,,
\end{aligned}
$$

où le passage à la dernière ligne vient de ce que $\mathbb{E}(N^{(j)}N^{(\ell)}) = \delta_{j,\ell}$. Ceci montre que X est bien un vecteur gaussien de moyenne m et de matrice de dispersion D. ■

Corollaire. (Densité gaussienne) *Soit m un vecteur quelconque de $\mathbb{R}^d$ et D une matrice symétrique $d \times d$. Si le determinant de la matrice de dispersion, $det(D)$ est non nul, et si D^{-1} désigne l'inverse D, alors la loi gaussienne d-dimensionnelle $\mathcal{N}(m, D)$ est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$, avec pour densité*

$$
(2\pi)^{-d/2} \cdot det(D)^{-1/2} \exp\left\{-\frac{1}{2}\langle (x-m), D^{-1}(x-m)\rangle\right\}\,.
$$

Preuve: On effectue un changement de variables dans la construction précédente. ■

Corollaire. *Soit $X = \left(X^{(1)}, \cdots, X^{(d)}\right)$ un vecteur gaussien. Pour que les v.a. $X^{(1)}, \cdots, X^{(d)}$ soient indépendantes, il faut et il suffit que la matrice de dispersion de X soit diagonale.*

Preuve: La condition est trivialement nécessaire. Pour la réciproque, on utilise la construction donnée dans la preuve de la proposition précédente si la matrice de dispersion D est inversible, et on voit que la densité de la loi de X s'exprime comme le produit des densités des marginales. Les coordonnées sont donc indépendantes.

Dans le cas général où D n'est pas nécessairement inversible, on peut faire le même raisonnement avec la fonction caractéristique Φ_X. ■

En particulier, si un couple (X, X') de v.a. réelles est gaussien, alors X et X' sont indépendants si et seulement si $\mathbb{E}(XX') = \mathbb{E}(X)\mathbb{E}(X')$. On se gardera bien de croire que si *individuellement* X et X' sont des v.a. gaussiennes telles que $\mathbb{E}(XX') = \mathbb{E}(X)\mathbb{E}(X')$, alors X et X' sont indépendantes. Pour qu'il en soit ainsi, il est nécessaire que le *couple* (X, X') soit un vecteur gaussien, ce qui n'a rien d'automatique. Par exemple, si on reprend l'exemple où N est une variable aléatoire réelle de loi $\mathcal{N}(0,1)$ et ε une variable indépendante de loi de Bernoulli, $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$, on a $\mathbb{E}(N\varepsilon N) = 0$, mais les variables N et εN ne sont pas indépendantes (bien sûr, on a vu que $(N, \varepsilon N)$ n'est pas un vecteur gaussien).

7.5 Le théorème central limite multi-dimensionnel

Soit $X = (X^{(1)}, \cdots, X^{(d)})$ une v.a. à valeurs dans $\mathbb{R}^d$. On suppose que $\mathbb{E}\left(||X||^2\right) < \infty$. On appelle le vecteur

$$m \;=\; (m^{(1)}, \cdots, m^{(d)})\,, \quad m^{(i)} \;=\; \mathbb{E}(X^{(i)})$$

la moyenne de X. La matrice $d \times d$, $D_X = (D_{i,j})$ donnée par

$$D_{i,j} \;=\; \mathbb{E}(X^{(i)}X^{(j)}) - \mathbb{E}(X^{(i)})\mathbb{E}(X^{(j)})$$

est appelée la matrice de covariance, ou de dispersion de X. Le même calcul que dans la première partie montre que D est une matrice symétrique positive. En conséquence, il existe une unique loi sur $\mathbb{R}^d$ qui soit gaussienne de moyenne nulle et de matrice de dispersion D; on la notera $\mathcal{N}(0, D)$.

On sait d'après le théorème central limite uni-dimensionnel que pour chaque coordonnée j, si on se donne une suite de v.a. $X_1^{(j)}, \cdots, X_n^{(j)}, \cdots$ indépendantes et de même loi que $X^{(j)}$, alors

$$\frac{X_1^{(j)} + \cdots + X_n^{(j)} - nm^{(j)}}{\sqrt{n}} \to \mathcal{N}(0, D_{j,j}) \qquad \text{(en loi)}\,.$$

En revanche, le théorème central limite uni-dimensionnel ne permet pas de conclure quant à la convergence des vecteurs aléatoires (convergence conjointe des coordonnées). Ce point fait l'objet la version multidimensionnelle du théorème central limite:

Théorème. *Soient $X_1, \cdots, X_n, \cdots$ une suite de vecteurs indépendants, ayant tous la même loi que X. Alors*

$$\frac{X_1 + \cdots + X_n - nm}{\sqrt{n}}$$

converge en loi quand $n \to \infty$ vers $\mathcal{N}(0, D)$.

Preuve: La preuve du théorème central limite multidimensionnel est très proche de celle que nous avons donnée en dimension 1. Nous esquisserons donc juste les lignes principales.

Première étape: Recentrage. Quitte à travailler avec $X - m$ au lieu de X, on peut supposer que X est centré, i.e. $m = 0$. Les coefficients de la matrices de dispersion de X sont alors simplement donnés par $D_{i,j} = \mathbb{E}\left(X^{(i)}X^{(j)}\right)$.

Deuxième étape: Estimation de la fonction caractéristique Φ_X de X. On écrit tout d'abord

$$\Phi_X(\lambda) \;=\; \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\langle\lambda, x\rangle} P_X(dx)\,, \qquad \lambda \in \mathbb{R}^d\,.$$

L'hypothèse que X admet un moment d'ordre 2 nous permet de 'dériver sous le signe intégral' et on obtient que

$$\frac{\partial^2 \Phi_X(\lambda)}{\partial\lambda^{(i)}\partial\lambda^{(j)}} \;=\; -\int_{\mathbb{R}^d} x^{(i)}x^{(j)}\mathrm{e}^{\mathrm{i}\langle\lambda, x\rangle} P_X(dx).$$

D'une part, on voit aisément par convergence dominée que le terme de droite est une fonction continue en λ, autrement dit Φ_X est une fonction de classe $\mathcal{C}^2$. D'autre part, la dérivée partielle seconde en $\lambda = 0$ vaut $-D_{i,j}$, autrement dit, la différentielle

seconde de Φ_X en $\lambda = 0$ est l'opposé de la matrice de dispersion D. On a donc le développement de Taylor

$$\Phi_X(\lambda) \;=\; 1 - \frac{1}{2}\langle \lambda, D\lambda \rangle + o(|\lambda|^2)\,, \qquad (|\lambda| \to 0).$$

Troisième étape: Calcul de la fonction caractéristique de $S_n/\sqrt{n}$. En notant $S_n = X_1 + \cdots + X_n$ et en utilisant le fait que les X_k sont i.i.d., on tire

$$\Phi_{S_n}(\lambda) \;=\; \Phi_X(\lambda)^n \quad , \quad \Phi_{S_n/\sqrt{n}}(\lambda) \;=\; \Phi_X(\lambda/\sqrt{n})^n\,.$$

En utilisant l'estimation précédente pour Φ_X, on déduit sans peine que

$$\lim_{n\to\infty} \Phi_{S_n/\sqrt{n}}(\lambda) \;=\; \exp\left\{-\frac{1}{2}\langle \lambda, D\lambda \rangle\right\} \quad \forall \lambda \in \mathbb{R}^d\,.$$

Dernière étape: On sait que la matrice de dispersion est symétrique positive (même calcul que celui effectué dans la preuve de la première proposition). Ainsi, le terme de droite dans la limite ci-dessus est bien la fonction caractéristique d'un vecteur gaussien centré de matrice de covariance D. Il ne reste qu'à appliquer le critère de convergence en loi avec les fonctions caractéristiques. ■

Chapitre 8

Quelques notions de Statistique

8.1 Introduction

Jusqu'à présent, nous avons beaucoup étudié dans ce cours des suites de variables i.i.d. dont nous connaissions la loi. La théorie statistique s'intéresse au problème inverse: on observe une suite de valeurs dont on sait qu'elle est donnée par une suite de variables aléatoires i.i.d. (on parle alors d'un échantillon) mais dont on ignore la loi, et à partir de l'échantillon, on voudrait soit estimer la loi inconnue, soit prendre décider si on faut accepter ou rejeter une hyptohèse la concernant.

La première question qu'il est naturel de se poser est de savoir si effectivement on peut espérer obtenir des informations précises sur une loi de probabilité inconnue à l'aide d'un échantillon de cette loi. Le raffinement suivant de la loi des grands nombres nous dit que c'est bien le cas.

Théorème de Glivenko-Cantelli *Soit $X_1, \ldots$ une suite de variables aléatoires réelles indépendantes, toutes de même loi dont on notera F la fonction de répartition.*

On appelle **fonction de répartition empirique** *de rang $n \geq 1$ la fonction aléatoire en la variable réelle x donnée par*

$$F_n(x) \;=\; \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{X_k \leq x\}}.$$

Alors, avec probabilité un, la suite des fonctions F_n converge vers F, uniformément sur $\mathbb{R}$.

En conséquence, on voit que la mesure empirique

$$dF_n(x) \;=\; \frac{1}{n} \sum_{k=1}^{n} \delta_{\{X_k\}}(dx)\,,$$

qui est une mesure de probabilités aléatoire sur $\mathbb{R}$, converge p.s. vers la loi dF qui nous intéresse.

Preuve: Pour simplifier, on n'établira que la convergence simple p.s. Soit Δ l'ensemble des points de discontinuité de la fonction de répartition F; on sait que Δ est au plus dénombrable. On note $D = \Delta \cup \mathbb{Q}$, qui est lui aussi un ensemble dénombrable, et de plus est dense dans $\mathbb{R}$.

Fixons d'abord $x \in D$. Les variables aléatoires $\mathbf{1}_{\{X_k \leq x\}}$ pour $k = 1, \ldots$ forment une suite de variables indépendantes, toutes de loi de Bernoulli de paramètre $F(x)$. La loi forte des grands nombres énonce donc que p.s., on a $\lim_{n\to\infty} F_n(x) = F(x)$. Comme D est dénombrable, l'évènement

$$\{F_n(x) \text{ converge vers } F(x) \text{ pour tout } x \in D\}$$

a donc une probabilité un.

On travaille maintenant sur l'évènement précédent et on considère $y \in \mathbb{R}$, un réel arbitraire. Si $y \in D$, il n'y a plus rien à démontrer. Sinon, on peut trouver une suite croissante $(y_k, k \in \mathbb{N})$ et une suite décroissante $(y'_k, k \in \mathbb{N})$ de réels dans D tels que $\lim y_k = \lim y'_k = y$. On a pour tout entier k fixé $F_n(y_k) \leq F_n(y) \leq F_n(y'_k)$, ce qui conduit en faisant d'abord tendre n vers l'infini à

$$F(y_k) \leq \liminf_{n\to\infty} F_n(y) \leq \limsup_{n\to\infty} F_n(y) \leq F(y'_k)\,.$$

On fait ensuite tendre k vers ∞ pour déduire que

$$F(y-) \leq \liminf_{n\to\infty} F_n(y) \leq \limsup_{n\to\infty} F_n(y) \leq F(y)\,.$$

Comme on n'a supposé que $y \notin D$, a fortiori $y \notin \Delta$, de sorte que $F(y-) = F(y)$ et le théorème est démontré. ■

La solution donnée par le Théorème de Glivenko-Cantelli est très générale dans la mesure où on n'a fait aucune hypothèse quant à la loi à estimer. Le prix à payer est que pour que la fonction de répartition empirique F_n soit une "bonne approximation" de la fonction de répartition F, on a besoin concrètement de prendre l'entier n grand (ce qui se révèle coûteux).

En pratique, il arrive souvent que l'on sache a priori que la loi cherchée appartient à une certaine famille de lois de probabilités dépendant d'un paramètre θ, ce qui permet d'utiliser des approches bien plus efficaces. Donnons un exemple typique très simple. Une société voudrait commercialiser un nouveau produit. Avant de lancer la fabrication, la société a besoin d'estimer la proportion $\theta \in [0, 1]$ de la population susceptible d'acheter ce produit, et pour cela, elle commande un sondage. Supposons pour simplifier qu'on choisit au hasard dans la population 1 000 individus à qui on demande s'ils acheteront ce produit. On notera 1 pour la réponse "oui" et 0 pour "non" et X_i la réponse de la i-ème personne interrogée. Pourvu que la population totale soit très supérieure à 1000, on peut supposer que la suite des réponses $X_1, X_2, \ldots, X_{1000}$ sont des variables i.i.d., toutes de loi de Bernoulli de paramètre θ. Les variables $X_1, \ldots, X_{1000}$ forment un échantillon de taille 1000 de la loi de Bernoulli de paramètre θ, paramètre inconnu qui nous intéresse. Il est intuitivement évident qu'une bonne estimation de θ est donnée par la moyenne empirique $(X_1 + \cdots + X_n)/1000$. Bien entendu, donner une estimation du paramètre θ ne suffit pas à la société qui voudrait savoir si on peut avoir confiance en l'estimation donnée. L'institut de sondage doit donner un "intervalle de confiance" dans lequel se trouve le vrai paramètre θ avec un très grande probabilité (souvent 0,95). Bien évidemment, la société demande un intervalle de confiance le plus petit possible, dans lequel se trouve le vrai paramètre

avec une probabilité la plus grande possible, et pour cela il faut que la taille de l'échantillon soit assez grande. De façon plus prosaïque, la société doit prendre la décision de lancer la fabrication du produit si le paramètre θ est plus grand qu'une valeur minimale θ_0; on doit donc tester l'hypothèse $\theta > \theta_0$ sans nécessairement avoir à connaitre la valeur précise de θ.

Nous allons maintenant formaliser ce qui précède en introduisant les notions de base de la théorie de l'estimation de paramètre.

8.2 Estimation

(a) Premières définitions

On considère un *modèle statistique*, c'est-à-dire un espace Ω muni d'une tribu $\mathcal{F}$, et une famille de lois de probabilités $(P_\theta)_{\theta\in\Theta}$. On appelle Θ l'espace des paramètres. Dans ce cours, on supposera toujours pour simplifier que Θ est une partie de $\mathbb{R}^d$. Par exemple, on peut prendre $\Theta = [0,1]$ et noter P_θ la loi de Bernoulli de paramètre θ, ou $\Theta =]0,\infty[$ et noter P_θ la loi exponentielle de paramètre $1/\theta$, c'est-à-dire de moyenne θ. Un troisième exemple classique est celui du modèle gaussien (réel pour simplifier) qui correspond à $\Theta = \mathbb{R}\times\mathbb{R}_+$ et pour $\theta = (m,\sigma^2)$, $P_\theta = \mathcal{N}(m,\sigma^2)$.

On appelle *échantillon de taille n* d'une loi de probabilités P, une suite $X_1,\ldots,X_n$ de variables aléatoires indépendantes toutes de loi P. Enfin, un *estimateur* est une application d à valeurs dans l'espace des paramètres Θ qui dépend de l'échantillon, i.e. de la forme

$$d(X_1,\ldots,X_n)\,.$$

On parlera d'estimateur *sans biais* lorsque

$$E_\theta\left(d(X_1,\ldots,X_n)\right) \;=\; \theta$$

pour toutes les valeurs possibles du paramètre $\theta\in\Theta$. Il est implicitement supposé dans cette définition que les lois P_θ admettent un moment d'ordre un fini.

Voyons deux premiers examples parmi les plus usuels d'estimateurs.

• Si l'échantillon est à valeurs réelles, la *moyenne empirique*

$$\overline{X}_n \;=\; \frac{X_1+\ldots+X_n}{n}$$

est un estimateur classique de la moyenne. Il est immédiat que c'est estimateur sans biais. C'est un estimateur *consistant* dans le sens où, d'après la loi forte des grands nombres, on a

$$\lim_{n\to\infty}\overline{X}_n \;=\; E_\theta(X)\,,\qquad P_\theta-\text{p.s.}$$

pour toute valeur du paramètre pour laquelle la moyenne existe.

Lorsque les lois P_θ ont un moment d'ordre deux, le théorème central limite permet de préciser asymptotiquement l'erreur commise en estimant la moyenne par la moyenne empirique de l'échantillon: on a sous la loi P_θ que

$$\sqrt{n}\left(\overline{X}_n - m\right) \xrightarrow{(loi)} \mathcal{N}(0,\sigma_\theta^2)\,,$$

où σ_θ^2 désigne la variance sous la loi P_θ.

• **Le maximum de vraisemblance:** Supposons que le modèle statistique est formé de lois discrètes (i.e. P_θ est une mesure de probabilités sur $\mathbb{Z}$) et notons $f(x \mid \theta) = P_\theta(X = x)$ pour $x \in \mathbb{Z}$. On appelle maximum de vraisemblance la fonction qui aux entiers $x_1, \ldots, x_n$ associe une valeur $d(x_1, \ldots, x_n) \in \Theta$ en laquelle la fonction

$$\theta \;\to\; P_\theta\left(X_1 = x_1, \ldots, X_n = x_n\right) \;=\; \prod_{k=1}^{n} f(x_k \mid \theta)$$

est maximale. On a une définition analogue lorsque les lois P_θ ont une densité $f(\cdot \mid \theta)$ par rapport à la mesure de Lebesgue.

Par exemple, quand on considère comme modèle statistique la famille $(P_\theta, \theta \in [0,1])$ des lois de Bernoulli, la moyenne empirique fournit un estimateur qui converge p.s. vers la valeur du paramètre θ quand la taille de l'échantillon tend vers ∞ (loi des grands nombres). Déterminons ensuite l'estimateur du maximum de vraisemblance pour les valeurs $x_1, \ldots, x_n$ de l'échantillon. Si on note $k = x_1 + \ldots + x_n$ le nombre de valeurs 1 prises par l'échantillon, on cherche donc à trouver la valeur $\theta \in [0,1]$ pour laquelle $\theta^k(1-\theta)^{n-k}$ est maximal. La dérivée en θ de l'expression précédente vaut $k\theta^{k-1}(1-\theta)^{n-k} - (n-k)\theta^k(1-\theta)^{n-k-1}$, et s'annule uniquement pour $\theta = k/n$. On voit donc que dans cette situation, l'estimateur du maximum de vraisemblance coincide avec la moyenne empirique.

(b) Risque quadratique

De façon informelle, on a interet à travailler avec les meilleurs estimateurs possibles. Pour donner un sens précis, on évalue le *risque quadratique* d'un estimateur d par

$$\mathcal{R}(d, \theta) \;=\; E_\theta\left(\left(d(X_1 \ldots, X_n) - \theta\right)^2\right) .$$

C'est une fonction qui dépend du paramètre; lorsque l'estimateur d est sans biais, le risque quadratique coincide avec la variance de cet estimateur. Par exemple, le risque quadratique pour la moyenne empirique dans le modèle statistique des lois de Bernoulli est

$$\mathcal{R}(\overline{X}_n, \theta) \;=\; \frac{\theta - \theta^2}{n} .$$

On voit que dans ce cas, le risque quadratique est maximal pour $\theta = 1/2$ et que $\max_{\theta \in [0,1]} \mathcal{R}(\overline{X}_n, \theta) = 1/(4n)$. En particulier on peut réduire le risque quadratique en augmentant la taille de l'échantillon.

Il est clair qu'il ne peut pas exister d'estimateur qui minimise le risque quadratique pour toutes les valeurs possibles du paramètre (l'estimateur constant $d(X_1, \ldots, X_n) = \theta_0$ a un risque quadratique nul pour $\theta = \theta_0$, donc si on voulait minimiser le risque pour toutes les valeurs du paramètre, il faudrait un estimateur qui donne toujours la valeur exacte du paramètre, ce qui est impossible). Comme on cherche à minimiser le risque indépendamment du paramètre, on est amené à introduire la notion suivante: on dira que d est un *estimateur minimax* si

$$\max_{\theta \in \Theta} \mathcal{R}(d, \theta) \;=\; \min \max_{\theta \in \Theta} \mathcal{R}(\delta, \theta)$$

où le minimum est pris sur l'ensemble des estimateurs.

Par exemple, si on cherche à estimer le paramètre d'une loi de Bernoulli à partir d'un échantillon de taille 1, on peut montrer (exercice) que l'estimateur minimax est $d(X) = \frac{1}{2}X + \frac{1}{4}$. Plus généralement, si on dispose d'un échantillon de taille n, l'estimateur

$$d(X_1, \ldots, X_n) \;=\; \frac{X_1 + \ldots + X_n + \sqrt{n/4}}{n + \sqrt{n}}$$

est minimax. On observera que cet estimateur a un biais et on pourra vérifier (exercice) que son risque quadratique est constant.

(c) Intervalle de confiance

Considérons un vote avec un assez grand nombre d'électeurs. Quand le scrutin est clot, on commence à dépouiller les bulletins, et assez vite on est en mesure de donner une estimation du résultat final. En pratique, on ne donne pas une estimation numérique (telle liste obtient 18 % des votes), mais une *fourchette*, c'est-à-dire un petit intervalle dans lequel on estime que le pourcentage exact figure.

La taille de la fourchette dépend de la confiance qu'on souhaite avoir dans l'estimation. Par exemple, on peut vouloir que la probabilité que le pourcentage exact d'une liste soit bien dans la fourchette dépasse 0,95. L'aléas vient bien sûr de la bonne ou mauvaise répartition des bulletins dépouillés. Si par exemple sur 100 000 electeurs, 200 on votés pour la liste A, il est possible (mais très improbable) que sur les 1000 premiers bulletins dépouillés figurent 100 de la liste A, auquel cas on attribura en première estimation le pourcentage erroné de 10% à cette liste. La seule façon d'être certain à 100% que le pourcentage exact figure bien dans une petite fourchette est de dépouiller la quasi-totalité des bulletins, ce qui n'est pas intéressant si on est impatient.

Quand on s'est fixé un niveau de confiance (par exemple 0,95), plus le nombre de bulletins dépouillés est grand, plus les fourchettes sont étroites. On arrive à des estimations très précises bien avant d'avoir fini le dépouillement. Bien entendu, plus on exige un haut le niveau de confiance, plus les fourchettes sont larges.

En statistique, on parle souvent *d'intervalle de confiance* au lieu de fourchette. Plus précisément, considérons un modèle statistique $(P_\theta, \theta \in \Theta)$, un échantillon de taille n, $X_1, \ldots, X_n$. On se donne un niveau de confiance $1 - \alpha$, où $\alpha \in]0, 1[$ represente la probabilité de se tromper qu'on tolère. On appelle intervalle de confiance de niveau $1 - \alpha$ un intervalle $I(X_1, \ldots, X_n) = [a(X_1, \ldots, X_n), b(X_1, \ldots, X_n)]$ tel que

$$P_\theta \left(\theta \in I(X_1, \ldots, X_n)\right) \geq 1 - \alpha \qquad \text{pour tout } \theta \in \Theta\,.$$

Bien sûr, pour un échantillon de taille donnée, on souhaite avoir un niveau de confiance haut et une taille de l'intervalle petite, et ces deux conditions sont antagonistes.

Concrètement, pour trouver un intervalle de confiance, on part d'un estimateur $d(X_1, \ldots, X_n)$ et on cherche à mesurer l'erreur de cet estimateur pour déterminer la taille d'un voisinage de $d(X_1, \ldots, X_n)$ qui donnera l'intervalle de confiance au niveau souhaité. Par exemple, considérons encore le modèle de la loi de Bernoulli de paramètre $\theta \in [0, 1]$ et l'estimateur de la moyenne empirique $\overline{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$. L'inégalité de Bienaymé-Chebichev conduit à l'encadrement

$$P_\theta \left(|\overline{X}_n - \theta| \geq \varepsilon \right) \;\leq\; \frac{\mathrm{Var}_\theta(\overline{X}_n)}{\varepsilon^2} \;=\; \frac{\theta - \theta^2}{n\varepsilon^2} \;\leq\; \frac{1}{4n\varepsilon^2}\,,$$

où l'on a noté $\mathrm{Var}_\theta(\overline{X}_n)$ la variance de $\overline{X}_n$ sous la probabilité P_θ. La majoration en $1/(4n\varepsilon^2)$ ne dépend pas du paramètre θ; on a fixé le niveau de confiance à $1-\alpha$, c'est-à-dire qu'on doit prendre $1/(4n\varepsilon^2) \leq \alpha$. On conclut que

$$[\overline{X}_n - \frac{1}{2\sqrt{n\alpha}}, \overline{X}_n + \frac{1}{2\sqrt{n\alpha}}]$$

est un intervalle de confiance de niveau α pour le modèle de Bernoulli.

En pratique, lorsque la taille de l'échantillon est grande, on obtient souvent de meilleures majorations de l'erreur en n'utilisant non plus l'inégalité de Bienaymé-Chebichev, mais soit des estimations gaussiennes issues du théorème central limite, soit encore des estimations de grandes déviations.

8.3 Etude d'un modèle gaussien

(a) Préliminaires.

Les lois gaussiennes jouent un rôle fondamental dans la théorie des probabilités, dû au theorème central limite. Heuristiquement, elles apparaissent souvent en physique à cause de "bruits" qui perturbent de façon aléatoire les mesures effectuées.

On introduit d'abord trois lois de probabilités sur $\mathbb{R}$ très importantes dans l'étude du modèle gaussien. On considère ici $X_1, \cdots, X_n$ un échantillon de taille n de la loi normale centrée réduite $\mathcal{N}(0,1)$.

• On appelle loi du χ^2 (chi-deux) à n degrés de liberté et on note $\chi^2(n)$ la loi de $X_1^2 + \ldots + X_n^2$. Par changement de variables en coordonnées polaires, on voit que sa densité est

$$\frac{1}{\Gamma(n/2)} 2^{-n/2} \mathrm{e}^{-x/2} x^{n/2-1}\,, \qquad x > 0\,,$$

où $\Gamma(q) = \int_0^\infty \mathrm{e}^{-t} t^{q-1} dt$ désigne la fonction gamma. Cette loi a pour moyenne n et variance $2n$. Pour $n = 2$, elle coincide avec la loi exponentielle de paramètre $1/2$.

• La loi de **Fisher** à (n_1, n_2)-degrés de libertée est celle donnée par la variable quotient $(n_2 Y_1)/(n_1 Y_2)$ où Y_1 et Y_2 désignent deux variables indépendantes de lois $\chi^2(n_1)$ et $\chi^2(n_2)$, respectivement. Elle est notée $F(n_1, n_2)$.

• La loi de **Student** à n degrés de liberté est notée $t(n)$. Elle est donnée par le quotient $\sqrt{n}X/\sqrt{Y}$ où X et Y sont deux deux variables indépendantes, de lois $\mathcal{N}(0,1)$ et $\chi^2(n)$, respectivement.

Le résultat élémentaire suivant présente un estimateur standard d'un échantillon gaussien réel.

Proposition. *Soit $X_1, \ldots, X_n$ un échantillon de taille n de la loi gaussienne $\mathcal{N}(m, \sigma^2)$, i.e. de moyenne $m \in \mathbb{R}$ et de variance $\sigma^2 > 0$. On appelle*

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n} \quad , \quad S_n^2 = \frac{(X_1 - \overline{X}_n)^2 + \ldots + (X_n - \overline{X}_n)^2}{n-1}$$

la moyenne empirique et la variance empirique. Alors $\overline{X}_n$ et S_n^2 sont indépendants et

$$\sqrt{n}(\overline{X}_n - m) \stackrel{\mathcal{L}}{=} \mathcal{N}(0, \sigma^2) \quad , \quad \sigma^{-2} \sum_{k=1}^n (X_k - m)^2 \stackrel{\mathcal{L}}{=} \chi^2(n)$$

$$\sigma^2 \sum_{k=1}^{n} (X_k - \overline{X}_n)^2 \stackrel{\mathcal{L}}{=} \chi^2(n-1) \quad , \quad \sqrt{n}(\overline{X}_n - m)/S_n \stackrel{\mathcal{L}}{=} t(n-1)$$

Preuve: Un calcul immédiat de variance montre que $\overline{X}_n$ est indépendant de $(X_1 - \overline{X}_n), \ldots, (X_n - \overline{X}_n)$, et en particulier $\overline{X}_n$ et S_n sont indépendants. La première identité en loi est évidente, la seconde découle de la définition même de la loi $\chi^2(n)$. La troisième est un exercice facile sur les lois gaussiennes (observer que la moyenne empirique d'un échantillon est obtenue par projection orthogonale du vecteur dans $\mathbb{R}^n$ de coordonnées $(X_1, \ldots, X_n)$ sur la diagonale de vecteur directeur $(1, \ldots, 1)$). La dernière en découle alors grâce à l'indépendance de $\overline{X}_n$ et de S_n^2 et à la définition même de la loi de Student. ■

(b) Estimation de paramètres

Nous allons utiliser ce résultat précédent pour étudier un modèle statistique gaussien. Considérons d'abord le modèle $(P_m = \mathcal{N}(m, \sigma^2), m \in \mathbb{R})$ où la variance σ^2 est connue et la moyenne un paramètre. La moyenne empirique $\overline{X}_n$ est un estimateur sans biais de m. Il est exponentiellement consistant puisque d'après la proposition précédente, la loi de $\sqrt{n}(\overline{X}_n - m)$ est celle de σN (N étant toujours une variable normale $\mathcal{N}(0,1)$). Il s'ensuit que pour tout $\varepsilon > 0$ on a

$$P_m\left(|\overline{X}_n - m| > \varepsilon\right) = \mathbb{P}(\sigma|N| > \varepsilon\sqrt{n}),$$

et on sait bien que le terme de droite est de l'ordre de $\exp\{-n\varepsilon^2/(2\sigma^2)\}$ quand n est grand. On peut montrer que parmi les estimateurs sans biais, il est celui de risque quadratique minimum. Le fait que $\overline{X}_n - m$ suit une loi gaussienne centrée de variance σ^2/n permet également de déterminer les intervalles de confiance pour la moyenne empirique. On trouve que si on note $\phi(y)$ la valeur x pour laquelle $\mathbb{P}(N > x) = y$, alors

$$\left[\overline{X}_n - \frac{\sigma\phi(\alpha/2)}{\sqrt{n}}, \overline{X}_n + \frac{\sigma\phi(\alpha/2)}{\sqrt{n}}\right]$$

est un intervalle de confiance de niveau $1 - \alpha$.

Considèrons ensuite le modèle $(P_{\sigma^2} = \mathcal{N}(m, \sigma^2), \sigma 2 > 0)$ où la moyenne est connue et la variance σ^2 est un paramètre. On peut estimer la variance empiriquement par

$$\overline{\sigma^2}_n = \left((X_1 - m)^2 + \ldots + (X_n - m)^2\right)/n$$

(on observera que $\overline{\sigma^2}_n$ coincide avec l'estimateur de la moyenne empirique pour l'échantillon statistique $(X_1 - m)^2, \ldots, (X_n - m)^2$). On peut à nouveau utiliser la proposition pour trouver des intervalles de confiance pourvu que l'on sache que l'espace des paramètres pour la variance est borné.

Enfin, pour le modèle plus général $(P_{m,\sigma^2} = \mathcal{N}(m, \sigma^2), m \in \mathbb{R}, \sigma^2 > 0)$ où la moyenne et la variance sont inconnue, on peut estimer la moyenne par la moyenne empirique, et la variance par la variance empirique. Là encore la proposition est utile pour déterminer les intervalles de confiance. On peut également estimer moyenne et variance par le principe du maximum de vraisemblance. Il est facile de vérifier que l'estimateur du maximum de vraisemblance de la moyenne coincide avec la moyenne

empirique, mais que l'estimateur du maximum de vraisemblance de la variance est $\frac{n-1}{n}S_n^2$ (observer que cet estimateur est biaisé). Un troisième estimateur très utilisé pour le modèle gaussien consiste à estimer la moyenne en cherchant la valeur μ pour laquelle $\sum_{k=1}^n (X_k - \mu)^2$ est minimal. On trouve encore que le minimum est atteint pour la moyenne empirique, et la valeur de ce minimum est $(n-1)S_n^2$, où S_n^2 est la variance empirique. Ce principe d'estimation porte le nom de *méthode des moindres carrés*. Il est à la base de la théorie des erreurs.

(c) Tests

Dans un grand nombre de problèmes statistiques concrets, on ne cherche pas nécessairement à estimer précisément la loi d'une variable, mais seulement à savoir si elle vérifie telle ou telle hypothèse. Par exemple, lors du second tour d'une election présidentielle en France, on souhaite avant toute chose savoir qui a gagné, c'est-à-dire si le candidat A a obtenu plus de 50 % des suffrages. Ensuite seulement on s'intéressera à son score exact. Autrement dit, si p désigne le pourcentage des electeurs qui aura voté pour A, on veut tout d'abord tester l'hypothèse $p > 0,5$. Comme dans le cas de l'estimation de paramètre, on a besoin de connaître la confiance qu'on peut avoir dans ce test. Donnons deux exemples classiques pour le modèle gaussien $(\mathcal{N}(m, \sigma^2), m \in \mathbb{R}, \sigma^2 > 0)$.

• **Test de Student** On se fixe une valeur $m_0 \in \mathbb{R}$ et on veut tester l'hypothèse H(0): "$m \leq m_0$" contre l'hypothèse H(1): "$m > m_0$". Pour cela, on utilise le fait que $T := \sqrt{n}(\overline{X}_n - m)/S_n$ suit la loi de Student à $(n-1)$ degrés de liberté, $t(n-1)$. Fixons un niveau $\alpha \in [0,1]$ et notons β le quantile d'ordre $1-\alpha$ de $t(n-1)$, c'est-à-dire que la probabilité que $T \leq \beta$ vaut $1-\alpha$. On a donc que sous P_{m,σ^2}, la probabilité que $\overline{X}_n \leq m + \beta S_n/\sqrt{n}$ vaut exactement $1-\alpha$. Le test de Student consiste à accepter l'hypothèse H(0) si

$$\overline{X}_n \leq m_0 + \beta S_n/\sqrt{n}\,.$$

Par monotonie, on voit donc que la probabilité d'accepter H(0) quand $m \leq m_0$ est toujours supérieure à $1-\alpha$, alors qu'elle est toujours inférieure à $1-\alpha$ quand $m > m_0$.

• **Test de Fisher** Cette fois on veut tester l'hypothèse sur la variance H(0): "$\sigma^2 \leq \sigma_0^2$" contre H(1): "$\sigma^2 > \sigma_0^2$". On raisonne de même que pour le test de Student en se fixant un niveau α. On sait que sous P_{m,σ^2}, $(n-1)S_n/\sigma^2$ suit la loi $\chi^2(n-1)$. On note γ le quantile d'ordre $1-\alpha$ de cette dernière, et le test de Fisher consiste à accepter H(0) si $S_n^2 \leq \gamma\sigma_0^2/(n-1)$.

Nous allons conclure ce chapitre en présentant une application importante en statistique appliquée. Considérons une variable aléatoire qui ne prend qu'un nombre fini de valeurs, disons $\{1, \ldots, k\}$. On se donne un échantillon de taille n de cette loi, $X_1, \ldots, X_n$, et on note

$$\overline{p}_n(i) \;=\; \frac{1}{n}\sum_{j=1}^n \mathbf{1}_{\{X_j = i\}}\,, \qquad i \in \{1, \ldots, k\}$$

l'estimateur empirique de $p(i)$. **Le test du** χ^2 s'appuie sur le résultat suivant:

Proposition. *Si pour chaque entier $n \geq 0$, $X_1, \ldots, X_n$ est un échanitllon de taille n d'une loi discrète p à valeurs dans $\{1, \ldots, k\}$, avec $p(i) > 0$ pour tout $i = 1, \ldots, k$.*

Alors la suite

$$n \sum_{i=1}^{k} \frac{(p(i) - \overline{p}_n(i))^2}{p(i)}$$

converge en loi vers $\chi^2(k-1)$.

Preuve: Posons pour tout entier $n \geq 1$

$$Y_n^{(i)} = \mathbf{1}_{\{X_n = i\}} \quad , \quad Y_n = (Y_n^{(1)}, \ldots, Y_n^{(k)}) \, .$$

Les variables $Y_1, \ldots$ forment une suite i.i.d. de moyenne $(p(1), \ldots, p(k))$ et de matrice de dispersion

$$D_{i,i} = p(i) - p(i)^2 \quad , \quad D_{i,j} = -p(i)p(j) \text{ pour } i \neq j \, .$$

Le théorème central limite multidimensionel énonce la convergence en loi lorsque $n \to \infty$ du vecteur

$$\frac{Y_1^{(i)} + \cdot + Y_n^{(i)} - np(i)}{\sqrt{n}} = \sqrt{n} \, (\overline{p}_n(i) - p(i)) \, , \qquad i = 1, \ldots, k$$

vers un vecteur gaussien $(G_1, \ldots, G_k)$ centré de matrice de covariance $D = (D_{i,j})$. En conséquence $n \sum_{i=1}^{k} (p(i) - \overline{p}_n(i))^2 / p(i)$ converge en loi vers $\sum_{i=1}^{k} G_i^2 / p(i)$, et le calcul matriciel usuel conduit alors au résultat. ■

On construit grâce à ce résultat des tests pour accepter ou rejeter des hypothèses du type "la loi correspondant à l'échantillon $X_1, \ldots, X_n$ est p" en décidant si l'écart observé entre la loi présumée p et la loi empirique $\overline{p}_n$ est de l'ordre ou non de ce qui est prédit par la théorie. De même, on peut tester l'indépendance de caractères observés.